VAET: A Visual Analytics Approach for E-transactions Time-Series

Cong Xie, Wei Chen, *Member, IEEE*, Xinxin Huang, Yueqi Hu, Scott Barlowe, and Jing Yang

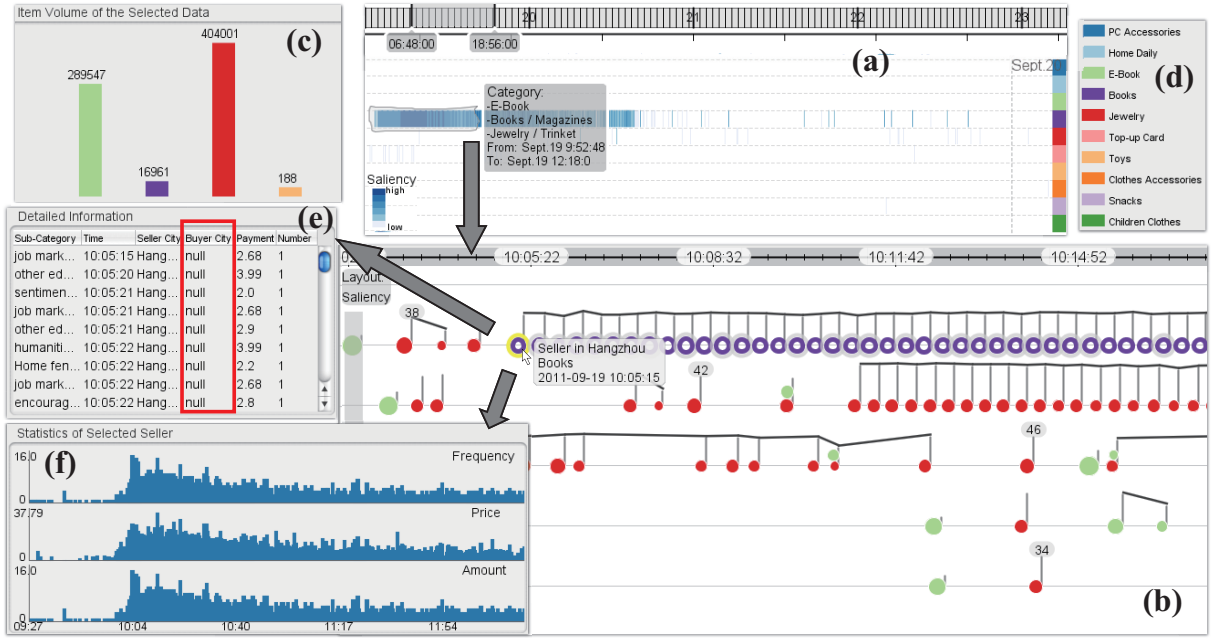

Fig. 1. The visual analysis interface of the VAET system. (a) The time-of-saliency (TOS) map overviews the saliency of each transaction computed with a probabilistic decision tree learner. (b) The KnotLines view shows the detailed information of transactions. The unfilled knots indicate fake transactions. (c) The legend of the sales category and the (d) bar chart shows the item volume of the selected transactions in TOS map. (e) Detailed transaction information and (f) statistical information are shown in auxiliary views.

Abstract—Previous studies on E-transaction time-series have mainly focused on finding temporal trends of transaction behavior. Interesting transactions that are time-stamped and situation-relevant may easily be obscured in a large amount of information. This paper proposes a visual analytics system, Visual Analysis of E-transaction Time-Series (VAET), that allows the analysts to interactively explore large transaction datasets for insights about time-varying transactions. With a set of analyst-determined training samples, VAET automatically estimates the saliency of each transaction in a large time-series using a probabilistic decision tree learner. It provides an effective time-of-saliency (TOS) map where the analysts can explore a large number of transactions at different time granularities. Interesting transactions are further encoded with KnotLines, a compact visual representation that captures both the temporal variations and the contextual connection of transactions. The analysts can thus explore, select, and investigate knotlines of interest. A case study and user study with a real E-transactions dataset (26 million records) demonstrate the effectiveness of VAET.

Index Terms—Time-Series, Visual Analytics, E-transaction

1 Introduction

Massive web data, such as online transactions, are being collected and stored each second [19]. The E-transaction time-series contains transactions among multiple users in a time range (e.g., different buyers purchase commodities from the same seller). Each record contains a time stamp, the IDs of the seller and buyer, and the associated attributes of the commodities. Each record is an atomic element representing an online transaction among a seller and a buyer.

Analyzing E-transaction time-series in a temporal context is critical for understanding transaction behavior, learning user preferences, and discovering temporal trends. This work is motivated by interviews with several analysts in an online customer-to-customer retail store. They pointed out that the following are important questions that they are often unable to answer when analyzing large E-transaction time-series:

- What are the temporal and contextual connections among multiple transactions of a seller? For example, a large number of transactions involving a single seller may occur in a short time. If the sales originate from the same buyer, there can be a special relationship between the buyer and seller. The analysts need to study related transactions whose time stamps, commodities, payment amounts, buyer locations, and other attributes are associated.

- *Cong Xie, Xinxin Huang, and Wei Chen are with State Key Lab of CAD&CG, Zhejiang University. E-mail: {xiecong, huangxinxin, chenwei}@cad.zju.edu.cn.*
- *Wei Chen is the corresponding author. He is also with the Cyber Innovation Joint Research Center, Zhejiang University.*
- *Yueqi Hu and Jing Yang are with Dept. of Computer Science, University of North Carolina at Charlotte. E-mail: {yhu12, Jing.Yang}@uncc.edu.*
- *Scott Barlowe is with Western Carolina University. E-mail: sabarlowe@email.wcu.edu.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; posted online xx xxx 2014; mailed on xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

- What are the most common transaction patterns? For example, transactions of a seller are usually sparse in an ordinary work day. However, most sellers can have frequent transactions with large commodity numbers on Christmas Day when they launch sales promotions. The analysts need to go through the dataset to find such patterns.
- How to identify transactions with interesting patterns? For example, fake transactions made by colluding buyers to accumulate seller credits are of particular interest to the analysts. With the interesting pattern defined, the analysts need to discover transactions with particular attribute values and correlations in a large dataset.
- How to examine a single transaction within context? For example, a transaction with a small payment amount and a large commodity number could be a fake transaction used to improve the seller's rank on an e-commerce website. To confirm that a transaction is fake, the analysts need to associate the transaction with information about the buyer and seller.

We argue that automatic data mining processes are not adequately flexible and precise to answer the above questions, because E-transaction time-series can be subtle, inter-leaving, and varied. There is a dire need for visual analytics approaches which allow the analysts to flexibly form and test hypotheses with instant visual feedback by integrating computation power, human perceptual capabilities, and domain knowledge. However, there are no ready-to-use visualization approaches for the above scenarios. Existing research efforts on multivariate time-series visualization mainly focus on summarizing global and/or temporal trends of multiple dimensions or discovering patterns of individual dimensions, such as Sparklines [23]. Most of the existing approaches are not suitable for the aforementioned tasks.

This paper presents a novel visual analytics system, called Visual Analysis of E-transaction Time-Series (VAET), that seeks to explore e-transaction time-series for analysis of transaction patterns among multiple users in a temporal context. VAET has the following two main visual analytics components: (1) Overview: This component helps the analysts effectively identify salient transactions from a large dataset. Here, the saliency refers to the relevancy of a transaction to a certain analysis task. VAET uses a probabilistic decision tree learner to first calculates a saliency value for each transaction to reveal its relevance to the analysis target (e.g., the possibility of being a fake transaction). Then, the saliency values are displayed in a pixel-oriented display [10], called a time-of-saliency (TOS) map (Figure 1 (a)). The map provides a workspace to explore and select potentially interesting transactions at different time granularities; (2) Detail view: This component allows the analysts to conduct detailed examination on interesting transactions for insights. In particular, the transactions selected from the overview are displayed using a new visual metaphor called KnotLines. In this view, lines reveal the connections among transactions and temporal trends. Knots along the lines encode the detailed information of the associated transactions (Figure 1 (b)). Distinctive attributes as well as temporal and contextual correlations of multi-user transactions are thus intuitively presented to the analysts for insight discovery. The TOS map and KnotLines are coordinated so that the analysts can quickly identify interesting transactions from a large dataset.

Evaluations have been conducted to test the capabilities of VAET in analyzing multi-user transaction patterns. A case study and a user study with a real online transaction dataset demonstrated that VAET was effective in supporting a variety of analysis tasks (see Section 8 and Section 9).

The main contributions of this paper include:

- A visual analytics system that allows the analysts to effectively analyze a large E-transaction time-series in a temporal context;
- An approach to detecting and visualizing salient transactions from large datasets;
- A novel visual metaphor to compactly place and encode distinctive attributes as well as temporal and contextual correlations of multi-user transactions.

2 Related Work

2.1 Visual Analysis of E-transaction Data

The transaction data contains various types of attributes, such as numerical, temporal and categorical. The Sparklines [23] can be used to visualize multiple trends in financial data. Liu et al. [11] proposed a visualization system called SellTrend for analyzing airline travel purchase requests. WireVis [3] was proposed to search on predefined patterns in large wire transaction datasets. Visual analytics approaches have been proposed to explore web clickstreams of online transactions [26]. Our approach is among the earliest visual analytics approaches for the exploration of temporal and contextual connections in multi-user transactions.

Transaction data often have multi-dimensional attributes. Analyzing them often requires the integration of well-designed data mining models. Probabilistic models are employed to model user behavior [12], resulting in user clusters. This scheme has been successfully applied to classify E-transaction data into different types [2]. Association analysis is another widely used model for transaction data. Hao et al. [7] proposed the DAV system to visualize the relationships of associated products.

2.2 Visual Analysis of User Behavior Time-Series

There are many previous works on the analysis and visualization of user behavior time-series. Here we only summarize the most relevant ones and categorize them as techniques for analyzing individual behaviors, user interactions, and group behaviors. More details can be found in [1], [9] and [14].

Temporal Individual Behavior Patterns

Many visualization approaches designed to analyze user behavior data are focused on exploring the temporal behavior patterns of individuals. For example, TimeSearcher [8] allows users to select interesting time-series using a rectangular query region. LifeLines [16] visualizes health-related incidents of patients along a timeline. Most previous works utilize high-dimensional visual exploration tools such as parallel coordinates [4] to explore extracted patterns. Density-based display techniques [6], [10] are capable of showing large time-series datasets for real-time monitoring. Additional visual exploration techniques include time trajectory [21] and [13].

User Interaction Patterns

To discover user interaction patterns, conventional solutions consider the user network as a social network and analyze its global structure. For example, Sallaberry et al. [20] provide an overview of dynamic network evolution over time. Other approaches emphasize the user interaction characteristics such as email connections [25] and instant messages [27]. However, these methods are focused on the structural changes rather than the temporal variations of the interactions. Other approaches aim to reveal the relationships among multiple users in a temporal context. For instance, Storyline [22] shows the narrative threads that form a plot or a subplot in works of fiction. The history flow approach successfully reveals author collaboration patterns [24]. Code Swarm [15] visualizes the animated histories of software project evolution. VAET reveals both the temporal patterns of multi-user behavior and their atomic level correlations. It improves the above approaches by allowing the analysts to explore a large number of transactions at different granularities.

3 Problem Definition

Multi-user transaction data is a special type of user behavior data with a focus on characterizing raw, detailed, and subtle inter-user transactions. An E-transaction time-series dataset contains information about each E-transaction, including information about transaction time, the buyer, and the seller. Each E-transaction records a transaction between a buyer and a seller.

In general, an E-transaction contains the following attributes:

- **User information** includes the IDs and other information about the buyer and the seller who make the transaction, e.g., their age group, gender, and location.

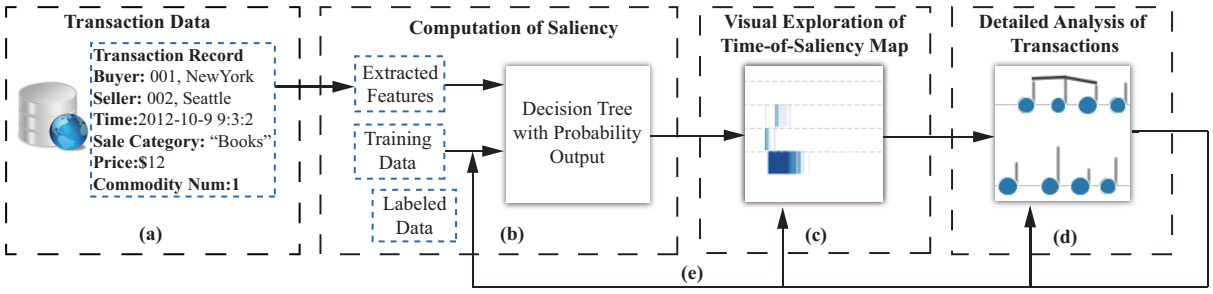

Fig. 2. Conceptual overview of VAET. (a) A single transaction is characterized with appropriate feature descriptions and analyzed with (b) decision tree approach. (c) The saliency values of all transactions are selected and explored. (d) Interesting transactions are further studied within a detailed view. (e) The entire exploration process is iterative.

- **Transaction information** includes the time stamp and other information about the commodities, e.g., the payment amount, the number, and the sales category of the commodity.

The above attributes can be numerical, ordinal, categorical, textual, or temporal.

The analysts usually conduct a complex task through a set of low level tasks. These tasks typically focus on the behavior of the seller, such as:

T1 Identifying time periods and/or sales categories of interest.
T2 Identifying transactions with interesting patterns in specific attributes (e.g., payment amount $\geq$ 500) and examining their detailed information.
T3 Identifying sellers with interesting transaction patterns, such as a seller making frequent transactions with small payment amounts.
T4 Examining the transaction patterns of a specific seller.

We use the term "saliency" to quantitatively describe the degree of relevance of a transaction to an analyst-defined target. According to our interview, identifying and examining salient transactions are critical yet challenging tasks in E-transaction time-series exploration. Typically, the analysts need to manually identify salient transactions by iteratively querying a dataset and examining the attribute values and relationships among the retrieved transactions. Moreover, the analysts often need to examine salient transactions together with information such as the users' historical data to justify their behavior or reveal interesting patterns. This process is typically laborious and tedious. VAET is designed to ease this process and improve the overall operation efficiency.

4 Approach Overview

The goal of VAET is to identify and explore interesting transactions by selecting those with high saliency and studying them. This is accomplished by integrating the capabilities of both data mining and visualization techniques within the following iterative visual exploration pipeline.

Step 1 Saliency computation with decision tree: A set of features are extracted from each transaction. The analysts manually label the features of some transactions as the training data. Using these features, a probabilistic decision tree learner is constructed upon the training data. It is then employed to produce the saliency values for each unlabeled transaction (Figure 2 (b)).

Step 2 Browsing and selection using the TOS map: The saliency values of all transactions are mapped to a compact, density-based Time-Of-Saliency (TOS) map. In this map, transactions are ordered by time and categories and represented by pixels whose colors correspond to saliency values. The analysts can interactively explore the map, investigate the global distribution and local patterns, and select interesting transactions according to the saliency values from this view. (Figure 2 (c)).

Step 3 Detailed analysis using KnotLines: The analyst-selected transactions are visualized with a novel visual metaphor, KnotLines, that allows the study of multiple attributes and contextual connections (Figure 2 (d)). The transactions identified as salient by the analysts can be labeled and fed back into Step 1 to continue the iterative process (Figure 2 (e)).

The analysts can iteratively loop over the above steps by adjusting the labeled dataset, navigating the map, and exploring the interesting transactions. Both the TOS map and the KnotLines visualization provide scalable exploration, such as time interval selection and detail review.

5 Saliency Computation with Decision Tree

Computing the saliency values is inherently context-aware and task-oriented. For many tasks, the saliency values cannot be directly derived from the transaction attributes. For instance, when the analysts search for abnormal transactions, the transaction frequency of the sellers often need to be considered. It is also impractical to let the analysts manually specify the saliency value of each transaction. Thus, we propose to compute the saliency value of each record by defining and computing a set of features for transactions.

In particular, our approach computes the saliency values as a *probability estimation problem* by means of a *probabilistic decision tree* [17]. We chose decision trees because it can handle both continuous and categorical attributes, and is easy to explain. The decision tree is initially constructed with the features of a set of analyst-determined training data. Applying the decision tree to the features of each unlabeled transaction yields a probability ranging from 0 to 1, which is used as the saliency value of the underlying transaction. The transactions manually labeled by the analysts as salient can be added into the training dataset in the subsequent analysis (Figure 2 (e)).

5.1 Feature Extraction

VAET computes a set of analyst-specified temporal and contextual characteristics for each transaction as a set of features. In general, three types of features are defined:

Basic Features One straightforward way to determine whether a transaction is interesting is to use the values of specified attributes as basic features, such as the payment amount of a commodity. In addition, the analysts can define new attributes. For example, if a seller is in the interesting list given by the analyst, he or she is considered as a salient seller, as shown in Figure 3. The collection of these attributes constructs a set of basic features.

Textual Features A transaction may contain textual information, such as the comment of a commodity. VAET examines whether the textual information contains sensitive words in a analyst-specified list. The analysts keep a dictionary for sensitive words and phrases collected manually from the past several months. For example, in a kind of fraud transactions, the buyers want their cash back as soon as possible. "cash back" is a sensitive phrase here. Sensitive words vary in different situations, and can be regarded as textual features.

Temporal Features Temporal patterns of a sequence of transactions are essential for identifying interesting patterns in the datasets. For example, the transaction amount of a seller in a time interval indicates

his or her popularity. However, time-oriented relations are difficult to discover with conventional decision tree approaches.

To address this problem, VAET uses the transaction frequency of the seller in every time interval as a measure of the temporal trend. The size of the time interval depends on the data collection configuration. For example, a typical choice is 5 minutes. Figure 3 provides an example of feature extraction.

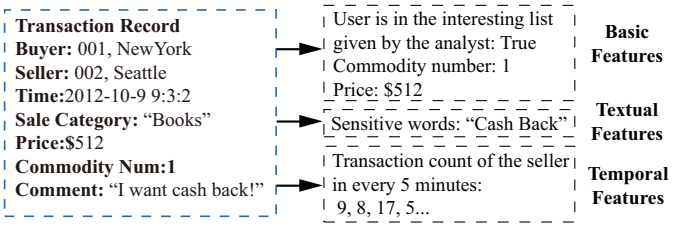

Fig. 3. VAET extracts the basic, textual, and temporal feature descriptions of a transaction. The features are specified by the analysts.

5.2 Estimating saliency using Probabilistic Decision Tree

A decision tree is initially built with a training dataset, which consists of extracted features from analyst-labeled transactions. The training dataset can be manually updated in the visual exploration process by adding analyst-identified transactions, as indicated in Figure 2 (e). In our approach, the decision tree is automatically constructed from the training data using the well-established C4.5 algorithm [18], which recursively splits the training set into subsets based on the features. In the decision tree (see Figure 4 for an example), a leaf represents a class (salient or non-salient) and an interior node corresponds to a feature. At each interior node, C4.5 splits the samples into subsets based on the feature which produces the highest normalized information gain and assigns the feature to that node.

The constructed decision tree [5] divides the dataset into two classes by identifying each unlabeled transaction x as either salient or non-salient based on the extracted features. Because a non-salient transaction can be classified into the salient class (false positive), each transaction is assigned a probability of being salient, namely the saliency value, using the following algorithm.

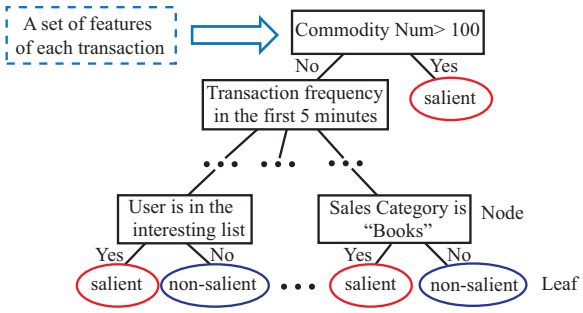

Fig. 4. The structure of the decision tree.

The probabilities are estimated based on the transactions at the decision tree leaves [17]. We denote FP as the number of false positives at a leaf and TP as the number of true positives (see the confusion matrix in Table 1). The probabilistic distribution estimation at the leaf is given by:

$$P(y|x) = TP/(TP+FP) \quad (1)$$

where y denotes the estimated class (salient or non-salient).

Simply using the number of transactions in the salient class may not get sound probabilistic estimates [17], especially in a leaf with few transactions. This problem can be addressed by smoothing the probabilities with a Laplace estimate which introduces a prior probability of $1/C$ for each class. As shown in Equation 2, the saliency value of a transaction, $S(x)$, is the probability at a decision tree leaf for the salient class y. (Note that the probability distribution is estimated with an unpruned decision tree.)

Table 1. The confusion matrix at each decision tree leaf.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

$$S(x) = P(y|x) = (TP+1)/(TP+FP+C) \quad (2)$$

6 Time-Of-Saliency Map: Browsing a Large Set of Transactions

6.1 Generation of Time-Of-Saliency Map

To allow for exploring the saliency values of a large set of transactions, VAET visually displays them in a *Time-Of-Saliency (TOS) map* (see Figure 5 (a)). A TOS map is a 2D density-based display, with time along the horizontal axis and the vertical axis organized by sales categories (e.g., "Electronics Accessories" and "Clothes").

The TOS map is evenly split into rows, each of which represents a sales category. In Figure 5 (a), the color rectangles highlighted by the blue box provides a visual index of the categories along the vertical axis. Moreover, each row is divided horizontally according to the time interval. Each transaction is projected to the corresponding cell according to its time stamp and sales category.

The saliency values of all transactions projected to the same cell are summed, and the sum is mapped to the color of the cell. A default color scale or analyst-specified color scales can be used for the color mapping. The resulting TOS map visually encodes the relevance of the transactions for the analysis task. A dark region implies a set of potentially interesting transactions. In particular, a continuous dark band in a row indicates highly salient transactions in the corresponding sales category over a period of time (see the selected region in the TOS map in Figure 1 (a)).

6.2 Time-Of-Saliency Exploration

The following interactions are provided in the TOS map view and can be used to accomplish T1:

Time Windowing The TOS map shows the transactions in an analyst-adjustable time interval. An additional time windowing widget can be used to locate a specific region of the view for further and detailed study. Analysts can click and drag on the time selection bar to set the time window of the TOS map, as highlighted over the top of the TOS map in Figure 5 (a). Figure 5 (b) shows the TOS map after the analyst sets the time window.

Region-Of-Interest Selection The analysts can click on the category index (the blue box in Figure 5 (a)) to choose the transactions of the same category. A lasso tool can also be used to select interesting regions. When a region is selected, a floating text box will appear to show the information about the region. The detailed information of selected transactions can be further visualized and explored in the KnotLines view described in Section 7. In addition, a bar chart view (Figure 1 (c)) is provided to show the sales volume of categories in the selected data.

7 KnotLines: Examining Transactions in Detail

KnotLines allows the analysts to conduct detailed analysis on salient transactions selected from the TOS map. It is designed to tackle T2 - T4. KnotLines visually presents two types of information: attributes and the temporal trend of the transactions.

7.1 Data Organization and Visual Layout

7.1.1 Three-Level Organization

To study attribute similarity and temporal correlations among transactions, the selected transaction set is organized into a three-level hierarchical tree (Figure 6). At first, we used a matrix form to visualize the organization of the transactions, as shown in Figure 7 (a).

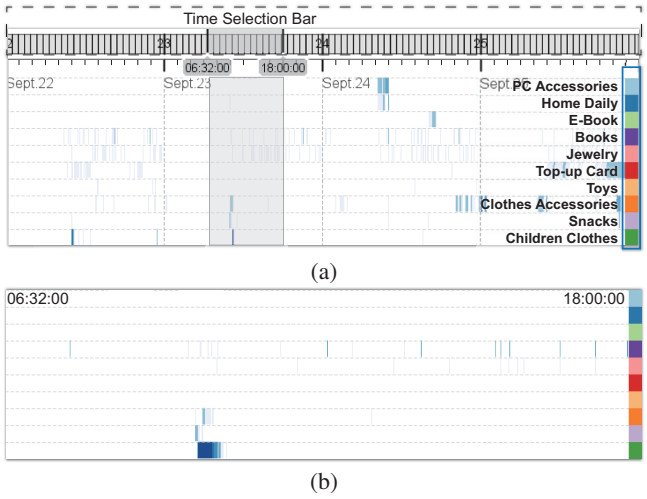

Fig. 5. (a) The TOS map visualizes the saliency distribution of transaction data along the time axis. (b) The analyst zooms in on the TOS map by selecting the time window in (a).

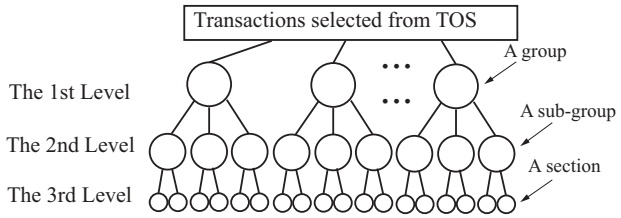

Fig. 6. Organizing a set of transactions with a three-level hierarchical tree, here $M = 3$ and $K = 2$.

Level One The whole selected transaction set is divided into N groups (level 1) according to different sellers. Each row in Figure 7 (a) represents a group. A group contains all transactions of a seller. The groups are listed from top to bottom along the vertical axis.

Level Two The transactions in a group are further divided into subgroups (level 2) according to their time stamps. The horizontal axis in Figure 7 (a) represents the time. Each row is split into M squares along the time axis which correspond to M time intervals. The lengths of all intervals are the same and can be adjusted to explore the data at different granularities. Transactions of a seller which fall into the same time interval form a sub-group (level 2).

Level Three A subgroup is further divided into sections (level 3) according to the sales categories (e.g., "Books"). In Figure 7 (a), each square is segmented into K cells, each of which represents a section. Transactions in the same section are made by the same seller, take place in the same time interval, and belong to the same sales category.

7.1.2 Compact Representation of the Visual Layout

Because the transaction volume of most sellers may be high only during parts of the day, the transaction density in the matrix shown in Figure 7 (a) can be very sparse. In addition, the number of groups N can be huge (e.g., 1 million). To make the exploration more effective, the matrix-like layout should be reformulated to be more compact.

VAET employs a simple two-step heuristic scheme that operates on each group. In the first step, empty sub-groups before the first non-empty sub-group and after the last non-empty sub-group are removed. This step results in many groups that only cover a small portion of the horizontal space, because most of them have a short time span.

To increase the space efficiency, the placement of the groups is heuristically optimized in the second step. An iterative layout strategy is used to satisfy the following principles:

- Uncluttered: groups should not overlap;
- Compact: space utilization should be high;

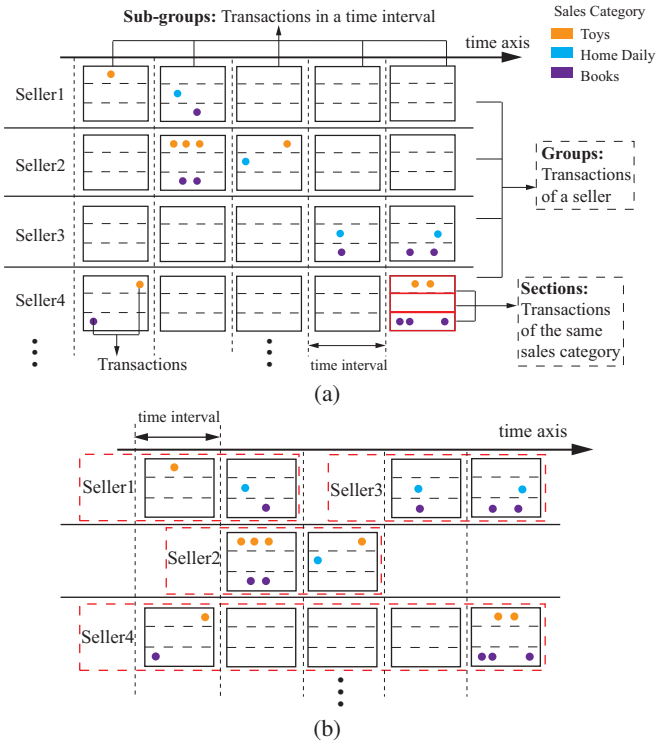

Fig. 7. (a) The visual layout of the three-level organization. (b) The compact layout derived from (a).

- Representative: important groups should have a display priority.

There are several considerations when reformulating the visual layout. First, if the time spans of two groups have one or more common time intervals, they should be placed in different rows to avoid overlap. Second, groups that do not overlap with each other can be placed in the same row to make a more compact layout, such as the groups of seller1 and seller3 in Figure 7 (b). Third, group importance can be specified by the analysts and important groups are placed in important regions. By default, important regions are located at the top portion of the view. The saliency or category information is used to determine the importance of a group as follows:

- Saliency-based: groups are ranked according to the total saliency values of the transactions they contain;
- Similarity-based: groups are clustered and arranged according to category similarity.

To satisfy the above requirements, KnotLines heuristically performs the layout calculation using the greedy algorithm illustrated in Algorithm 1. Figure 7 (b) shows the compact layout derived from Figure 7 (a) using that algorithm.

The visualization in (Figure 7 (b)) is a scatter plot presenting the distribution of the sellers' transactions. Each transaction is visualized as a dot whose color encodes its sales category. The same color scheme is used in the bar chart (Figure 1 (c)).

The prototype was demonstrated to the analysts with a real dataset. The analysts pointed out that there were several major drawbacks in this design: (1) The visualization was seriously cluttered since a section may contain hundreds of transactions. The analysts suggested aggregating transactions within the same section; (2) The analysts agreed that the compact layout was necessary for high space efficiency. However, it was difficult for them to identify transactions made by the same seller from this view. Additional visual attributes were desired to emphasize this important relationship; (3) Important information about the transactions such as payment amount, whether a transaction had missing values, and if identical transactions occurred frequently, was not presented in this view.

Algorithm 1 Compact layout generation

Input: The group list $G = \{g_1, g_2, g_3, ..., g_N\}$, the associated importance measure $I = \{i_1, i_2, i_3, ..., i_N\}$. The row list $R = \{r_1, r_2, r_3, ..., r_N\}$ in the view.

```
1: Order the group list G according to the importance measure I,
   where the most important group is at the beginning of the list.
2: for Each group g_i in the sorted list G do
3:     j = 1;
4:     while j <= N do
5:         if k_i does not overlap with any placed groups in r_j then
6:             Place k_i in v_j;
7:             break;
8:         else
9:             j++;
10:        end if
11:    end while
12: end for
```

7.2 KnotLines

To address the above issues, we designed an enhanced visual metaphor call KnotLines. It is inspired by musical notation which can be regarded as an improved scatter plot that places different types of dots (notes) along the time axis. It is a complex visual representation of a time-series (e.g., beat and rhythm) and its connections.

All transactions in the same section are aggregated into a bigger dot whose size represents the number of transactions. It is analogous to a note in musical notation, and is called a **knot** (Figure 8 (a)). If any transactions in the section contain missing values in important attributes (e.g., the delivery location of a buyer is null), the knot is unfilled. Unfilled/filled is a pre-attentive visual attribute and is used here to draw the analysts' attention to transactions with missing values. In addition, if the frequency of transactions in a knot was abnormally high, the frequency was displayed using a tag.

Knots of all the sections in a sub-group are placed sequentially on the left side of a stem (Figure 8 (b)). The knots and the stem compose a visual symbol of a sub-group. This visual representation of a sub-group is called a **knotbunch**. The length of the vertical stem can be used to encode an attribute of the sub-group, and we use it to present the total payment amount of the sub-group since the payment amount is the most important transaction information. The horizontal position of a knotbunch is decided by the time interval of the sub-group.

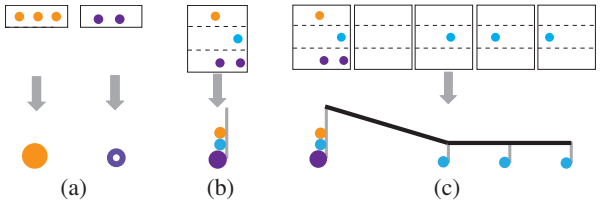

Fig. 8. Visual encoding (bottom) of the three-level organization (top) in the KnotLines view. (a) Two knots represent two sections. The unfilled knot indicates that the section contains empty values of attributes. (b) A knotbunch with multiple knots is used to encode a sub-group. (c) A knotline consisting of several knotbunches indicates a sequence of successive transactions of a seller.

To emphasize the relationships among the transactions made by the same seller, line segments are used to sequentially connect the tops of knotbunches of all sub-groups in a group along the time axis. The knotbunches and line segments connecting them form a **knotline** (see Figure 8 (c) for an example). Each knotline represents a group. If the group contains more than one sub-group, its line segments form a waving curve. The curve not only brings knotbunches of the same group together, but also reveal the vibration of the payment amount of the transactions of a seller along the time axis.

It should be noted that the visual encodings can be specified by the analysts subject to the analysis task. The analyst-specified visual encoding of this view is summarized in Table 2. Many knotlines form the KnotLines view. Figure 9 (a) depicts a KnotLines view with saliency-based layout.

Table 2. Visual Encoding for E-transaction Time-Series

Visual Encoding	Transaction Data
A knotline	Transactions from the same seller in different time (a group)
A knotbunch	Transactions from the same seller in a time interval (a sub-group)
The stem length	The total payment amount of transactions from the same seller in a time interval
A knot	Transactions from the same seller with the same sales category in a time interval (a section)
The knot color	The sales category of the knot
The knot size	The number of commodities in the knot
An unfilled knot	A transaction with abnormal seller or buyer locations

7.3 Visual Exploration

In addition to the specification of layout modes and the investigation of detailed knots, KnotLines provides a suite of interactions for analyzing multiple knotlines.

Saliency Modulation Each transaction shown in the KnotLines view contains a saliency value. The analysts can show either all transactions selected from the 2D TOS map or only the selected transactions whose saliency values are larger than a given threshold (e.g., 0.8). This filtering operation is useful when there are many knotlines shown in the view, so it supports T2 and T3. Figure 9 demonstrates the effect of saliency modulation.

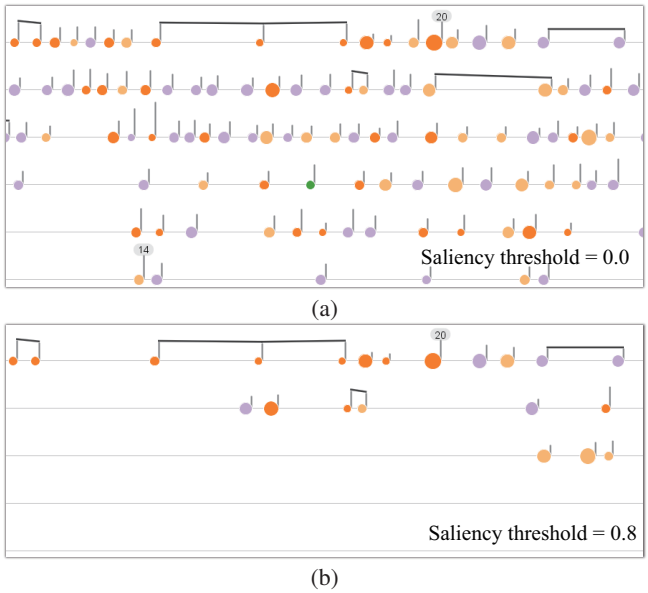

Fig. 9. Visualizing a set of transactions in the KnotLines view (a) without saliency modulation and (b) with saliency modulation.

View navigation The KnotLines view can be zoomed horizontally for clear illustration, a function helpful for T2 and T3. The lengths of the time intervals will be adjusted accordingly. Analysts can scroll vertically or horizontally to see more knotlines.

Knots of Interest Selection We can select a set of knots by clicking, or dragging using a lasso tool. When a knot is specified, it is highlighted with a yellow ring. Related knots which are made by the

same buyer are also highlighted with grey rings to draw the analysts' attention (Figure 1 (b)). A floating text box will appear displaying the detailed information of the selected knot such as the location of the seller, the payment amount, and the sales category. Analysts can check the information of the transactions (e.g., the location of the buyer and the seller, the sub-category, and the commodity number) in the selected knot (section) in the detail view (Figure 1 (e)), which is designed for T2. A statistic view (Figure 1 (f)), which is helpful for T4, is used to present statistical information for the selected knot, such as the trend of the transaction frequency and the payment amount of the seller.

Labeling When a specific transaction is identified as salient by analysts, it can be added to the labeled data for iterative visual analysis and exploration.

8 Case Study

An analyst from the data department in our collaborating customer-to-customer (C2C) retail business participated in this study. The company provided a dataset containing 26 million online E-transactions from which they would like to detect fake transactions. About 9.3 million sellers and buyers are involved in the dataset. He was interested in identifying when a seller accumulates credits by creating fake transactions with partner buyers. Some of the indicators of abnormal transaction behaviors can be an unusually large number of commodities, large variations of payment amount, frequent transactions between a specific seller and buyer, and attributes with values out of their normal range.

8.1 Construction of Decision Tree

Each E-transaction record includes the buyer and seller locations and transaction attributes such as payment amount, number of commodities, sales category, and transaction time. These attributes were used as basic features for each transaction. The transaction frequencies of a seller in each time interval were computed and used as a temporal feature. Key words and phrases in the comments were extracted (e.g. "credits") according to a sensitive dictionary provided by the analyst and used as textual features. We labeled the features of about 300 transactions, which were chosen from each category using stratified sampling. We trained the decision tree with the labeled data. The extracted feature descriptions of transactions to be analyzed were used as input to the decision tree and a saliency value was yielded for each of them. We evaluate the efficiency of the decision tree using the precision p and recall r: $p = TP/(TP+FP) = 0.89$, $r = TP/(TP+FN) = 0.92$, where TP, TN, FP, and FN were counted from the prediction results in the training data (see Table 1). The analyst agreed that the precision p and recall r were adequate for our dataset because the decision tree is only required to find potential salient transactions, which will be investigated in the subsequent process.

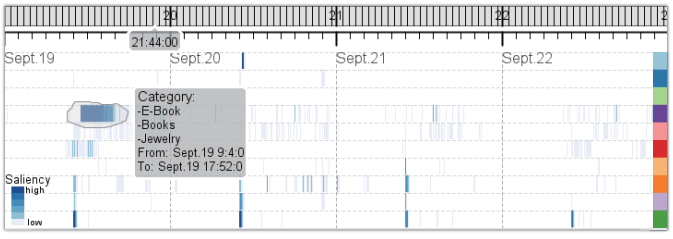

Fig. 10. TOS map of an E-transaction time-series. The selected region indicates highly salient transactions.

Below we show how VAET helped the analyst detect abnormal transactions. Section 8.2 and Section 8.3 show how VAET can support T1 - T4.

8.2 Abnormal Frequency and Locations of Transactions

The analyst started the exploration in the TOS map after a short training session. He noticed a long region with high saliency (Figure 10). Subsequently, the analyst specified the time window, and zoomed in to the desired region. To further study the transaction behaviors, the analyst selected this region and found that many transactions were categorized as "Books" at approximately 10 am on September 19.

The analyst noticed continuous red knots connected in a knotline in Figure 1 (b). He told us that this pattern indicated frequent transactions of a seller in the selected time interval. After checking the detailed information, the analyst found that these transactions belonged to the "Top-up Card" category and were made with different buyers. He commented that it might be a sales promotion event because he did not find any abnormal information in those transactions.

The analyst increased the saliency threshold to 0.8 with the saliency modulation slider. This allowed the analyst to effectively filter out many less salient transactions. The analyst immediately located multiple unfilled knots indicating transactions with missing values. He commented that using unfilled knots to present transactions with missing values was effective in drawing his attention. To further investigate whether these records indicated a sales promotion or fake transactions, the analyst conducted further analysis. When he clicked an unfilled knot in this knotline, many other knots in the same knotline were highlighted (Figure 1 (b)), indicating that most of the transactions in these knots are made by the same buyer and seller. By viewing the detailed information of the knots in this knotline (Figure 1 (e)), the analyst noticed that the delivery addresses of the buyers are null in the detailed information view. He commented that this was suspicious since the buyer would never get the commodity if he or she did not fill his or her address. By viewing the transaction history of the knot (Figure 1 (f)), the analyst found that the sales amount of the seller dramatically increases in a time interval.

The analyst concluded that these transactions are likely associated with earning credits. This conclusion was verified by analysts in the business intelligence department of the data providers who checked additional information related to the transactions, such as the IP addresses of the seller and buyers. They explained that the transactions were conducted by a group of buyers who helped the seller increase credits and the seller did not really deliver the products. The transactions in this knotline were labeled as salient by the analyst, and they were added to the training dataset.

8.3 Abnormal Attribute Values of Transactions

The analyst chose another time window in the TOS map. By checking the sales category information in the bar chart view (Figure 11 (a)), the analyst found that the total number of commodities sold in the "Electronics Accessories" category was much larger than those in other sales categories. The analyst believed that this would be a sales promotion due to its large commodity number. He investigated this hypothesis in the knotline view but didn't find any kontline containing frequent and continuous knots of "Electronics Accessories".

The analyst selected this category in the TOS map to filter out irrelevant categories. By carefully checking the remaining knots, the analyst found an extremely large knot with a short stem (Figure 11 (b)), indicating a large commodity number with a low total payment amount. The analyst told us that "through the pattern of this knotLine, I noticed the abnormal relationships between the payment and commodity number at the first sight of it". A detailed view of the knot (Figure 11 (c)) showed that this section contained a single transaction with a payment amount of only ten cents but a commodity of 220,000 units. By further investigating the transaction history of the seller (Figure 11 (d)), the analyst eliminated the probability of a sales promotion because the seller made few transactions over a period of time, indicating few goods were sold. The analyst identified it as an event where the seller attempted to increase their internet search ranking according to the number of commodities that had been sold. The analyst also labeled the transaction as salient and added it to the training dataset.

9 User Study

We conducted a user study to evaluate VAET's ability to support low level analysis tasks, namely T1 - T4 discussed in Section 3. The dataset used was the transaction dataset explored in Section 8.

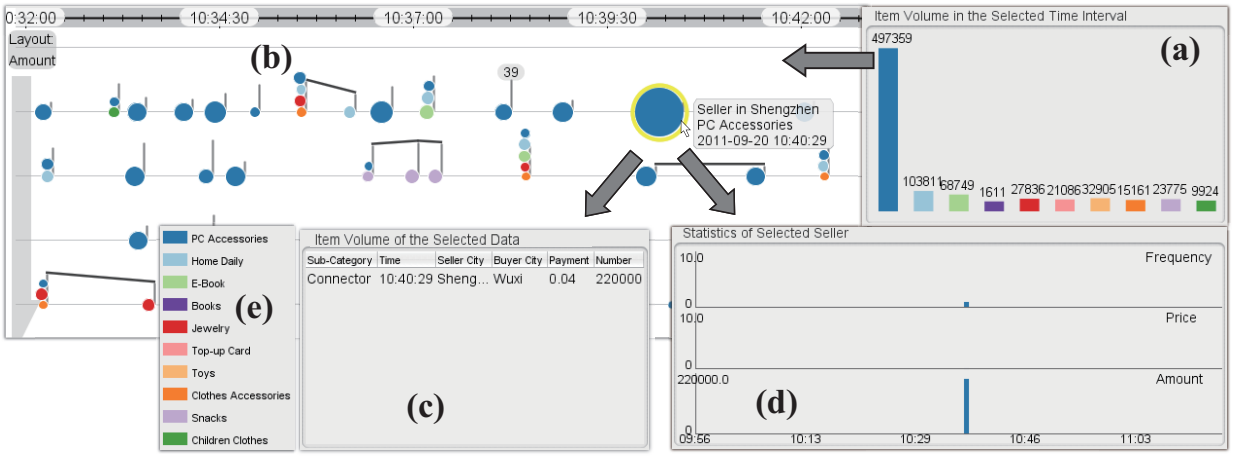

Fig. 11. A transaction with unusual attribute values. (a) The bar chart shows the number of commodities in different categories. (b) The big knot with a yellow ring indicates an unusually large number of commodities in a transaction. (c) Detailed information and (d) statistical information of the selected knot. (e) Sales category legend.

9.1 Design

Ten participants (6 males and 4 females) whose ages ranged from 21 to 35 took the user study. Two of them were analysts and the others were graduate students. All participants had prior experience using online commerce and were comfortable with computers. The majors of the students included computer science, design, math and biology. None of them had used VAET before.

9.1.1 Procedure

The participants took the study one by one. For each participant, a short training session was conducted before a test section. In the training section, the instructor first gave a 25 minute demo to the participant to introduce VAET. During the demo, the instructor explained the visual design and function of VAET. After the demo, the participant practiced the interactions provided in VAET with the help of the instructor for 5 minutes.

In the test section, the participants were asked to complete 9 exercises similar to those encountered during an actual analysis without instructor help. They were then asked to evaluate the system by answering a questionnaire and providing subjective feedback.

9.1.2 Exercises and Questionnaire

The 9 exercises were designed for three different specific situations, two of which were described in Section 8. Each low level task was evaluated by one or more exercises. The exercises and the tasks they evaluated were as follows (the tasks are shown in brackets):

E1 "Choose the sales category with the highest saliency value from 9 am to 10 am on September 21 using the TOS map." Objective: Identify time periods and sales categories of interest in the TOS map (T1).

E2 and E3 were from the case described in Section 8.3.

E2 "Choose the sales category with the largest commodity number." Objective: Interpret the bar chart and identify sales categories of interest (T1).

E3 "Find the transaction with the largest number of commodities in KnotLines." Objective: Interpret the visual encoding of a knot and identify transactions with interesting patterns in specific attributes (T2).

To finish E4 - E6, participants were asked to set the time between 18 pm and 19 pm on September 19, and choose "Top-up Card".

E4 "Find the seller (knotline) with the highest transaction frequency from the KnotLines view." Objective: Interpret the visual encoding of a knotline and identify sellers with interesting transaction patterns (T3).

E5 "In the KnotLines view, which transaction pattern of the seller does not occur? (a) Single transaction with a large number of commodities but a small payment amount. (b) Continuous transactions with low frequency. (c) Continuous transactions with high frequency and a small payment amount. (d) I don't know." Objective: Interpret the knotlines and identify interesting seller transaction patterns (T3).

E6 "Which is a feature of the seller transaction history of the knot in E4? (a) Continuous, frequent transactions. (b) Occasional transactions. (c) Sudden, frequent transactions." Objective: Examine the seller's behavior in the statistic view (T4).

E7 - E9 were the same case described in Section 8.2.

E7 "Find unfilled knots in the KnotLines view and report the buyer cities of them." Objective: Examine the attribute values of the transactions using the detailed information view (T2).

E8 "What is the main feature of the seller's transaction history of the knot identified in E7? (a) Continuous, frequent transactions. (b) Occasional transactions. (c) Sudden ,frequent transactions." Objective: Examine the seller's behavior using the statistic view (T4).

E9 "What is the transaction behavior of the seller identified in E7? (a) A single transaction with a large commodity number. (b) Frequent transactions with low payment amounts. (c) Frequent transactions with abnormal buyer cities." Objective: Interpret and examine the seller's transaction patterns from KnotLines (T4).

After conducting the exercises, the participants completed a questionnaire consisting of 6 questions (Q1 - Q6). They were asked to rate how easy it was to learn the system and the efficiency of VAET on a scale from 1 to 5 (1 = very easy or efficient, 5 = very hard or inefficient). The questions also gathered subjective feedback. The six questions were as follows:

Q1 Is it easy or hard to learn the TOS map?
Q2 Is it efficient or not to explore salient data with the TOS map?
Q3 Is it easy or hard to interpret the visual encoding of a single knot?
Q4 Is it easy or hard to interpret the visual encoding and layout of KnotLines?
Q5 Is it efficient or not to analyze the user transaction patterns with KnotLines?
Q6 Is it easy or hard to analyze multi-user behavior with VAET as a whole?

9.2 Results

The accuracy and time of the 9 exercises were collected for evaluation.

9.2.1 Accuracy and Speed

Overall, the participants completed the exercises, yielding 5 mistakes out of the 90 total exercises (94.4% accuracy). The analysts answered all questions correctly. As for the student participants, two of them erred on E5, two erred on E8 and one erred on E9. We interviewed the participants who erred on E5. They both said that they "only noticed the main patterns of the knotlines and ignored the patterns with fewer occurrence". However, they had no problem interpreting the user behavior from them. The participants who answered E8 incorrectly mentioned that they forgot to check the seller history information shown in the statistic view (Figure 1(f)) and answered the question based on the KnotLines view instead. The participant who erred on E9 thought that null buyer cities are normal for transactions of virtual commodities such as E-books, which do not need delivery addresses. In fact, all commodities in the "Books" category are real items. Although some participants had problems with T3 and T4, the overall accuracy indicates that VAET supports the tasks well.

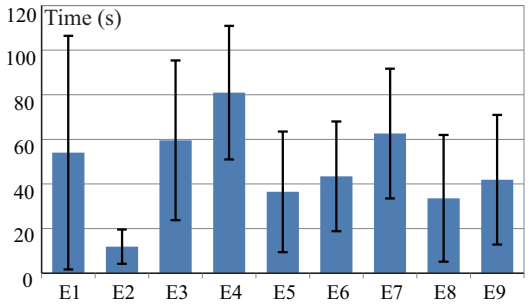

Fig. 12. Time spent for each exercise.

Figure 12 shows the average and standard deviation of completion time for each exercise. The time for E3, E4 and E7 were longer than those of other questions. These three exercises asked the participants to search for knotlines with specific characteristics which may require more time to examine the view in detail. The participants were able to finish the other 6 questions quickly.

The completion time ranged from 5.14s to 44.22s. The analysts spent much less time than the student participants. Overall, VAET allowed most of the participants to finish complex exercises and tasks (such as E4 and T3) in 90 seconds. This was quite fast when considering various attributes of a transaction and the relations among them.

9.2.2 Questionnaire

The ratings by the participants are shown in Figure 13. Most participants thought that the TOS map and the knotlines were intuitive and helpful as most of them rated the system as a 1 in Q1 and Q5. One participant gave a high rating (4) in Q6 and provided suggestions on improving the usability of VAET. He mentioned that "the small notes in the KnotLines view were hard to click when checking the detailed information" and that there should be more tips for the different views. The average ratings for all questions were between 1.20 - 2.30, indicating most participants found VAET easy to learn and efficient for analysis.

9.2.3 Analyst Feedback and Discussion

We interviewed the analysts and other participants at the end of the user study. Both analysts were satisfied with the result of classification. They commented that the integration of feature extraction and decision trees was efficient and flexible when analyzing different types of transactions. We asked them to compare the decision tree with logistic regression which they used previously. They thought our model had several advantages: 1. Decision tree is simple and easily understood for the analyst. 2. Decision tree is more powerful when dealing with missing values of attributes than logistic regression. 3. Classification using decision tree is faster than logistic regression. He thought that we could improve our method by using more information such as the IP addresses of the buyers and sellers in the future.

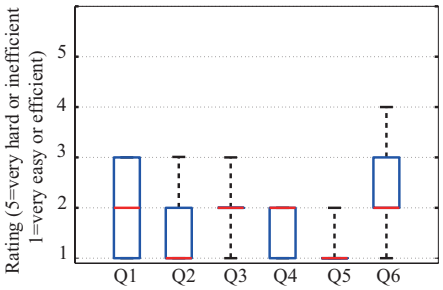

Fig. 13. Questionnaire results for each of the 6 questions (Q1-Q6) answered on a 5-point scale by the participants. The bottom and top of the blue box are the first and third quartiles of the ratings, and the red band indicates the median. The ends of the whiskers represent the minimum and maximum of the ratings.

They commented that the TOS map was intuitive and the exploration was convenient. One analyst stated that "The visualization of multi-user behaviors by KnotLines is creative and vivid." Although one analyst thought that the system was difficult to learn, few of the participants had difficulty interpreting the visualization and most were able to find interesting transactions after a brief training session.

Both analysts mentioned that VAET was able to help them explore undiscovered transaction patterns. This is crucial because "people who cheat always keep changing their measures to trick". They thought that VAET had the capacity to find new and emerging transaction patterns and to help them refine data models. The analysts were eager to use VAET for practical multi-user applications where they discovered contextual and temporal correlations among many high dimensional transactions.

Interestingly, some participants initially assumed that the usage of our system requires basic musical knowledge. Other participants mentioned that their knowledge of music notes affected the comprehension of the design. For example, the stem length is fixed in music notes but it is variable in our design. In addition, some people thought a single knotbunch not connected with other knotbunches (similar to a quarter note) lasts shorter than a connected knotbunch (similar to an eighth note). The fact is that the transactions have no duration because they are all made online instantaneously. The participants told us that after they learned how to interpret the design, this difference did not hinder their analysis.

10 Conclusion

This paper presents a novel visual exploration scheme for identifying elementary transaction data and studying the temporal or collective behavior from large pieces of fragmented records. Prior to detailed exploration and reasoning of the chosen transactions with KnotLines, a decision tree algorithm and a filtering process by TOS map are performed to choose potentially interesting transactions from a huge amount of records. The case study and the user study verify that VAET can effectively support most of the tasks. According to the result, some tasks such as T3 need to be better addressed as the patterns of transactions can be dynamic and multi-level. To make VAET easier to learn and use, we would like to make the design of TOS map and KnotLines more intuitive. For future work, we also would like to extend our approach for more datasets.

Acknowledgments

This work was supported in part by the Major Program of National Natural Science Foundation of China (61232012), the Google Faculty Research Grant, the National Natural Science Foundation of China (61202279), the National High Technology Research and Development Program of China (2012AA12090), the Zhejiang Provincial Natural Science Foundation of China (LR13F020001), the Doctoral Fund of Ministry of Education of China (20120101110134), the Fundamental Research Funds for the Central Universities, and the NUS-ZJU SeSama center.

REFERENCES

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Human-Computer Interaction. Springer Verlag, 1st edition, 2011.

[2] I. V. Cadez, P. Smyth, and H. Mannila. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *ACM SIGKDD*, pages 37–46, 2001.

[3] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *IEEE VAST 2007*, pages 155–162. IEEE, 2007.

[4] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE TVCG*, 12(6):1461–1474, 2006.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1), Nov. 2009.

[6] M. Hao, D. A. Keim, U. Dayal, D. Oelke, and C. Tremblay. Density displays for data stream monitoring. In *Computer Graphics Forum*, pages 895–902, 2008.

[7] M. C. Hao, U. Dayal, M. Hsu, T. Sprenger, and M. H. Gross. *Visualization of directed associations in e-commerce transaction data*. Springer, 2001.

[8] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[9] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE TVCG*, 16(6):927–934, 2010.

[10] D. F. Jerding and J. T. Stasko. The information mural: A technique for displaying and navigating large information spaces. *IEEE TVCG*, 4(3):257–271, 1998.

[11] Z. Liu, J. Stasko, and T. Sullivan. Selltrend: Inter-attribute visual analysis of temporal transaction data. *IEEE TVCG*, 15(6):1025–1032, 2009.

[12] E. Manavoglu, D. Pavlov, and C. L. Giles. Probabilistic user behavior models. In *IEEE International Conference on Data Mining*, pages 203–210, 2003.

[13] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. In *ACM SIGCHI conference on Human factors in computing systems*, pages 1483–1492, 2008.

[14] W. Müller and H. Schumann. Visualization for modeling and simulation: visualization methods for time-dependent data - an overview. In *Conference on Winter simulation: driving innovation*, pages 737–745, 2003.

[15] M. Ogawa and K.-L. Ma. code_swarm: A design study in organic software visualization. *IEEE TVCG*, 15(6):1097–1104, 2009.

[16] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: using visualization to enhance navigation and analysis of patient records. In *AMIA Symposium*, page 76, 1998.

[17] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52(3):199–215, 2003.

[18] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[19] A. Rajaraman and J. Ullman. *Mining of Massive Dataset*. 2012.

[20] A. Sallaberry, C. Muelder, and K.-L. Ma. Clustering, visualizing, and navigating for large dynamic graphs. In *Proceedings of Graph Drawing*, Sepetember 2012.

[21] T. Schreck, T. Tekušová, J. Kohlhammer, and D. Fellner. Trajectory-based visual analysis of large financial time series data. *ACM SIGKDD Explorations Newsletter*, 9(2):30–37, 2007.

[22] Y. Tanahashi and K.-L. Ma. Design considerations for optimizing storyline visualizations. *IEEE TVCG*, 2012.

[23] E. R. Tufte. *Beautiful evidence*, volume 1. Graphics Press Cheshire, CT, 2006.

[24] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *ACM SIGCHI conference on Human factors in computing systems*, pages 575–582, 2004.

[25] F. Vigas, D. Nguyen, J. Potter, and J. Donath. Digital artifacts for remembering and storytelling: Posthistory and social network fragments. In *HICSS-37*, 2004.

[26] J. Wei, Z. Shen, N. Sundaresan, and K.-L. Ma. Visual cluster exploration of web clickstream data. In *IEEE VAST 2012*, Oct. 2012.

[27] R. Xiong and J. Donath. Peoplegarden: Creating data portraits for users. In *ACM symposium on User interface software and technology*, pages 37–44, 1999.