Detecting Business Process Anomaly Using Graph Similarity Based on Dice Coefficient, Vertex Ranking and Spearman Method

Cecilia Stephanie
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ciliata7@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Abstract—**Graph similarity is a method to check the graphic equations between two graphs. Anomaly detection in a business process can use the graph similarity method as well. The graph of the ongoing business process compared to the Standard Operating Procedure's graph (SOP's graph) as a reference. Standard Operating Procedures (SOPs) can definitely as a benchmark of a company running a business process. In this study, the case taken to detect anomaly of business processes in the process of replacing parts in wireless device distributors. The method used is dice coefficient to detect vertex or edge overlap and vertex ranking using the Spearman method to detect missing or altered vertices. The result of detecting business process anomaly using dice coefficient found three business process anomalies. While in the vertex ranking method using Spearman coefficient found two anomalies. Anomalies were found in vertex ranking method are the same as the expert judgement's results.**

Keywords—graph similarity;business process anomaly;dice coefficient;Vertex Ranking, Spearman method

I. INTRODUCTION

Business processes that exist in the company must compared with the applicable Standard Operating Procedure (SOP). SOP serves as a guide how business processes should be working. In fact, the business processes that run in the field are not always the same as the existing SOP. Deviation distance between SOP and prevailing business processes called anomaly. A deviation is definitely an anomaly when it exceeds the firm-specified thresholds. Furthermore, anomaly also used as a guide to update SOP or even find fraud from the current business process.

One way that used to detect anomaly is by the graph similarity method. Graph similarity is a method by comparing the similarities of SOP graphs and business processes. The SOP graph consists of vertices and edges. Vertex describes every task that acts as a complex object connected by edges. Edge is the interaction between two objects or vertices [1]. Graph similarity used to detect missing vertices, missing edge, vertices modification and edges modification [2].

The deviation of SOP graphics and business process can extracted with Vertex / Edge Overlap algorithm and Vertex Ranking algorithm. Vertex / Edge Overlap calculates the distance between SOP graph and business process graph with the Dice Coefficient method. On the other hand, Vertex Ranking calculates the distance between graphs with Spearman method. After calculated, the deviations of the three methods above compared. Before drawing a conclusion, it is necessary to establish a similarity threshold that will separate the anomaly and the graph that considered almost similar [3]. Determination of threshold in this study using Process-Based Fraud (PBF) rating. The results of the experiments will compared with the expert judgment of the company.

II. LITERATURE

A. Standard Operating Procedure (SOP)

Standard Operating Procedure (SOP) can define as "a set of written and detailed instructions that document a routine or repetitive activity followed by an organization to achieve uniformity of the performance of a specific function" [4]. SOP ensures all workers perform tasks in the same way and the process must produce consistent output. If it does not run correctly, then the whole operation will fail [5]. SOPs allow organizations to be consistent and consistent with cross-departmental processes that will impact on products and services. Unlike before, Edelson & Bennett argued that SOPs were a management effort to improve the results of the process consistently and efficiently [5]. If there is no SOP, then there is no measure for the company to determine whether the running process is the best work process. Therefore, SOP must be reviewed periodically whether it is still valid at this time or not.

B. Graph Similarity

Graph similarity is applied to many fields and is measured by various similarity measurement [6]. One of the areas that apply similarity is a business process model. The graph similarity technique can divided into three categories with distance, iterative and feature extraction. In this paper, we use it for determine distance between SOP graph and business process graphs. For example, we have two graphs that have similarity checked. These two graphs have the possibility of having different vertices and edges. The graph similarity approach will produce similarity measurement values between "0" and "1" [7]. The graph differs or suspected as anomaly if value is close or equal to "0". Conversely, if the value is close or equal to "1" can be defined the same graph.

C. Anomaly

Anomalies can definitely use patterns that do not conform to normal behavior [8]. There are several factors from [8], which make the anomaly difficult to detect, such as:

- The boundary between the anomalous case and the normal case is not too obvious.
- Normal state influenced by many factors causing the change of anomaly boundary in the future.
- Data confounding is difficult to identify and causes anomaly indications to become chaotic.

That challenges makes anomaly detection difficult. Most existing anomaly detection techniques identify anomalies based on the nature of the data, availability of label data and anomaly types to be detected. In this paper, we try to detect anomalies by setting threshold according to PBF rating reference. Furthermore, the anomalous results of the graph similarity will compared with the expert judgment of the company.

D. PBF rating

Detecting process-based fraud (PBF) can see from three different angles, there are the point of view of the business process, business role and organization [9]. Comparing different business processes with respective models is the first one. Second, PBF analyzes the deviation of the business role. Third, PBF analyzes the people who do the tasks according to the division of tasks. The superiority of PBF can compare business processes and SOPs [10]. PBF can also detect missing vertices or edges. In this paper, we only use the first approach to check the process (the point of view business process). The threshold can be determined from the score of risk. The risk scores obtained from the number of business processes detected, divided by a hundred then multiplied by the weight factor [11]. Weight factor usually set by expert judgment in company because that person knows the risk factors in their business processes. There are three approaches to determine the weight factor is pressure, opportunity, and rationalization [12].

III. PROPOSED METHOD

A. Vertex / Edge Overlap

The vertex / edge overlap method approach is to calculate the number of vertices or edges of two graphs [2]. The advantage of this method is that it can detect the missing vertices or edges of the graphs tested with SOP graphs created as a comparison. It is undeniable that there are deficiencies in the vertex / edge overlap method of looking for ignore the sensitivity and coverage requirements. Similarity can be measured using Jaccard index or Dice coefficient. The result range of Jaccard and Dice method is equally between 0-1. The difference between Jaccard index and Dice coefficient is on the dice method of comparing the intersection to the total number tested. While the Jaccard index, compare the intersection with the disjunction of the data tested. The intersection of the graph is the data that is on both graphs. Disjunction is a non-slice diagram. Union is a composite data of two graphs or more precisely the sum of intersection and disjunction. The difference of intersection, union, and disjunction described in Figure 1.

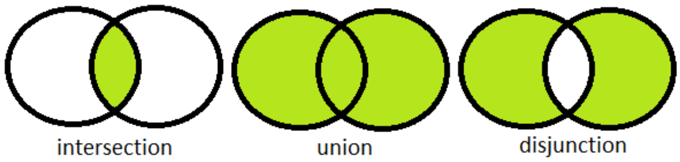

Figure 1. The difference of intersection, union, and disjunction

In this paper, we choose to use Dice coefficient because it's more to describe the similarity of graph than the Jaccard index because there may be value outside disjunction. The formula used is twice the sum of the vertices and edge intersection and then divided the total vertices and edges of the two graphs. The formula Vertex/Edge Overlap between SOP's graph (G) and business process case's graph (G')

$$Dice\,(G, G') = 2\,\frac{(V \cap V') + (E \cap E')}{(V + V') + (E + E')} \qquad (1)$$

Where:
- V = total vertices in graph G
- V' = total vertices in graph G'
- E = total edges in graph G
- E' = total edges in graph G'

B. Vertex Ranking

The definition of vertex ranking is a method of comparing vertices of two graphs by ranking. Vertex on both graphs will rank based on whether or not there is a vertex in the business process. In paper [2], Spearman method usually applied to connect a ranking variable. The use of such methods is still rare for graph similarity. First, vertex will sort based on the priority ranking of vertices that must exist. Then, look for the distance between the SOP's vertex and the business process case's vertex. Distance from the vertices of SOP and business process vertices squared and entered into superman equations. Spearman method formula comparing SOP's graph and business process cases described in

$$r_s = 1 - \frac{6\sum di^2}{n(n^2 - 1)} \qquad (2)$$

Where:
- r_s = Spearman coefficient
- di^2 = square distance between SOP's vertex and business process case's vertex
- n = total vertex that has the same rank

In the above equation, distance symbolized by d_i, while n is the number of vertices in the graph. Equation in (2) only be used if there is no same ranking value on one graph, if there is the same ranking value in the graph use formula

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum d_i^2}{2\sqrt{\sum x^2 . \sum y^2}} \quad (3)$$

Where:

x^2 = square of total vertices in the SOP graph
y^2 = square of total vertices in the business process case graph
d_i^2 = square distance between SOP's vertex and business process case's vertex

The x^2 in equation obtained from n squared three, then reduced by n, detail equations can be look in (4) and (5).

$$\sum x^2 = \frac{n^3 - n}{12} - \sum Tx \quad (4)$$

$$\sum y^2 = \frac{n^3 - n}{12} - \sum Ty \quad (5)$$

Where:

x^2 = square of total vertices in the SOP graph
y^2 = square of total vertices in the business process case graph
n = total vertex that has the same rank
$\sum Tx$ = total of the average SOP's vertices twin rank
$\sum Ty$ = total of the average business process case's vertices twin rank

Then, divided it by twelve then reduced by the average twin ranking value ($\sum Tx$). In equation (6), $\sum Tx$ is defined by the number of twin rankings for each ranking (t), it squared by three and reduce with (t), last the result divided by twelve.

$$Tx \,/\, Ty = \frac{t^3 - t}{12} \quad (6)$$

Where:

Tx = the average SOP's vertices twin rank
Ty = the average business process case's vertices twin rank
t = total vertices that have twin rank

Vertices that do not exist on other graph would be easy to detect. The challenge of this method is to rank appropriate for business processes.

IV. EXPERIMENT AND RESULTS

Business process taken for this experiment is the replacement parts of wireless devices on the distributor of wireless devices. The study will compare SOPs with 10 cases of running business processes. In SOP, there are 23 tasks with 28 vertex and 58 edges. Task of SOP's replacement parts of wireless devices described in Table 1. SOP's graph illustrated in Figure 2.

TABLE 1. Task of SOP's replacement parts of wireless devices

Task	Activity
t1	Request parts replacement
t2	Check device
t3	Inform status device
t4	Create device receipt
t5	Fill the parts replacement registration
t6	Deliver device to the technician
t7	Service device
t8	Check service results
t9	Calculate cost
t10	Make a service offer
t11	Run billing SOP
t12	Create delivery order
t13	Check the warranty
t14	Make notification of damaged device
t15	Inform replacement device to the customer
t16	Request device replacement to vendor
t17	Replacement parts
t18	Check replacement stock of parts
t19	Ask for approval replacement
t20	Fill in the parts request form
t21	Check sales stock of parts
t22	Request stock release
t23	Create job costing

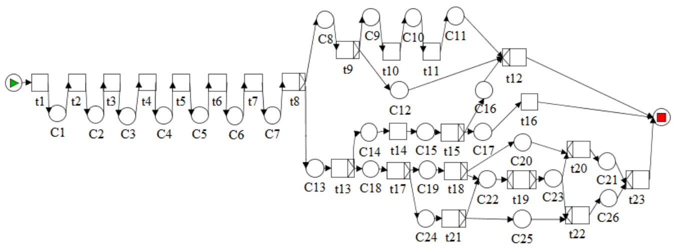

Figure 2. Graph for SOP's replacement parts of wireless devices

The SOP graph illustrates all possible business processes at the company. This means not all tasks must missed. One case study of anomaly detection is Figure 3. The first step of anomaly detection Vertex / Edge Overlap by Dice coefficient method is calculate the number of vertices and edges of the SOP graph (Figure 2). The second step is to calculate the vertex and edge intersections of the SOP charts and business processes. Finally, enter the results of the first and second steps into the formula (1). Example for case study 1 can see in Figure 3.

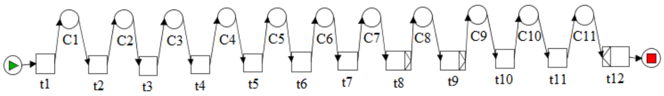

Figure 3. Case study 1

Before the experiment, we must set the threshold which one business process stated anomaly. Anomalies detect by the risk factor. Each risk factor in business process given weight score. In one hundred experiment, ten cases determined as anomaly which weight factor for that risk factor is five. Then, we get 0.5 for threshold. When the result of an experiment with dice coefficient methods or vertex ranking using Spearman method greater or equal to 0.5, we identify similar business processes, and less than 0.5 for anomaly.

Dice coefficient used to find overlapping vertices and edges. The number of vertices in the SOP is 28 vertices, while the number of edges in the SOP is 58 edges. The number of vertices and edges in the case study explained in Table 2. Column $V \cap V'$ shows the number of vertices overlap between SOP and each case studies. Column $E \cap E'$ shows the number of edges overlap between SOP and case studies. The higher the value of vertices and edges overlap, then the degree of similarity SOP's graph and business process case study's graph higher. Then the data in Table 2 entered into the equation (1) to get the dice coefficient.

TABLE 2. The number of vertices and edges in the case study

	V	V'	E	E'	$V \cap V'$	$E \cap E'$
SOP vs case 1	28	11	58	24	11	24
SOP vs case 2	28	9	58	20	9	20
SOP vs case 3	28	11	58	24	11	24
SOP vs case 4	28	11	58	24	11	24
SOP vs case 5	28	12	58	26	12	26
SOP vs case 6	28	13	58	29	13	29
SOP vs case 7	28	13	58	29	13	29
SOP vs case 8	28	8	58	18	7	16
SOP vs case 9	28	7	58	16	3	8
SOP vs case 10	28	12	58	26	7	16

The result of graph similarity calculation using the Dice Coefficient method is in the range 0.2 to 0.65. With the determination of the threshold of 0.5, cases in eighth, ninth, and tenth can categorized as anomaly. Once checked, there are many missing vertices and edges, also many modification vertices and edges. Overall, result of low similarity maximum only 0.65. This is because there are many branches in the SOP that only need to be selected one branch.

The first step in the Vertex Ranking method is rank for each task in business process. After that, we can get the distance, which uses in Spearman method. A rank of SOP and each task in that business process case study showed in Table 3. Distance and square distance from each task in business process case to SOP described in Table 4 and Table 5. After getting the square distance, we input the value to the equation (3).

TABLE 3. Rank of SOP and each task in that business process case study

Task	SOP	1	2	3	4	5	6	7	8	9	10
t1	12	9	7	8	8	10	11	11	7		11
t2	14	11	9	9	9	11	12	12			12
t3	11	8	6	6	6	9	10	10			10
t4	1	1	1	1	1	1	1	1	1		
t5	8	6	4	4	4	7	8	8	4	4	8
t6	10	7	5	5	5	8	9	9	5	5	
t7	13	10	8	10	10	12	13	13	8		
t8	9	5	3						3		
t9	2	2									
t10	3	3									
t11	4	4									
t12	4		2						2	2	
t13	8			5	5	5	7	7			
t14	2			2	2						
t15	7			7	7						
t16	4			3							
t17	3				3						
t18	2					2	2	2			
t19	5					4	4				4
t20	6						5				5
t21	4						6				3
t22	3					3	3				
t23	7					6					
t24	5							5			
t25	6							6			
t26	4							4			
t27	3							3			
t28	7										
t29	0								6	6	
t30	0									7	
t31	0									1	
t32	0									3	
t33	0										1
t34	0										2
t35	0										7
t36	0										9
t37	0										6
#case	28	11	9	11	11	12	13	13	8	7	12

TABLE 4. Distance from each task in business process case to SOP

Task	1	2	3	4	5	6	7	8	9	10
t1	3	5	4	4	2	1	1	5	12	1
t2	3	5	5	5	3	2	2	14	14	2
t3	3	5	5	5	2	1	1	11	11	1
t4	0	0	0	0	0	0	0	0	1	1
t5	2	4	4	4	1	0	0	4	4	0
t6	3	5	5	5	2	1	1	5	5	10
t7	3	5	3	3	1	0	0	5	13	13
t8	4	6	9	9	9	9	9	6	9	9
t9	0	2	2	2	2	2	2	2	2	2
t10	0	3	3	3	3	3	3	3	3	3
t11	0	4	4	4	4	4	4	4	4	4
t12	4	2	4	4	4	4	4	2	2	4
t13	8	8	3	3	3	1	1	8	8	8
t14	2	2	0	0	2	2	2	2	2	2
t15	7	7	0	0	7	7	7	7	7	7
t16	4	4	1	4	4	4	4	4	4	4
t17	3	3	3	0	3	3	3	3	3	3
t18	2	2	2	2	0	0	0	2	2	2
t19	5	5	5	5	1	1	5	5	5	1
t20	6	6	6	6	6	1	6	6	6	1
t21	4	4	4	4	4	-2	4	4	4	1
t22	3	3	3	3	0	0	3	3	3	3
t23	7	7	7	7	1	7	7	7	7	7
t24	5	5	5	5	5	5	0	5	5	5
t25	6	6	6	6	6	6	0	6	6	6
t26	4	4	4	4	4	4	0	4	4	4
t27	3	3	3	3	3	3	0	3	3	3
t28	7	7	7	7	7	7	7	7	7	7
t29	0	0	0	0	0	0	0	22	22	0
t30	0	0	0	0	0	0	0	0	21	0
t31	0	0	0	0	0	0	0	0	27	0
t32	0	0	0	0	0	0	0	0	25	0
t33	0	0	0	0	0	0	0	0	0	27
t34	0	0	0	0	0	0	0	0	0	26
t35	0	0	0	0	0	0	0	0	0	21
t36	0	0	0	0	0	0	0	0	0	19
t37	0	0	0	0	0	0	0	0	0	22

TABLE 5. Square distance (di^2) from each task in business process case to SOP

	1	2	3	4	5	6	7	8	9	10
t1	9	25	16	16	4	1	1	25	144	1
t2	9	25	25	25	9	4	4	196	196	4
t3	9	25	25	25	4	1	1	121	121	1
t4	0	0	0	0	0	0	0	0	1	1
t5	4	16	16	16	1	0	0	16	16	0
t6	9	25	25	25	4	1	1	25	25	100
t7	9	25	9	9	1	0	0	25	169	169
t8	16	36	81	81	81	81	81	36	81	81
t9	0	4	4	4	4	4	4	4	4	4
t10	0	9	9	9	9	9	9	9	9	9
t11	0	16	16	16	16	16	16	16	16	16
t12	16	4	16	16	16	16	16	4	4	16
t13	64	64	9	9	9	1	1	64	64	64
t14	4	4	0	0	4	4	4	4	4	4
t15	49	49	0	0	49	49	49	49	49	49
t16	16	16	1	16	16	16	16	16	16	16
t17	9	9	9	0	9	9	9	9	9	9
t18	4	4	4	4	0	0	0	4	4	4
t19	25	25	25	25	1	1	25	25	25	1
t20	36	36	36	36	36	1	36	36	36	1
t21	16	16	16	16	16	4	16	16	16	1
t22	9	9	9	9	0	0	9	9	9	9
t23	49	49	49	49	1	49	49	49	49	49
t24	25	25	25	25	25	25	0	25	25	25
t25	36	36	36	36	36	36	0	36	36	36
t26	16	16	16	16	16	16	0	16	16	16
t27	9	9	9	9	9	9	0	9	9	9
t28	49	49	49	49	49	49	49	49	49	49
t29	0	0	0	0	0	0	0	484	484	0
t30	0	0	0	0	0	0	0	0	441	0
t31	0	0	0	0	0	0	0	0	729	0
t32	0	0	0	0	0	0	0	0	625	0
t33	0	0	0	0	0	0	0	0	0	729
t34	0	0	0	0	0	0	0	0	0	676
t35	0	0	0	0	0	0	0	0	0	441
t36	0	0	0	0	0	0	0	0	0	361
t37	0	0	0	0	0	0	0	0	0	484
	497	626	535	541	425	402	396	1377	3481	3435

The results of calculations using Spearman method vary greatly between 0.35 to 0.89. The difference is due to ranking first on the vertex, so the determination of missing vertex and vertex modification is more accurate. At the time of ranking, there are some difficulties if a vertex exists only one diagram and if there is the same ranking value. The solution to such difficulties is to rank the most recent SOPs for vertices that exist only in the case of business processes. If there is the same ranking value, then use Spearman method formula on (3). With the determination of the threshold value of 0.5 then obtained cases ninth and tenth have anomaly tendencies.

The results of the Vertex Ranking using Spearman method in this study are very good because it is able to produce data that really contrast. It cannot be denied that the vertex ranking is very dependent on the selection of task necessity exists as a ranking reference. If ranking is not correct mapping, then Spearman method failure rate will be very high. The results of Dice Coefficient and Spearman method calculations shown in Table 6.

TABLE 6. Results of computing graph similarity

Case Study	Dice Coefficient	Vertex Ranking using Spearman method
1	0.57851	0.86341

2	0.50435	0.82796
3	0.57851	0.85295
4	0.57851	0.85130
5	0.61290	0.88320
6	0.65625	0.88952
7	0.65625	0.89117
8	0.41071	0.65954
9	0.20183	0.35958
10	0.37097	0.42352

In the case study ninth, the lowest dice coefficient is 0.20183. This is because the number of intersections between vertices between SOP and business process case study having the smallest value is three (see in Table 2). The number of intersections between the edges between SOP and business process case study having the smallest value too, is eight. In contrast, in the case studies sixth and seventh had the highest number of vertex overlaps with SOPs, thirteen. The highest number of edge overlaps also in the case studies sixth and seventh, was twenty-nine compared to the SOP.

The vertex ranking value based on the Spearman coefficient has the highest value, 0.89117 in the case study seventh. That is because the ranking of tasks in case study seventh has a rating that is almost the same as SOP's tasks (see in Table 3). The lowest value of the vertex ranking based on the Spearman coefficient is in case study ninth, which is 0.35958. In case ninth, there were four tasks were not in the SOP. Tasks that are not in the SOP's graph ranked as null or zero. This causes the distance value between the ninth business process's graph and the SOP's graph to be high and produces a low value of Spearman coefficient.

V. Conclusion

The conclusion from the research that the algorithm that used to compare business process graph is dice coefficient and Spearman method. Dice coefficient is suitable for binary vectors or two sets of data, X and Y. The result of comparing ten cases of business processes with dice coefficient found there are three anomaly processes with threshold 0.5, cases eighth, ninth and tenth.

In other hand, Spearman method obtained by ranking for every vertex in business process and SOP and then get distance from it. Spearman method measurements also found two cases of anomaly business processes (under threshold 0.5), cases ninth and tenth. After comparison with expert judgement, found that only case ninth and tenth were anomalies.

It was concluded that use several methods can obtain more accurate anomaly indicators than one method. In the next study, we can add the method of Hamming distance and Jaccard index.

References

[1] Chen, H. H., Gou L., Zhang X.L., and Giles, C.L. "Discovering Missing Links in Networks Using Vertex Similarity Measures". *Proceedings of the Twenty-Seventh Annual ACM Symposium on Applied Computing (SAC'12), Trento* (2012):138–143. **DOI:** 10.1145/2245276.2245305

[2] Papadimitriou, P., Dasdan, and A., Garcia-Molina, H. "Web graph similarity for anomaly detection." *J Internet Serv Appl* 1.1(2008):1167. **DOI:** 10.1007/s13174-010-0003-x

[3] Sarno, R., and Sinaga, F. "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning". *International Conference on Computer, Control, Informatics and its Applications (IC3INA)* 2015: 12-17. **DOI:** 10.1109/IC3INA.2015.7377738

[4] Amare, Gidey. "Reviewing the Values of a Standard Operating Procedure." *Ethiopian Journal of Health Sciences* 22.3 (2012): 205–208. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3511899/

[5] Treville, S.D., Antonakis, J., and Edelson, N.M. "Can Standard Operating Procedures be Motivating? Reconciling Process Variability Issues and Behavioural Outcomes". *Total Quality Management* 16.2 (2005):231–241. **DOI:** 10.1.1.549.8898

[6] Koutra, D., Parikh, A., Ramdas, A., and Xiang, J. "Algorithms for graph similarity and subgraph matching". Technical report. Carnegie-Mellon-University (2011). **DOI:** 10.1.1.377.4291

[7] Broder, A.Z., Glassman, S.C., Manasse, M.S., and Zweig, G. "Syntactic Clustering of the Web". In Proc. of Int. World Wide Web Conf. (WWW), Apr 1997: 393-404. **DOI:** 10.1016/S0169-7552(97)00031-7

[8] Sensarma, D., and Sarma, S.S. "A Survey on Different Graph Based Anomaly Detection Techniques". *Indian Journal of Science and Technology* 8.31 (2015). **DOI:** 10.17485/ijst/2015/v8i31/75197

[9] Huda, S., Sarno, R., and Ahmad, T. "Fuzzy MADM Approach for Rating of Process-Based Fraud". *Journal of ICT Research and Applications* 9.2 (2015): 111-128. **DOI:** 10.5614/itbj.ict.res.appl.2015.9.2.1

[10] Huda, S., Sarno, R., and Ahmad, T. "Increasing Accuracy of Process-based Fraud Detection Using a Behavior Model". *International Journal of Software Engineering and Its Applications* 10.5 (2016): 175-188. **DOI:** 10.14257/ijseia.2016.10.5.16.

[11] Rahmawati, D., Yaqin, M.A., Sarno, R.. "Fraud Detection on Event Logs of Goods and Services Procurement Business Process Using Heuristics Miner Algorithm". *International Conference on Information & Communication Technology and Systems (ICTS)* 2016: 249 – 254. **DOI:** 10.1109/ICTS.2016.7910307

[12] Rahmawati, D., Sarno, R., Fatichah, C., Sunaryono, D. "Fraud Detection on Event Log of Bank Financial Credit Business Process using Hidden Markov Model Algorithm". *International Conference on Science in Information Technology (ICSITech)* 2017: 35-40. **DOI:** 10.1109/ICSITech.2017.8257082