A Multiple-Phased Modeling Method to Identify Potential Fraudsters in Online Auctions

Wen-Hsi Chang/TamKang University
Graduate Institute of Management Sciences
Taipei, Taiwan
e-mail: wenhsi.chang@gmail.com

Jau-Shien Chang/TamKang Univeristy
Department of Information Management
Taipei, Taiwan
e-mail: jschang@mail.im.tku.edu.tw

Abstract—**Loopholes in online auction sites enabled fraudsters to easily hide themselves. To reduce the odds of being defrauded, online auction traders usually use reputation systems for estimating a trading partner's credit. However, reported dollar losses of online auction fraud have hit recorded height for years that implies existing reputation systems may not prevent fraud effectively as expected. To reduce the risk of being defrauded, an ideal fraud detection mechanism should be not only to identify current fraudster but also potential ones. Therefore, this paper proposes a multiple-phased modeling method integrating with decision trees for enhancing the capability of fraud detection. To demonstrate the effectiveness of the proposed method, real transaction data were collected from Yahoo!Taiwan for training and testing. The experimental results show that the recall rate of identifying a potential fraudster before transitioning into his criminal phase was up to 86%.**

Keywords-fraud detection; online auction; decision tree; e-commerce

I. INTRODUCTION

Convenience of online auctions not only satisfied online traders but also created massive cash turnover annually in the past several years. However, this same feature caused convenient loopholes for fraudsters. The reported dollar losses in online auctions have been top 2 types of Internet frauds for years, according to the statistics of the annual report of NW3C in recent 5 years [1].

Most trading participants usually estimate the reputation of a trading partner by his feedback score in reputation systems from trading partners. The score also is explained as credit level of an account. Originally, the score of a trader is used to estimate the risk of fraud and determine whether enter a deal or not. Therefore, some schemed fraudsters always inflate the score for deceiving naive victims. As a result, a high reputation score could become a trap as well for enticing target victims under certain circumstances.

Online auction fraud detection could be treated as the problem of anomaly detection [2]. The difficulty of anomaly detection on malicious actions is that most members of auction sites are legitimate accounts and only very small number of fraudsters in real world [3]. On the other hand, fraudsters always mimic regular behavior as camouflage that makes malicious intention not to appear apparently [4]. Fabricating transaction histories is one of common schemes for increasing credit. After target victims appearing, they always offer expensive items for swindling immediately.

Much previous research on online auction fraud detection has been proven effective in discovering fraudsters after a victim appears. Once a fraudster has activated a fraud, there are many obvious visual features exhibited of which most could be recognized more easily. However, prior to malicious actions were executed, the fraudster had to behave as normal as possible for disguising his intention.

To reduce the probability of being the first victim of a schemed fraudster, the capability of identifying potential fraudsters during the latency period should be improved. Therefore, this paper proposes a method to extract features in the earlier phases, before a fraud activated, for modeling latent behavior of fraudsters.

In order to demonstrate the effectiveness of the proposed methods, real transaction data were collected from Yahoo!Taiwan for training and testing. The experimental results demonstrate that the recall rate of fraudster identification is around 86%, and F-measure reaches 85%.

The rest of this article is structured as follows: Section 2 discusses decision trees, fraud detection and feature measuring attributes in previous research. The third section discusses the behavior of the latency period. Section 4 describes how to build potential fraudster detection models using decision trees. Section 5 presents the experimental results. Finally, conclusions and future research directions are presented in the last section.

II. RELATED WORKS

This section is to discuss the fraud detection methods by decision trees, the general characteristics of online auction fraudsters and featured measuring attributes.

A. Fraud Detection and Decision Trees

Decision trees finds extensive used in a wide variety of applications. Much of the previous work on fraud detection applying decision trees has achieved high performance on a

large of domain, such as credit cards, customs claims, insurance, health care and online auctions etc. [2][5][6][7].

Quinlan proposed C4.5 algorithm for handling both continuous and discrete attributes. C4.5 examined the normalized Information Gain Ratio, which is difference in entropy, theoretically instead of Information Gain [8]. There are many successful empirical cases of applying C4.5 algorithm for inducing decision trees, such as Donoho used C4.5 on early detection of insider trading in option market before news break out, of which scenarios are similar to potential fraudster detection [9]. In addition, decision trees outperform than other algorithm for converting into if-then rules with less computation complexity [10].

In this study, a C4.5 revised version named as J48 was employed, which is one of Weka3.6.0 classifiers [11]. To boost the performance of decision trees, AdaBoostM1 algorithm [12][13] was applied using a major voting for weighted majority to solve the conflicts in the result of classification. However, the criterion of selecting which attributes is one of critical component for building decision trees in real applications [14]. Hence, the measuring attributes we adopted will be discussed in next section.

B. Measuring Attributes in Online Auctions

From our observations, a large amount of monetary loss case usually is not a coincidence but a consequence of schemed preparation. Whether a concluded trade is normal or not, it is inevitable for any participants to leave some traits in transaction histories. From this perspective, every case of fraud also leave some fraudulent traits in transaction histories, such as concluding prices and purchased items that imply certain features of trading records occurred period being remained prior to a fraud being activated. Therefore, capturing the features before a fraud being activated is necessary to discover potential fraudsters. Furthermore, the methods of fabricating reputation score schemed swindlers, who used to appear legitimate, apply might be traced in transaction histories, such as selling cheap items, changing ID and changing location, etc. [15].

In order to identify a potential fraudster, the latent behavior models are necessary for being constructed. In the real world, the features of activating a fraud are identified easily even without any assistance. On contrary, the features in the latency are obscure since a fraudster disguised with similar normal actions. That make latent fraudulent behavior modeling is more difficult.

We consider the problem of fraudster detection in online auctions is a kind of anomaly detection, because the fraudulent behavioral features don't appear on the majority of members. Therefore, measuring attributes in modeling could impact the efficacy of a fraud detection model, much previous research proposed different measuring attributes for capturing the more sophisticated characteristics of a fraud. Wang used Boolean values for denoting particular status, such as being a shop owner [16].

Some researchers apply the values of K-core in transaction networks that are helpful in detecting reputation inflation [16][17][18]. The magnitude of price changes and the trends of purchase of fraudsters could be the features of a fraud. Therefore, Chau et al. devised 17 variables that use median and standard deviation of the concluding prices regarding transaction history information in order to observe abnormal behaviors, as well as the number of commodities, and the ratio of selling to all transactions. [6][7][19] (See Table 1). In addition, some particular numerical value might reveal something meaningful could be used as additional features, such starting bid, etc.

TABLE I. CHAU'S FEATURES

Features Description
Median prices of items sold within the first 15, first 30, last 30, and last 15 days
Median prices of items bought within the first 15, first 30, last 30, and last 15 days
Standard deviation of the prices of items sold within the first 15, first 30, last 30, and last 15 days
Standard deviation of the prices of items bought within the first 15, first 30, last 30, and last 15 days
Ratio of the number of items bought to that of all transactions

III. BEHAVIOR IN THE LATENCY PERIOD

Most of online auction frauds involving a large amount of money are the results of deliberate premeditation, which was mentioned in section 2. In addition, only the transaction history of a schemed fraudster apparently presents behavior changes after a fraud activated. In general, the entire lifespan of a schemed fraudster could be divided into the latency period and the execution period. In contrast, the transaction history of a legitimate user usually remains consistent, regardless of its length lifespan.

The main difficulty of discovering a potential fraudster is that they always disguise behavior as normal legitimate accounts until a victim appears. In order to detect the latent behavior of potential fraudsters, a model for detecting behavior in the latency period has to be constructed. Hence, first and foremost, identifying the end of the latency period of a fraudster, at which is the point of a fraud occurred, is to extract behavior features before the fraudster defrauding.

Given the complete transaction history of a proven fraudster, which equals the entire lifespan, is presented by phase 100% where the fraudster left the auction site or was suspended. Referring to Fig. 1, 80% of the transaction history doesn't comprise the time of a fraud occurrence; the 85% of the transaction history contain a little behavior in the execution period. The phase 80% behavior is complete latent behavior.

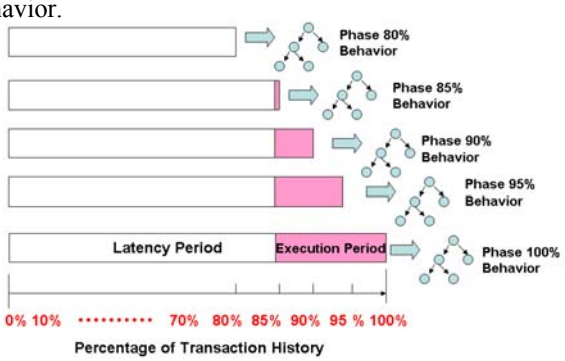

Figure 1. Behavior changes in the entire lifespan of a fraudster

To describe the latency period, a unit for partitioning transaction histories has to be determined. The simplest unit is the count of accumulated ratings. For example, if a fraudster obtaining 100 ratings were enlisted at the blacklist, then his phase 80% period would be at the time he got 80th ratings.

The latency period of a fraudster consists of planning and preparing work for swindles, and the execution period focuses on targeting victims. Therefore, there is a precise demarcation between the behavioral features of two different periods. However, the boundary of two period was determined by circumstances and the fraudster self. Any two fraudsters is very difficulty to activate a fraud at the same time.

Referring to Fig. 2, fraudster *A* activated a fraud at the starting point of phase 90%, fraudster *B* activated a fraud at the beginning of phase 80%, and fraudster *C* activated at phase 85%. Therefore, the Phased 85% behavior to fraudster *A* is the behavior being in the execution period, and to fraudster *B* being in the latency period. In addition, fraudster C enters his execution period at the phase 85%, but fraudster *A* and *B* activates their frauds at different phases. Unfortunately, each fraudster enters his execution period under a different situation that makes potential fraudster detection more difficult.

Fig. 2 also shows that the phased 85% model comprises of all features occurred in 85% of transaction history of each account. Realistically, we cannot ensure that the model contains the phased 85% latent behavior only, according to different fraud activated point for each fraudster. It is difficulty to identify the exact point of demarcation for the two periods. However, the earlier phased models contain much latent behavioral features at least.

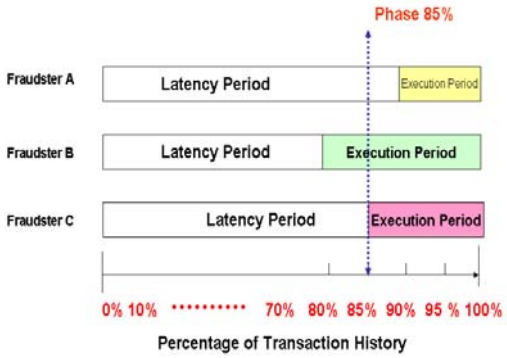

Figure 2 Transaction histories of different fraudsters

IV. Potential Fraudster Detection Models Construction

Building a potential fraudster detection model on complete history of a fraudster's accumulated ratings can represent the features of a fraud which has occurred, however through the proposed phased model, known behaviors of fraudsters in different stages are incorporated for predicting the legitimacy of users before a fraud occurs. In general, a set of measuring factors referring to extracted features of suspended members from the blacklists of auction sites to identify potential fraudsters.

A. Multiple-Phased Modeling

To keep the ratio of legitimate accounts to fraudsters at 2:1 in the training sets of each phased profiles for learning the capability of identification. Fig. 3 shows how to construct a single phased model.

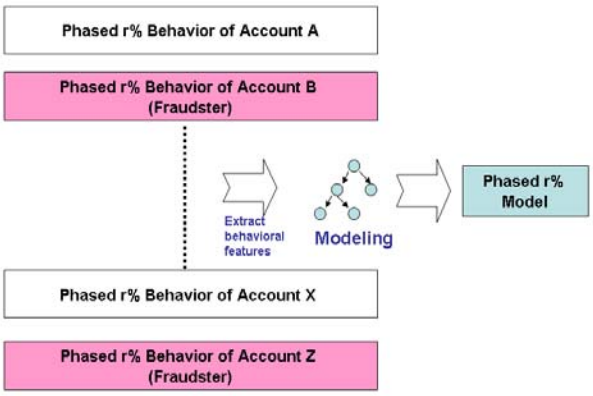

Figure 3 A single phased model construction.

To identify potential fraudsters, choosing an appropriate number of phased models for constructing a detection procedure is necessary. Therefore, an ideal potential fraudster detection mechanism could consist of a set of single-phased models (See Fig. 4).

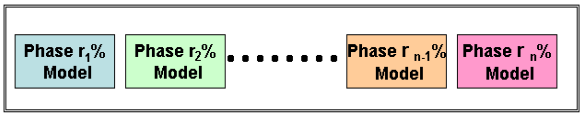

Figure 4 An ideal multiple-phased detection mechanism

Based on the previous concept, we demonstrate the ideal potential fraudster detection procedure in pseudo code (See List 1).

List 1 Identify Potential Fraudster Procedure

```
global variables
    Model[]: array of M(r%)
// === build M(r%) according to test  set TS ===

    procedure buildPhasedModels
    (lb: a real number labeling the lower bound of r%,
     ub: a real number labeling the upper bound of r%,
     gap: the increment from lb to ub,
     TS: the data set for building models,)

      n = (ub-lb)/gap ;
        for i = 0 to n
                Model[i]=buildModel(lb+gap*i, TS, ;
        end procedure
```

```
//=== determine whether U is a fraudster or not ==

    function detectFruad(U: the account under test,
Models: models for fraud detection)
          L = length of Models ;
          for i = L-1 to 0
          if (Models[i].testFruad(U) == true) return i ;
          return -1 ;
    end function
// === determine whether account U is a fraudster or not
```

The buildPhasedModel() procedure is to build different phased models. For instance, calling buildPhasedModel(*0.85, 1.0, 0.03, TS*) results in the system importing a designated test set, TS to build 5 phased models that include M(*85%*), M(*88%*), M(*91%*), M(*94%),* M(*97%)* and M(*100%*). And then store the 6 models in the global array respectively, such as Model[].

The function of detectFraud(*U, Models*) inspects all models in the Models[] array. If an account *U* matches one of models in Models[i], the system will return a model number. Otherwise, it will return back –1 to denote its innocence. All inspections are performed with the model number in descending order. Hence, the account *U* will be checked with M (100%), M (97%), M (94%) and so on. Unfortunately, it is impossible to have zero misclassification with the phased models. The procedure is helpful in refining suspicious accounts at the first stage.

B. Potential Fraudster Detection Model Construction by Multiple Phased Features

From our observations, there exists a common situation that the features of one fraudster in phase r_1% might be similar to the features of another fraudster in phase r_2%, where $r_1 \neq r_2$. For example, the phase 94% behavior of fraudster A is similar to the phase 80% behavior of fraudster B. In practice, the problem of overlapped phased features affects the previous potential fraudster detection model procedure. To solve this problem above, we extracted the features of each account from more different partitioned phases that were fed into decision trees altogether for reducing the probability of misclassification.

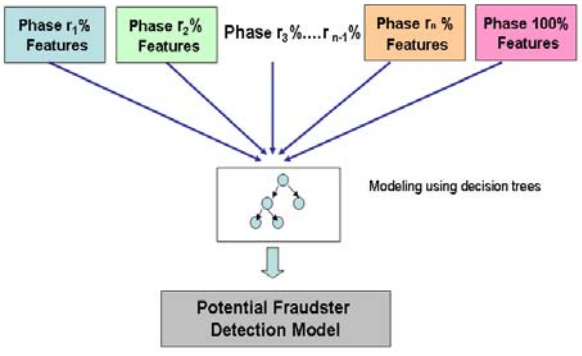

Figure 5 Constructing the potential fraudster detection model

Fig. 5 shows the potential fraudster detection model we proposed in this study. In addition, the features of different phases could be induced into rules by decision trees. The induced rules are helpful in explaining the differences between normal behavior and fraudulent behavior.

V. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the phased modeling approach, F-measure was used as the overall performance metrics for comparison and evaluation as follows:

$$F\text{-}Measure = (2 \times Recall \times Precision)/(Recall + Precision)$$

To prepare a data set for testing, we collected the transaction histories of 236 proved fraudsters from the blacklist of Yahoo!Taiwan. To keep the ratio of legitimate accounts to fraudsters at 2:1 in the training sets, we randomly sampled 300 legitimate accounts, and 150 from the 236 fraudsters. Then, we selected 75 from the remaining fraudsters and 150 legitimate accounts.

In order to extract the latent features of potential fraudsters, the transaction history of each account was partitioned into 5 different phases in this study, such as phase 80%, 85%, 90%, 95% and 100%. Each phase of an account is described by the Chau's 17 measuring attributes. So that, a training set consists of 2,250 ((300+150)*5) phased behavior profiles, and there are 935 ((150+75)*5) phased behavior profiles in a test set.

Table2 demonstrates the experimental results of potential fraudster identification using C4.5 algorithm by J4.8 classifier of Weka 3.6.0. In addition, each field in the Table 2 below is the averaged result of 10 trials. To optimize the accuracy of identification, a meta-learner applying AdaBoostM1 was adopted for enhancing the performance of J4.8 in this study. Referring to the row 3 of Table 2, the true positive rate of legitimate account identification was significantly improved to 92%. The F-Measure of fraudster identification is 85%. The results indicate that the overall performance is acceptable.

TABLE II. RESULTS OF IDENTIFYING POTENTIAL FRAUDSTERS

Boosted	TP rate	FP rate	Precision	Recall	F-Measure	Class
No	0.86	0.15	0.92	0.86	0.89	Legitimate
	0.85	0.14	0.75	0.85	0.8	Fraudster
Yes	0.92	0.14	0.93	0.92	0.92	Legitimate
	0.86	0.08	0.85	0.86	0.85	Fraudster

* Yes denotes the results of applying AdaBoostM1

* TP stands for true positive; FP stands for false positivve

VI. CONCLUSIONS AND FUTURE WORK

The fundamental concept of this study is to discover fraudsters them before they defraud, therefore the proposed method is not only to identify current fraudsters but also potential fraudsters who was preparing for swindling. The

experimental results show that the recall rate of detecting latent fraudsters is 86% and also present the practicality in online auctions. However, even the experimental result are quite encouraging, the proposed approach needs further improvement to satisfy new and upcoming types of fraudulent behaviors.

In this study, decision trees are top-down solutions in which most rules are generated by the extracted features of the majority of instances, which are proven legitimate accounts. On the contrary, fraudsters are far too few to influence how the trees grow, so that the predictive models generated by the decision trees detect the minority of fraudsters, of which results are not as accurate as the majority of instances. We are considering different lazy learners instead of decision trees for compensating the weakness in our future work.

In previous experiments, we did not take other contextual attributes such as the frequency and sequences of fraudulent behaviors into consideration. Most measuring features in this study only reflect part of irregular behavior. For example, if a trader successfully obtained many positive ratings within a short period of time, he could be inferred as a suspect of artificially raising feedback scores, which are not described by the previous measuring attributes directly. For further improvement, we are going to design the other set of measuring attributes for capturing contextual features.

In addition, the proposed method in this study focuses on abnormal preparatory behavior during the latency period, so that the earlier part of transaction history of a typical identity thief that being stolen from a legitimate account makes the features of the latency period incorrect. The weakness of the method will be improved in our future work.

REFERENCES

[1] Internet Fraud Complaint Center, "2008 Internet Crime Report – January 1- December 31," National White Collar Crime and the Federal Bureau Investigation, April 2009, http://www.ic3.gov/media/annualreport/2008_IC3Report.pdf

[2] V. Chandola, A. Banerjee and Kumar, V. 2009. "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 1-58, July 2009, doi: http://doi.acm.org/10.1145/1541880.1541882

[3] F. Angiulli, F. Fassetti, and L. Palopoli. "Detecting outlying properties of exceptional objects, " ACM Trans. Database Syst, vol. 34, no. 1, pp. 1-62, April 2009, doi: http://doi.acm.org/10.1145/1508857.15088

[4] S. Virdhagriswaran and G. Dakin, "Camouflaged fraud detection in domains with complex relationships," In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, pp. 941-947, doi: http://doi.acm.org/10.1145/1150402.1150532

[5] H. Shao, H. Zhao and G. Chang, "Applying Data Mining to Detect Fraud Behavior in Customs Declaration," in Proceedings of the First International Conference on Machine Learning and Cybernetics, pp. 1241-1244, Nov. 2002.

[6] D. H. Chauand C. Faloutsos, "Fraud Detection in Electronic Auction," in Pr oceedings of European Web Mining Forum (EWMF 2005) at ECML/PKDD, Oct. 3-7, 2005.

[7] D. H. Chau, S. Pandit and C. Faloutsos, "Detecting Fraudulent Personalities in Networks of Online Auctioneers," in Proceedings of PKDD 2006(LNAI 4213), pp. 103-114, Sep. 18-22, 2006.

[8] J. R. Quinlan, "C4.5:Programs for machine learning," San Mateo CA: Morgan Kaufmann, 1993

[9] S. Donoho, "Early detection of insider trading in option markets," In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Seattle, WA, USA, August 22 - 25, 2004). KDD '04. ACM, New York, NY, pp. 420-429, August 2004, doi: http://doi.acm.org/10.1145/1014052.1014100

[10] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data Mining Techniques for the Detection of Fraudulent Financial Statements," Expert Systems with Applications, vol. 32, issue 4, pp. 995-1003, 2007

[11] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques," San Francisco: Morgan Kaufmann, pp. 373-377, 2005

[12] D. Mease, A. J. Wyner and A. Buja, "Boosted Classification Trees and Class Probability/Quantile Estimation," J. Mach. Learn. Res. 8, pp. 409-439, 2007

[13] J. Gehrke, V. Ganti, R. Ramakrishnan and W. Loh, "BOAT—optimistic decision tree construction," In Proceedings of the 1999 ACM SIGMOD international Conference on Management of Data (Philadelphia, Pennsylvania, United States, May 31 - June 03, 1999). SIGMOD '99. ACM, New York, NY, pp. 169-180, doi: http://doi.acm.org/10.1145/304182.304197

[14] Y. Huang and V. T Hoa, "General criteria on building decision trees for data classification," In Proceedings of the 2nd international Conference on interaction Sciences: information Technology, Culture and Human (Seoul, Korea, November 24 - 26, 2009). ICIS '09, vol. 403. ACM, New York, NY, pp. 649-654. doi: http://doi.acm.org/10.1145/1655925.1656042

[15] B. Gavish and C. L. Tucci, "Reducing internet auction fraud," Commun. ACM 51, 5 (May. 2008), pp. 89-97. doi: http://doi.acm.org/10.1145/1342327.1342343

[16] J. Wang and C. Q. Chiu, "Detecting Online Auction Inflated-Reputation Behaviors using Social Network Analysis," in NAACSOS Conference 2005 Proceedings, June 26-28, 2005.

[17] Y. Ku, Y. Chen and C. Chiu, "A Proposed Data Mining Approach for Internet Auction Fraud Detection," Lecture Notes in Computer Science 4430 Springer, pp. 238-243, April 11-12, 2007.

[18] M. Kobayahi and T. Ito, "An Approach to Implement A Trading Network Visualization System for Internet Auctions," in Proceedings of the Second International Conference on Knowledge, Information and Creativity Support Systems, Ishikawa, Japan: (KICSS 2007), Nov. 5-7, 2007.

[19] S. Pandit, D. H. Chau, S. Wang and C. Faloutsos, "Netprobe: a fast and scalable system for fraud detection in online auction networks," In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 2007