Scalable Attack Analysis of Business Process based on Decision Mining Classification

Dewi Rahmawati[1], Riyanarto Sarno[2]
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dewi16@mhs.if.its.ac.id[1], riyanarto@if.its.ac.id[2]

***Abstract* — Banking crime is one of the widespread phenomena in 2016 are closely associated with the used of computer-based technology and internet networks that constantly evolving. One of them is the burglary of customer accounts through the internet banking facility. To overcome this, we need a method of how to detect a conspiracy of bank burglary case of customer accounts. The way to scalable is by get a mining decision to get a decision tree and from the decision tree to get a decision attribute value to determine the level of anomalies. Then of all the attributes decision point is calculated rate of fraud. The rate of fraud is classified through level of security of attack by the attacker then entropy gain is used to calculate the relative effort between the level of attacks in the decision tree. The results show that the method could classify three levels of attacks and the corresponding entropy gains. The paper uses decision trees algorithm, alpha++ and dotted chart analysis to analyze an attack that can be scalable. The results of the analysis show that the accuracy achieved by 0.87%.**

Keywords— security; bank; event logs; business process; fraud; dotted chart analysis; process mining; scalable.

I. INTRODUCTION

Cyber criminals can compromise networked systems by exploiting vulnerabilities, and such events impose a critical socioeconomic impact on enterprises and individuals [1]. An attack surface describes vulnerabilities that cyber criminals can exploit to penetrate through the networked system and so it is of paramount importance to secure the networked system by minimizing the attack surface (e.g., patching vulnerabilities). Security models, or also known as attack representation models (ARMs), are well-defined means of analyzing the security of networked systems in efforts to enhance the fundamental framework for network security [2]. These models can be used to analyze vulnerabilities in the networked system, and provide solutions to effectively manage them (e.g., network hardening). However, analyzing all possible attack paths using single-layered graph-based ARMs has a scalability problem (e.g., an attack graph (AG). This is an emerging problem as network systems are becoming large, such as the Cloud [3].

II. RESEARCH METHOD

A. Conspiration

Conspiration is the activity of secretly planning with other people to do something bad or illegal. A collaboration between people to do criminal or legal action but being forbidden to do a variety of actors. There are laws to deal with conspiracy in federal court cooperation only in violation of the law before the law. Conspiracy is a crime that is separate from the criminal act. Conspiracy is the preparation of criminal acts. Conspiracy form of solicitation is being made to engage in a criminal act. Conspiracy need an agreement between two or more people, while the solicitation can be done by one person alone [4]. Conspiracies can be called efforts. Efforts can be made by one person. The conspiracy existed before the crime actually carried out, no effort will succeed unless their efforts made. conspiracy in law is a separate substantive crime because when the person did criminal cooperation, the potential for increased criminal activity [5]. Therefore act with criminal intent agreement (in this case the real behavior) is considered quite dangerous. Conspiration is shown in Fig. 1. Colour label Green: describe dangerous person or mafia from cheater, Colour label Blue: describe people as a source of criminal data, Colour label pink: describe the intermediary cheater and Colour label orange: describe the most influential cheater.

B. Defraud

Encyclopedia of American Law explain that defraud is to make the wrong assumption of facts, making false or recklessly without regard to whether it is right or wrong, intending to make something be wrong and in a state in which the person is not responsible for its destruction [6]. Defraud is intended to deprive the property or benefit of any person in any case, real, or equal to fraud. Her intent means intent to deceive others, and to induce others like, dependence on such scams, consider, create, transfer, modify, or terminate the rights, obligations or powers with reference to the property [7].

C. Event Logs

The event log is a recording process in the form of transaction history or audit trail in an information system tools. Each system event log information definitely has evidence of ongoing transactions [8]. For example, only existing recording of the event logs on the Case Defrauding Conspiration of Customer Bank Account (see Fig.2). Event Log contains information about activities in the form of a case or a specific task [9]. The case itself is called the "process instance" is an ongoing activity. For example, the order to the supplier (purchasing), the order by the customer (customer order) and several other events. While the task is activity in the trace, could be stages of activity [10]. So the trace have many tasks. Event Log consists of several attributes including caseid, attributename, activity, user (originator) and time (timestamp) [11].

978-1-5386-0549-3/17/$31.00 ©2017 IEEE

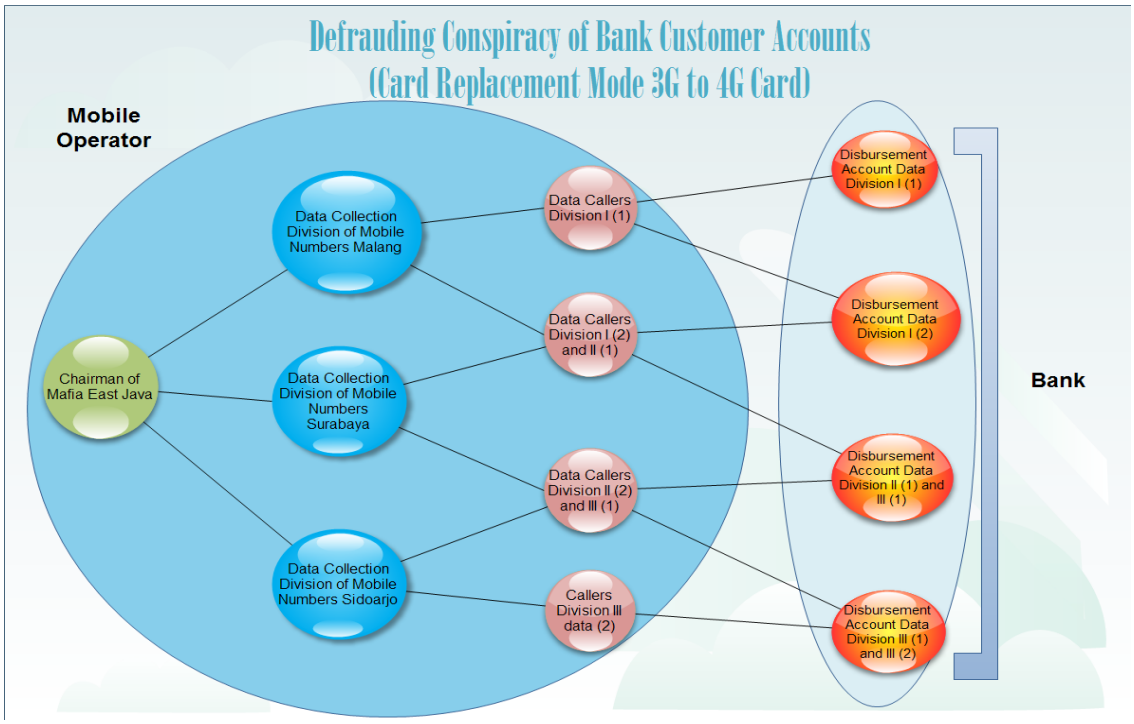


Fig. 1. Illustration of Network of Defrauding Conspiracy of Bank Customer Accounts

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- MXML version 1.0 -->
<!-- Created by Fluxicon Disco (http://fluxicon.com/disco/ -->
<!-- (c) 2016 Fluxicon - http://fluxicon.com/ -->
<WorkflowLog xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceS
    <Source program="Fluxicon Disco"/>
    <Process id="burglary bank account.burglary.mxml.gz" description="Converted to
        <ProcessInstance id="PP1">
            <Data>
                <Attribute name="Variant">Variant 1</Attribute>
                <Attribute name="Variant index">1</Attribute>
            </Data>
            <AuditTrailEntry>
                <Data>
                    <Attribute name="Amount">39</Attribute>
                    <Attribute name="CustomerID">C1</Attribute>
                    <Attribute name="PolicyType">Normal</Attribute>
                </Data>
                <WorkflowModelElement>collect customer phone numbers</WorkflowModel
                <EventType>complete</EventType>
                <Timestamp>2017-03-08T16:00:00.000+14:00</Timestamp>
                <Originator>Ali</Originator>
            </AuditTrailEntry>
            <AuditTrailEntry>
                <Data>
                    <Attribute name="TypeCollect">SM</Attribute>
                    <Attribute name="PhoneNumber">1</Attribute>
                </Data>
                <WorkflowModelElement>collecting phone number of customers via soci
                <EventType>complete</EventType>
                <Timestamp>2017-03-08T16:05:00.000+14:00</Timestamp>
                <Originator>Budi</Originator>
            </AuditTrailEntry>
            <AuditTrailEntry>
                <WorkflowModelElement>get more info</WorkflowModelElement>
                <EventType>complete</EventType>
                <Timestamp>2017-03-08T16:10:00.000+14:00</Timestamp>
                <Originator>Budi</Originator>
```

Fig. 2. Event Log .mxml for Case Defrauding Conspiration of Customer Bank Accounts

D. Dotted Chart Analysis

Dotted chart shows a visualization of an example of the process as colored dots on the schedule [12]. Whenever according to the specific event of the process being analyzed and each one colored dot on the timeline sequence corresponding to a special event to process. Color dot shows the events that occurred and the location of the point on the timeline is the time when it occurred (see Fig. 3). It indicates the level of difficulty dotted chart analysis contains information on all the cases recorded. for example, there is no possibility to determine the points that show the different events on the same model elements (eg, movement, change the name of special activities, creation) and did not investigate the technical attributes of the event log (ie, leaves no plug-ins). for it is some event log can be one focus for each (one for each sample) [13]. part of the event in the event log can then be grouped per model element. The dotted lines take an event log as input. it can be concluded that the dotted Chart can be used as a visual study of Parts Per Millions.

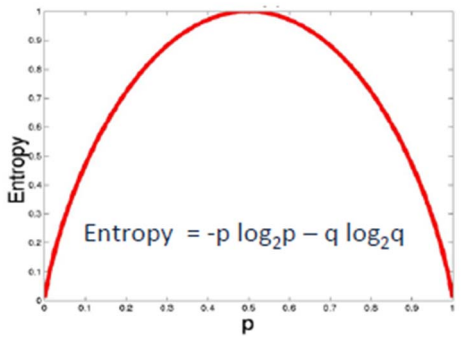


Fig. 3. Graph of Entropy Algorithm

E. Decision Tree Classification

Decision tree is how to establish a regression model or classification in such a tree structure visualization [14].

It can create a set of data into smaller sections of tree-related decisions that gradually developed in the same time [15] [16][17][18][19][20]. The end result of the decision tree is a tree with leaf nodes and nodes decisions.

Leaf node is a classification and decision. On this tree node section in accordance with the best prediction is called the root node. A decision tree can handle data both in numerical and categorical. Basic algorithm to construct a decision tree is ID3 (Iterative Dichotomiser 3) by J. R. Quinlan with a topdown, greedy search through the space without the possibility of backward branch. ID3 using Information Gain and Entropy to build a decision tree. A decision tree is built by using a top-down from the root node and involves partitioning the data into subsets that contain samples with the same value (homogeneous). ID3 algorithm in (1) is used to calculate the entropy of the homogeneity of the sample (see Fig. 4). That variable of p_i is meaning Probability. If different sample then homogenized entropy value is zero and if the value of entropy is one then the sample was equally divided.

$$E(S) = \sum_{i=1}^{c} p_i \, log_2 \, p_i \quad (1)$$

III. PROPOSED METHOD

A. Data Source

The data obtained by the authors is the data of the conspiration system in the form of defrauding customer bank accounts activity performed and data properties bank employees like in the Table I and then extracted to be .mxml format. The data in .mxml format consists of several data containing:

1. Case_id: contains user id procurement had been forged.
2. Activity_name: contains the name of the activity that has been performed.
3. Time_stamp: contains the date and start time activity.
4. Originator: contains resources that do activities.
5. Attribute_name: contain properties or data of activities.

B. Procedure Research

The procedures in this research starting from data collection event log from modelling simulation data in Tool Named YAWL that can see in Fig. 5. And then extraxted to be .xes and then extracted again to be .mxml format. Then use the .mxml formatted log event data to analysis with dotted chart analysis, decision point analysis and conformance checker using library in ProM that dotted chart analysis use for get decision atributed in each case. After that, calculate the value of entropy and information gain to calculate a rate of fraud (see Table. II) and using decision trees classification algorithm to find some rules and then make a structure of network conspiration that can be detect of scalable security. The formatted data in .mxml (can be seen Fig. 2). And then mining with ProM to look aoundness from yawl modelling to prom modelling (see Fig. 6). And then mining decision point analysis to find decicion trees of conspiration (see Fig. 7). And then minning dotted chart analysis to analysis distributed event in over time with dotted chart analysis. In dotted chart analysis that can be seen that brown circle is the succesful status with high amount and low amount attack or heavy attack. And other color circle is slight attack or failed status attack or conspiration (see Fig. 8).

TABLE I. TABLE EVENT LOG BANK FINANCIAL CREDIT BUSINESS PROCESS

Case ID	Policy Type	Status Call	Content SMS	PIN	Amount Account Bank	Total Money	Amount	TypeFM	Status
1	Normal	CanBe Calling	Ganti	Given	TheFirstIntermediary	100000000	Unknown	Unknown	Successful with High Amount
2	Normal	CantBe Calling	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Failed
3	Normal	CanBe Calling	Null	NotGiven	Unknown	Unknown	Unknown	Unknown	Failed
4	Normal	CanBe Calling	Ganti	Given	TheFirstIntermediary	50000000	Unknown	Unknown	Successful with Low Amount
5	Normal	CanBe Calling	Ganti	Given	TheSecondIntermediary	100000000	Unknown	Unknown	Successful with High Amount
6	Normal	CanBe Calling	Ganti	Not Given	Unknown	Failed	Unknown	Unknown	Failed
7	Reguler	CanBe Calling	Ganti	Given	TheFirstIntermediary	100000000	DistantRelatives	PublicNumbers	Successful with High Amount
8	Reguler	CantBe Calling	Null	Unknown	Unknown	Unknown	NearofKin	PrivateNumbers	Failed
9	Reguler	CanBe Calling	Null	NotGiven	Unknown	Unknown	DistantRelatives	PublicNumbers	Failed
10	Reguler	CanBe Calling	Ganti	Given	TheFirstIntermediary	100000000	NearofKin	PrivateNumbers	Successful with High Amount
11	Reguler	CanBe Calling	Ganti	Given	TheSecondIntermediary	100000000	DistantRelatives	PublicNumbers	Successful with High Amount
12	Reguler	CanBe Calling	Ganti	Given	TheFirstIntermediary	100000000	NearofKin	PrivateNumbers	Successful with High Amount
13	Premium	CanBe Calling	Ganti	Given	TheFirstIntermediary	50000000	Unknown	Unknown	Successful with Low Amount
14	Premium	CantBe Calling	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Failed

Case ID	Policy Type	Status Call	Content SMS	PIN	Amount Account Bank	Total Money	Amount	TypeFM	Status
15	Premium	Unknown	Null	Not Given	Unknown	Unknown	Unknown	Unknown	Failed
16	Premium	CanBe Calling	Ganti	Given	TheFirstIntermediary	50000000	Unknown	Unknown	Successful with Low Amount
17	Premium	CanBe Calling	Ganti	Given	TheSecondIntermediary	50000000	Unknown	Unknown	Successful with Low Amount
18	Premium	CanBe Calling	Ganti	Given	TheFirstIntermediary	Failed	Unknown	Unknown	Failed

TABLE II. TABLE ACTUAL AND OBTAIN ACTIVITY

Case ID	Policy Type	Status Call	Content SMS	PIN	Amount Account Bank	Total Money	Amount	Type FM	Amount	Rating	Status
10	0	2	2	2	0	2	2	2	12	1.5	Successful with High Amount
12	0	2	2	2	0	2	2	2	12	1.5	Successful with High Amount
11	0	2	2	2	2	2	0	0	10	1.25	Successful with High Amount
5	2	2	2	2	2	2	-2	-2	8	1	Successful with High Amount
7	0	2	2	2	0	2	0	0	8	1	Successful with High Amount
1	2	2	2	2	0	2	-2	-2	6	0.75	Successful with High Amount
17	-2	2	2	2	2	2	-2	-2	4	0.5	Successful with Low Amount
4	2	2	-2	2	0	2	-2	-2	2	0.25	Successful with Low Amount
13	-2	2	2	2	0	2	-2	-2	2	0.25	Successful with Low Amount
16	-2	2	2	2	0	2	-2	-2	2	0.25	Successful with Low Amount
6	2	2	2	0	-2	0	-2	-2	0	0	Failed
18	-2	2	2	2	0	0	-2	-2	0	0	Failed
9	0	2	0	0	-2	-2	0	0	-2	-0.25	Failed
3	2	2	0	0	-2	-2	-2	-2	-4	-0.5	Failed
8	0	-2	0	-2	-2	-2	2	2	-4	-0.5	Failed
15	-2	2	0	0	-2	-2	-2	-2	-8	-1	Failed
2	2	-2	-2	-2	-2	-2	-2	-2	-12	-1.5	Failed
14	-2	-2	-2	-2	-2	-2	-2	-2	-16	-2	Failed

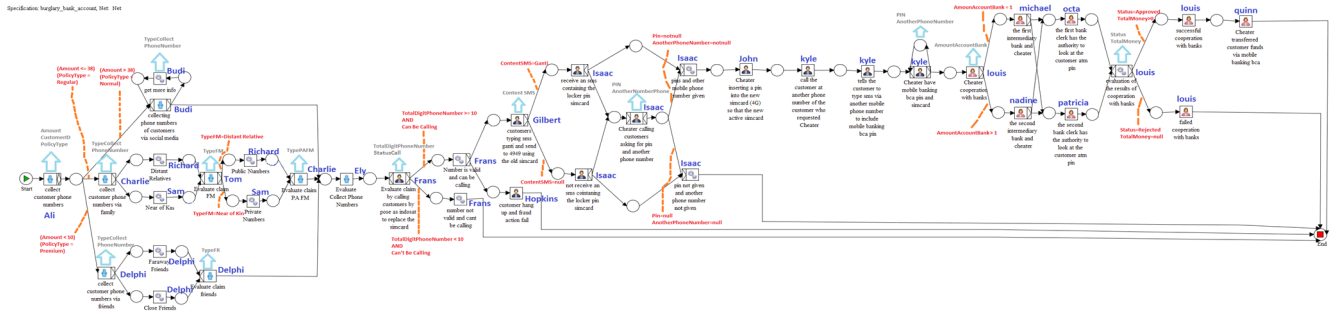
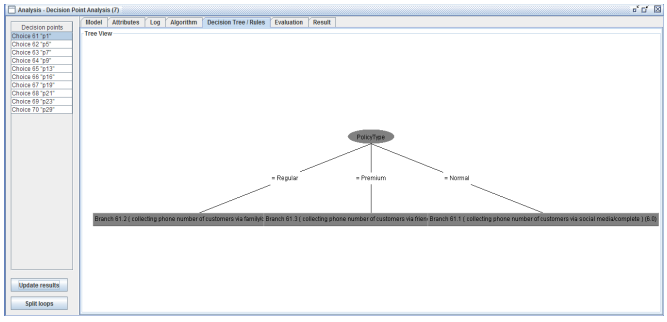


Fig. 4. Standard Operational Procedure Defrauding Conspiracy of Customer Bank Account (card replacement mode 3G to 4G card)



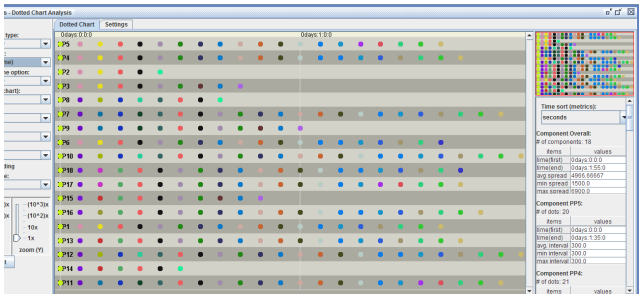


Fig. 5. Decision Tree Results in Decision Point Analysis of Case Defrauding

Fig. 6. Dotted Chart Analysis Defrauding Conspiracy of Case Customer Bank Account

IV. Result and Analysis

The experiments for this research is a defrauding conspiration of bank customer account data and process. This research using event log with 18 cases to determine which one is heavy attack and which one is slight attack. This paper uses 8 indicators or ways to detect attack of security scalable. The indicator i.e. PolicyType, StatusCall, ContentSMS, PIN, AmountAccountBank, TotalMoney, Amount, TypeFM and Status. First is Attribute value of Defrauding Conspiracy of Customer Bank Account (see Table. II). The Great Value Gain Then More Difficult For Thieves break into accounts, Small Value Gain Then The Easy thieves breaking into account, The Great Value Entropy then The Big Opportunity thief to break into accounts, Small Value Entropy So Small growing opportunities thief To break into an account. Attribute of each value of case defrauding conspiracy of customer bank account can be seen in Table. III:

TABLE III. Attribute value of Defrauding Conspiracy of Customer Bank Account

PolicyType	**Scoring**	**AmountAccountBank**	**Scoring**
Normal	2	TheSecondIntermediary	2
Reguler	0	TheFirstIntermediary	0
Premium	-2	Unknown	-2
StatusCall		**TotalMoney**	
CanBecalling	2	Successful high amount	2
CantBecalling	0	Successful low amount	0
Unknown	-2	Failed	-2
ContentSMS		**Amount**	
Ganti	2	NearofKin	2
Null	0	DistantRelatives	0
Unknown	-2	Unknown	-2
PIN		**TypeFM**	
Given	2	PrivateNumbers	2
NotGiven	0	PublicNumbers	0
Null	-2	Unknown	-2

TABLE IV. Calculate Entropyof The Target

Status	
Successful	**Failed**
1	4
Entropy(Status)	0.721928095

TABLE V. Calculate Entropy use the frequency table of two attributes

		Status	
		Successful	Failed
Total Money	Successful	10	0
	Failed	0	2
	Unknown	0	6
	Entropy (Status,TotalMoney)	0.388316669	

And then calculation of entropy and information gain for knowing information gain and entropy using two attributes of frequency table for each decision point is like Table IV and Table V:

TABLE VI. Information Gain using two attributes of frequency table

InformationGain	Value
Gain (Status,Amount)	0.018310782
Gain (Status,TypeFM)	0.018310782
Gain (Status,PolicyType)	0.018310782
Gain (Status,StatusCall)	0.225829531
Gain (Status,AmountAccountBank)	0.492478806
Gain (Status,ContentSMS)	0.557727779
Gain (Status,Pin)	0.557727779
Gain (Status,TotalMoney)	0.602759391

TABLE VII. Entropy using two attributes of frequency table

Entropy	Value
Entropy (Status,PolicyType)	0.972765278
Entropy (Status,Amount)	0.972765278
Entropy (Status,TypeFM)	0.972765278
Entropy (Status,StatusCall)	0.765246528
Entropy (Status,ContentSMS)	0.433348281
Entropy (Status,Pin)	0.433348281
Entropy (Status,AmountAccountBank)	0.498597254
Entropy (Status,TotalMoney)	0.388316669

TABLE VIII. Information Gain for

Rule	Attack Type	Value
Gain (Status,TotalMoney)	Heavy Attack	0.333611426
Gain (Status,StatusCall)	Slight Attack	0.288579814
Gain (Status, Pin)	Slight Attack	-0.043318433

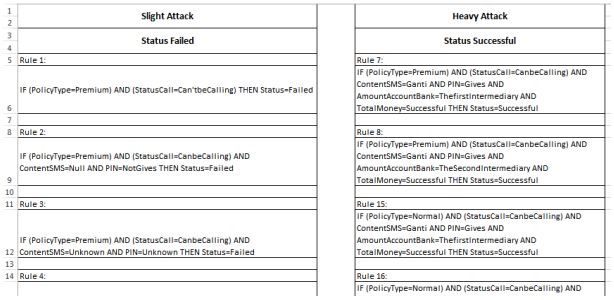


Fig. 7. Rules of Scalable Security Attack for slight and heavy attack

Based on Table. VI and Table. VII value of entropy and information gain, that value can be used for structuring decision trees, higher value is first brach and lower value is last brach. Decision trees for entropy and information gain that have been structured is like Fig. 10 The decision tree is simple modified to change to the conditions of the mapping of the root node to node per one leaf. Based on this case that make 33 rules with classification succesful with high amount and low amount or heavy attack is 8 rules and failed status or slight attack is 25 rules (see Fig. 9).

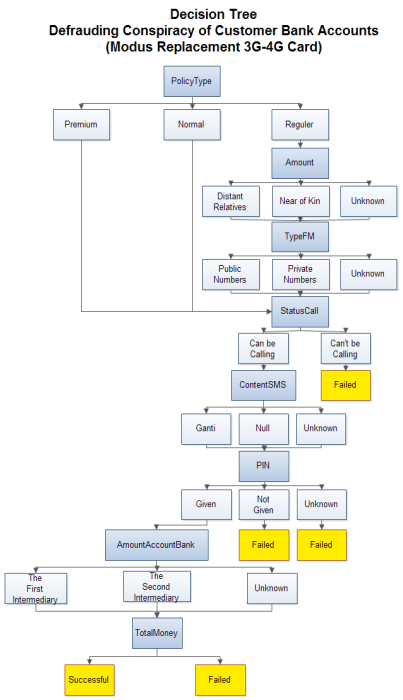


Fig. 8. Decision Treen Defrauding Conspiracy

TABLE IX. THE QUALITY OF ALPHA++ ALGORITHM

Defrauding Conspiracy of Customer Bank Account using	Fitness	Precision	Structure
Alpha++ Algorithm	1	0.87	1

V. CONCLUSION

This experiment identifies 18 cases of Defrauding Conspiracy of Customer Bank Account. The results show that: a) six cases successfully attack achieving high amount status; b) four cases heavily attack achieving low amount status; and c) eight cases lightly attack failing defrauding status. There is 33 rules consisting 8 rules for heavy attacks and 25 rules for light attacks. This paper concludes that the accuracy of the experiment by using Alpha++ is 0.87%.

REFERENCES

[1] J. B. Hong and D. S. Kim, "Towards scalable security analysis using multi-layered security models," Journal of Network and Computer Applications, vol. 75, pp. 156–168, Nov. 2016. https://doi.org/10.1016/j.jnca.2016.08.024

[2] X. Dong, H. Lu, Y. Xia, and Z. Xiong, "Decision-Making Model under Risk Assessment Based on Entropy," Entropy, vol. 18, no. 11, p. 404, Nov. 2016. https://doi.org/10.3390/e18110404

[3] F. Bezerra and J. Wainer, "Algorithms for anomaly detection of traces in logs of process aware information systems," Information Systems, vol. 38, no. 1, pp. 33–44, Mar. 2013. https://doi.org/10.1016/j.is.2012.04.004

[4] H. A. Reijers, M. Song, and B. Jeong, "Analysis of a collaborative workflow process with distributed actors," Information Systems Frontiers, vol. 11, no. 3, pp. 307–322, Jul. 2008. https://doi.org/10.1007/s10796-008-9092-5

[5] C. F. Graumann and S. Moscovici, Eds., "Changing Conceptions of Conspiracy," 1987. https://doi.org/10.1007/978-1-4612-4618-3

[6] T. B. Hadden, "Conspiracy to Defraud," The Cambridge Law Journal, vol. 24, no. 02, p. 248, Nov. 1966. https://doi.org/10.1111/j.1468-2230.1960.tb00638.x

[7] S. Huda, R. Sarno, and T. Ahmad, "Increasing Accuracy of Process-based Fraud Detection Using a Behavior Model," International Journal of Software Engineering and Its Applications, vol. 10, no. 5, pp. 175–188, May 2016. https://doi.org/10.14257/ijseia.2016.10.5.16

[8] M. Wynn, A. Rozinat, W. van der Aalst, A. ter Hofstede, and C. Fidge, "Process Mining and Simulation," Modern Business Process Automation, pp. 437–457, Sep. 2009. https://doi.org/10.1007/978-3-642-03121-2_17

[9] L. Wen, W. M. P. van der Aalst, J. Wang, and J. Sun, "Mining process models with non-free-choice constructs," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 145–180, Mar. 2007. https://doi.org/10.1007/s10618-007-0065-y

[10] D. Rahmawati, M. A. Yaqin and R. Sarno, "Fraud Detection on Event Logs of Goods and Services Procurement Business Process Using Heuristics Miner Algorithm," in The 10th International Conference on Information & Communication Technology and System (ICTS)., Surabaya, 2016.

[11] R. Sarno and K. R. Sungkono, "Coupled Hidden Markov Model for Process Mining of Invisible Prime Tasks," International Review on Computers and Software (IRECOS), pp. 539-547, 2016. https://doi.org/10.15866/irecos.v11i6.9555

[12] T. Molka, W. Gilani, and X.-J. Zeng, "Dotted Chart and Control-Flow Analysis for a Loan Application Process," Lecture Notes in Business Information Processing, pp. 223–224, 2013 https://doi.org/10.1007/978-3-642-36285-9_26

[13] Claes, J., Vanderfeesten, I., Pinggera, J. et al. Inf Syst E-Bus Manage (2015) 13: 147. doi:10.1007/s10257-014-0245-4

[14] A. Rozinat and W. M. P. van der Aalst, "Decision Mining in ProM," Business Process Management, pp. 420–425, 2006. https://doi.org/10.1007/11841760_33

[15] W. M. P. van der Aalst and C. W. Gunther, "Finding Structure in Unstructured Processes: The Case for Process Mining," Seventh International Conference on Application of Concurrency to System Design (ACSD 2007), Jul. 2007. https://doi.org/10.1109/acsd.2007.50

[16] R. Bergenthum, J. Desel, R. Lorenz, and S. Mauser, "Process Mining Based on Regions of Languages," Business Process Management, pp. 375–383. https://doi.org/10.1007/978-3-540-75183-0_27

[17] L. Rokach and O. Maimon, "Decision Trees," Data Mining and Knowledge Discovery Handbook, pp. 165–192 https://doi.org/10.1142/9097

[18] P. Máša and T. Kočka, "Finding Optimal Decision Trees," Intelligent Information Processing and Web Mining, pp. 173–181. https://doi.org/10.1007/3-540-33521-8_17

[19] L. E. Raileanu and K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria," Annals of Mathematics and Artificial Intelligence, vol. 41, no. 1, pp. 77–93, May 2004. https://doi.org/10.1023/b:amai.0000018580.96245.c6

[20] R. Sarno, Fernandes, D. Sunaryono and A. Munif, "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning," in The 2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2015. https://doi.org/10.1109/ic3ina.2015.7377738