Detecting Abnormal Patterns in Call Graphs Based on the Aggregation of Relevant Vertex Measures

Ronnie Alves[1], Pedro Ferreira[2], Joel Ribeiro[3], and Orlando Belo[4]

[1] Vale Technological Institute Sustainable Development, Brazil
ronnie.alves@vale.com
[2] Centre for Genomic Regulation, Spain
pedro.ferreira@crg.es
[3] Eindhoven University of Technology, The Netherlands
j.t.s.ribeiro@tue.nl
[4] University of Minho, Portugal
obelo@di.uminho.pt

Abstract. Graphs are a very important abstraction to model complex structures and respective interactions, with a broad range of applications including web analysis, telecommunications, chemical informatics and bioinformatics. In this work we are interested in the application of graph mining to identify abnormal behavior patterns from telecom Call Detail Records (CDRs). Such behaviors could also be used to model essential business tasks in telecom, for example churning, fraud, or marketing strategies, where the number of customers is typically quite large. Therefore, it is important to rank the most interesting patterns for further analysis. We propose a vertex relevant ranking score as a unified measure for focusing the search of abnormal patterns in weighted call graphs based on CDRs. Classical graph-vertex measures usually expose a quantitative perspective of vertices in telecom call graphs. We aggregate wellknown vertex measures for handling attribute-based information usually provided by CDRs. Experimental evaluation carried out with real data streams, from a local mobile telecom company, showed us the feasibility of the proposed strategy.

1 Introduction

Graphs have become increasingly important in modeling sophisticated structures and their interactions in a large variety of applications, ranging from chemical informatics to telecommunications [3]. Particularly in the latter, business analysts can make use of graph-based analysis for better understanding customer social behavior and thus devising proper business strategies. For instance, from a business point of view, it has been shown that it is more reasonable to retain (or maintain) existing customers rather than acquiring new ones [5]. If the company anticipates the intention of the customer to leave (typically called "churn"), proper measures can be taken to avoid such action. In the telecom

context, customers can be seen as vertices (nodes) of the network graph and the calls made between them the edges (arcs). An edge connecting two customers contains the information that for a given instant summarizes the calling pattern between them. This data structure is called a call graph being particularly large and sparse [7].

Let's assume a typical scenario where the business analyst wants to search for potential fraud situations, i.e., looking for customers presenting abnormal pattern behaviors. Chances are that their vertices act quite distinctively in the entire graph, and thus, relevant vertices, which could be acting as fraudsters should present a particular behavior. Indeed, a relevance vertex measure needs to take into consideration attribute-based and structural vertex measures. Classical graph-vertex measures usually expose quantitative information of the vertices in the graph. From the extensive literature in mining call graph patterns we highlight two works, which employ different graph measures for better understanding of call graphs [4,7]. More recently, in [2] was presented a complete survey about graph mining, covering algorithms, laws and generators. There are also other kinds of graph patterns based on frequent patterns analysis [9]. With respect to graph mining over CDRs, Cortes et al. [4,3,9] propose a data structure based on the union of small sub graphs (Top-K edges), called community of interests (COI) to handle large dynamics graphs. Different from all those previous studies we make a clear distinction between attribute-based and "purely" structural graph-vertex measures.

In this work we do not intend to compare the effectiveness of several graph-vertex measures for scoring interesting (abnormal) vertices. Rather, we propose a unified ranking strategy which makes use of enhanced classical vertex measures combining attribute-based information and graph structural information, aggregating vertex measures into a unified measure for revealing abnormal patterns in call graphs. This network was also explored in other studies on telecom fraud detection by exploring customer behavior using signatures [6] and dynamic clustering [1]. The main contributions of the proposed strategy are: 1) a dynamic model for mining evolving call graph networks, so the model can be up-to-date when new information is available; b) a set of relevant vertex measures devised for allowing attribute-based and structural evaluation of call graphs; and 3) a vertex ranking function for mining abnormal K-vertices by applying a unified strategy that aggregates distinct vertex measures of interestingness. The remainder of this paper is organized as follows. In Section 2 we present the concepts to define the structure and properties of call graphs. In Section 3 the interest measures used in this work are described. Next, in Section 4 the proposed mining strategy is explained followed by an empirical evaluation in Section 5. Conclusions and future work are provided in Section 6.

2   Evolving Call Graphs

Before presenting the proposed mining strategy we need first to highlight a few concepts to understand the problem of evaluating abnormal patterns by

aggregating attribute-based and structural graph vertex-measures in evolving call graphs.

Annotated Call Graph. An annotated call graph is a digraph $G' = (V, E, A)$, where

- V is a set of vertices or nodes,
- E is a set of ordered pairs of vertices, called directed edges,
- A is a powerset of attribute-based information that describes the edges.

Each edge $e(x, y, info) \in E$ (with $info \in A$) denotes a direct connection from x to y and contains a set of attribute values $(info)$.

Due to the dynamic nature of telecom networks, data is being obtained constantly from new calls. In order to capture and reflect the new data the concept of evolving call graph is introduced [3].

Evolving Call Graphs. An evolving call graph consists of an annotated call graph updated with new information. The weight of the new and the old information is defined by a weighting factor θ. For a given instant t, the new call graph G'_t reflects the information and the structure of the graph in the previous instant G'_{t-1} and the new information g'_t, as described in Eq. 1.

$$G'_t = \theta \cdot G'_{t-1} \oplus (1 - \theta) \cdot g'_t \tag{1}$$

The weighting factor θ models the longevity of the information, i.e., how long the information that represents a call is reflected in the graph. Since the model can be updated over time with new CDRs, we are able now to formulate the process of mining abnormal patterns (vertices or customers) in call graphs by aggregating attribute-based and structural graph vertex measures.

Problem: Mining Abnormal Usage Patterns in CDRs.
Given the information provided by a transaction database of CDRs T and the respective evolving graph G', find customers (vertices) that present an abnormal pattern. This should take into account the attribute and structure information from G'. Such aggregation model should assist the evaluation of vertex relevance in G' according to a unified vertex ranking measure.

3   Relevance Vertex Measures

A vertex is said to be relevant on a graph when its behavior distinguishes from the other vertices, either from a structural or attribute-based perspective on the entire graph. Attribute-based measures evaluate a vertex according to the information associated to its incident edges, for example call duration and billed time. Structural measures evaluate a vertex with respect to its interaction(s). A well-known structural measure used to evaluate the relevance of vertices (web

pages) in a web graph is the PageRank [8], used for instance in the Google rank-
ing scheme. In the telecom context, vertices with high page rank may reveal
clients who have a high social importance and due to their economical impor-
tance should be the target of particular inspection [7].

In this work we distinguish relevance vertex measures as purely structural
and attribute-based ones. Purely structural are vertex measures that take into
account the graph structure (e.g., degree of centrality). Attribute-based measures
can be also context-sensitive (e.g., closeness centrality) or context insensitive
(e.g., vertex-usage being explained in the next section).

3.1 Vertex Usage (M_1)

The Vertex Usage measure is based on the standard score (or z-score) used in
statistics, and indicates how much the behavior of a graph vertex (observation)
deviates from the mean of the entire graph. A vertex presenting a high vertex
usage score may be viewed as abnormal pattern, being potential indicative of
fraud or churning situations in telecom. Vertex Usage of a vertex v and an
attribute i is defined as follows:

$$M_1(v, i) = \frac{x_i - \overline{x}_i}{\sigma_i} \tag{2}$$

where $i \in info$, x_i is the observation of i in v, $\overline{x}_i$ is the mean of i, and σ_i is
the standard deviation of i. The computational cost to calculate this measure is
$O(I \cdot V)$, being I the set of attributes and V the set of vertices.

As an example of its application let's use the information of Table 1 and
assume that we want to evaluate the Vertex Usage for the entire vertices of the
graph G'_t. This table presents the statistics associated to the attributes *Air Time*
(AT) and *Charged Amount* (CA) used for the Vertex Usage score. Vertex Usage
results of each attribute, i.e., the values of each attribute for incoming, outgoing
and all edges of all vertices of the graph, are shown in Table 2.

Table 1. Example of statistics of two attributes of a call graph G'_t

Edge		Attribute-based *info*	
Origin	Dest.	AT	CA
1	2	51.00	13.60
2	3	285.00	152.05
2	4	8.25	2.40
	AVG	114.75	56.02
	STD	121.64	68.06

3.2 Degree of Centrality (M_2)

The degree of centrality is a structural measure defined as the ratio of the *in-
coming* and *outgoing* edges of a vertex, and the total number of edges on the

Table 2. Vertex usage (M_1) values of the vertices in Table 1

v	$M_1(v, AT)$			$M_1(v, CA)$		
	Ori.	Des.	All	Ori.	Dest.	All
1	-0.52	-0.94	-0.52	-0.62	-0.82	-0.62
2	1.47	-0.52	1.89	1.45	-0.62	1.65
3	-0.94	1.40	1.40	-0.82	1.41	1.41
4	-0.94	-0.88	-0.88	-0.82	-0.79	-0.79

graph. Vertices with high scores may be viewed as cases of popularity where a relationship exists with many other vertices. These relationships should not be seen as a unique factor of vertex importance since this measure only considers the links without their attribute-based information. The degree centrality of a vertex v is defined as follows:

$$M_2(v) = \frac{inDegree(v) + outDegree(v)}{count(edges)} \tag{3}$$

where $inDegree(v)$ and $outDegree(v)$ are the count of incoming and outgoing incident edges of v, and $count(edges)$ is the total number of edges in the graph. Table 3 depicts the computational process of this measure.

Table 3. Degree of centrality (M_2) values of the vertices in Table 1

	Vertex v			
	1	2	3	4
$inDegree(v)$	0	1	1	1
$outDegree(v)$	1	2	0	0
$M_2(v)$	0.33	1.00	0.33	0.33

3.3   Closeness Centrality (M_3)

Closeness centrality can be considered either as attribute-based (when using information associated to its edges) or structural-based (when counting only the existence of edges) graph vertex measure. Closeness centrality indicates how close a vertex is on average to all other vertices and it is defined as follows:

$$M_3(v, i) = \frac{\sum_{k=1}^{n} min\ Dist(v, reachableVertex(v, k), i)}{reachableVertices(v)} \tag{4}$$

where $i \in info$, $reachableVertices(v)$ is the count of reachable vertices of v, $reachableVertex(v, k)$ is the kth reachable vertex of v, and $min\ Dist(v_1, v_2, i)$ is the minimum distance between the vertices v_1 and v_2 with respect to i. Table 4 illustrates the calculation of this measure. For this measure we assume

the set of attributes $I = \{Occurrences, Air\ Time, Charged\ Amount\}$ where $Occurrences(OC)$ is the number of edges (when the number of reachable vertices is zero we directly assign this value as M_3 result).

Table 4. Closeness Centrality (M_3) values of the vertices in Table 1

	Vertex v			
	1	2	3	4
$reachableVertices(v)$	3	2	0	0
$M_3(v, OC)$	1.67	1.00	0.00	0.00
$M_3(v, AT)$	148.75	146.63	0.00	0.00
$M_3(v, CA)$	65.08	77.23	0.00	0.00

3.4 Vertex Interest (M_4)

This attribute-based measure is inspired on the PageRank measure. In telecom context, a vertex will have a high rank if receives many calls or alternatively if these calls have a high importance to the origin vertex. The vertex interest $M_4(v, i)$ of a vertex v and an attribute i is defined as follows:

$$\alpha \cdot \sum_{a=1}^{k} \frac{inEdgeValue(v, a, i)}{vertexAVG(v, i)} + \beta \cdot \sum_{b=1}^{k} \frac{outEdgeValue(v, b, i)}{vertexAVG(v, i)} \tag{5}$$

where $i \in info$, α and β are user-defined constants, $vertexAVG(v, i)$ is the average of the attribute i for all incident edges of v, $inEdgeValue(v, k, i)$ is the value of the kth incoming incident edge of v, $outEdgeValue(v, k, i)$ is the value of kth outgoing incident edge of v, a is the total number of the incoming incident edges of v, and b is the total number of the outgoing incident edges of v. An example of its application is given in Table 5.

Table 5. Vertex Interest (M_4) values of the vertices in Table 1

	Vertex v			
	1	2	3	4
$vertexMean(v, AT)$	51.00	114.75	285.00	8.25
$vertexMean(v, CA)$	13.60	56.02	152.05	2.40
$M_4(v, AT)$	1.00	3.00	1.00	1.00
$M_4(v, CA)$	1.00	3.00	1.00	1.00

4 Finding Abnormal Vertices

Since $info$ may contain several attributes, the computation of a unified relevance vertex measure can result in a set of different measures. Thus, the aggregation

function I_{AGG} to combine (aggregate) all relevance vertex attributes (information) is defined by:

$$I_{AGG}(v, I, m) = max \ [norm(M_m(v, i))] \text{ for all } i \in I \qquad (6)$$

where I is a set of attribute values, $m \in \{1, 2, 3, 4\}$ according to a relevance vertex measure, and $M_m(v, i)$ is the value of a measure of interest m of the vertex v. $norm(x) \in [0, 1]$ refers to the min-max normalization of x. Table 6 illustrates the I_{AGG} calculation.

Table 6. Aggregation of the attributes *Air Time* and *Charged Amount* with the measure M_3, according to Table 1

	Vertex v			
	1	2	3	4
$M_3(v, AT)$	148.75	146.63	0.00	0.00
$M_3(v, CA)$	65.08	77.23	0.00	0.00
$norm(M_3(v, AT))$	1.000	0.986	0.000	0.000
$norm(M_3(v, CA))$	0.843	1.000	0.000	0.000
$I_{AGG}(v, \{AT, CA\}, 3)$	1.000	1.000	0.000	0.000

Considering all vertices v, $I_{AGG}^k(I, m)$ refers to the top-k $I_{AGG}(v, I, m)$ values. Taking into account only the Top-K cases, it is possible to identify the most interesting vertices through Eq. 7.

$$I_{AGG}^k(v, I, m) = \begin{cases} I_{AGG}(v, I, m) & \text{if } I_{AGG}^k(I, m) \text{ contains } v \\ min \ I_{AGG}^k(I, m) & \text{otherwise} \end{cases} \qquad (7)$$

Telecom call graphs are usually big figures for being explored at one shot. Therefore, one should be able to focus on particular spots of the entire graph. To do so, the graph composition function M_{AGG} for measure aggregation according the Top-K vertices is defined (Eq. 8). Such function compares a specific vertex with the Top-K vertices for each relevance vertex measure using all vertex measures at once, being possible to constraint the search of (Top-K) abnormal patterns in the entire graph.

$$M_{AGG}(v, I, k) = \prod_{m=1}^{n} I_{AGG}^k(v, I, m) \qquad (8)$$

Remark that only Top-K values of each measure are used. An example of M_{AGG} calculation is presented in Table 7 (it is assumed that $k = 2$), where vertices 2 and 3 are the most interesting ones.

The first step to calculate the M_{AGG} score is evaluating I_{AGG}^2 values for each measure $m = \{1, 2, 3\}$. For Vertex Usage ($m = 1$) we refer to Table 2 to observe all scores (see column "All") for all vertices, where I_{AGG}^2 is evaluated as follows:

- $I_{AGG}^2(v = 1, \{Air\ Time, Charged\ Amount\}, 1)$
  $= max(0.130, 0.070) = 0.130$
- $I_{AGG}^2(v = 2, \{Air\ Time, Charged\ Amount\}, 1)$
  $= max(1.000, 1.000) = 1.000$
- $I_{AGG}^2(v = 3, \{Air\ Time, Charged\ Amount\}, 1)$
  $= max(0.823, 0.902) = 0.902$
- $I_{AGG}^2(v = 4, \{Air\ Time, Charged\ Amount\}, 1)$
  $= max(0.000, 0.000) = 0.000$

Finally the two highest scores for Vertex Usage are $I_{AGG}^2 = \{1.000, 0.902\}$. The I_{AGG}^2 scores for Closeness Centrality ($m = 3$; Table 4) are $I_{AGG}^2 = \{1.000, 0.986\}$. The Vertex Interest scores ($m = 4$; Table 5) are $I_{AGG}^2 = \{1.000, 0.333\}$. The M_{AGG} for $v = 1$ is then evaluated as:

- $M_{AGG}(v = 1, \{Air\ Time, Charged\ Amount\}, 2)$
  $= 0.902 \times 1.000 \times 0.333 =$
  $= 0.300$

Table 7. Measure (M_{AGG}) and information (I_{AGG}) aggregation for the measures {Vertex Usage, Closeness Centrality, Vertex Interest} and the attributes {*Air Time*, *Charged Amount*}, according to Table 1

	Vertex v			
	1	2	3	4
$I_{AGG}^2(v, \{AT, CA\}, 1)$	0.130	1.000	0.902	0.000
$I_{AGG}^2(v, \{AT, CA\}, 3)$	1.000	0.986	0.000	0.000
$I_{AGG}^2(v, \{AT, CA\}, 4)$	0.333	1.000	0.333	0.333
$M_{AGG}(v, \{AT, CA\}, 2)$	**0.300**	**0.986**	0.296	0.296

Table 8. Statistics about all call graphs in the related week sample. Each sample corresponds to a particular day.

Sample	Vert.	Edges	Comp.	Diam.	Path	AvgNeig.
1	35726	20356	15564	2	1.014	1.131
2	22886	12270	10688	3	1.010	1.067
3	22377	11896	10531	3	1.007	1.059
4	21743	11508	10287	3	1.005	1.054
5	21956	11598	10426	2	1.008	1.051
6	22100	12016	10142	2	1.005	1.083
7	20380	11234	9237	2	1.004	1.097

5   Detecting Potential Fraud Situations in Call Graphs

In this section we present a case studied using real data streams from a mobile telecom company. The main goal was to highlight potential fraud situation using

the proposed strategy. It was provided a list of fourteen fraud cases obtained from a specific week of CDRs. For each day of the week there are approximately 2.5 millions of records (CDRs) and 700,000 customers. In this empirical study only around 5% of the entire dataset containing both fraud (fourteen situations) and potential unidentified cases were selected for furthers analysis. Table 8 provides statistics about the call graphs obtained for each day of the week. As one can observe these graphs are quite sparse, being a great challenge the detection of abnormal call patterns.

Table 9. Results of the vertex ranking on the final evolving call graph for the known cases

	Blacklist Cases													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
M_1						16		6	62	33	84	82	5	10
M_2		67	54	31		24		38	50		31	21		67
M_3														
M_4		71	46	34		23		36	48		39	21		73
M_{AGG}			49	51		16		17	30		26	23	50	42

Further discussion about variables related to this dataset can be found in previous works [1,6]. In order to assess effectiveness, all relevance vertex measures are computed for all vertices of the 5% sample. Then, it was verified whether fraud cases are in the Top-100 results or not. The Degree of Centrality is able to highlight eleven cases of fraud maybe due to the increasing number of calls in that week. On the other hand the Closeness Centrality measure does not detect any case of fraud consequence to the lower diameter and average neighborhood of the call graphs (Table 8).

Table 10. Results of the vertex ranking on the daily call graphs for the known cases

	Blacklist Cases													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
M_1					69	6		8	7	11	40	40	2	1
M_2	27	11	15	9		5		11	15	29	14	9	37	20
M_3					52		45	74		5				6
M_4	26	13	14	12		4		10	16	42	12	9	31	23
M_{AGG}	41	19	19	12		3		5	6	22	8	9	11	5

Tables 9 and 10 present the results of applying all relevance measures employed in the proposed aggregation strategy. For each of the fraud cases, it is identified their ranking in the Top-100 of the different measures. Table 9 refers the application of vertex ranking taking into account the preference selection function (Eq. 8) on the final evolving call graph (i.e., aggregating all samples).

Table 11. Results of the detection of abnormal patterns using different graph-based metrics. I = Information, S = Structure, B = Both

	Blacklist Cases													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Evolving call graph		B	B	B		B		B	B	I	B	B	I	B
Daily call graphs	B	B	B	B	S	B	S	B	B	B	B	B	B	B

Similarly, Table 10 refers the application of vertex ranking but for each of the samples (i.e., each sample as a separate call graph). Each value represents, for a given case and measure, the highest ranking for the different samples.

The results suggest that there are different types of fraud situations. One of our conclusions is that some cases can be grouped according their rankings. Probably different groups identify distinct types of fraud and should be differently handled by the fraud analysts. One strategy should be the selection of other similar sub-graphs based on such groups identified by the proposed model [10] [11]. Indeed, for detecting abnormal patterns in weighted call graphs one should not set aside attribute-based information about the calls. This observation explains why Vertex Usage and Vertex Interest are more sensitive to this type of problem.

The application of the vertex ranking on the final evolving call graph (Table 9) identified successfully 9 out of 14 (around 65%) of the given cases as high potential fraud cases. The same application on the daily call graphs (Table 10) improved the detection rate to 86%. Finally in Table 11 we summarized the fraud detection analysis. The conclusion is that the evolving call graph (G'_t) should be taking into account together with the daily call graphs (g'_t).

6 Conclusions

In this work we have presented enhancements on well-known graph-vertex measures in order to improve selection and ranking of abnormal patterns over telecom call graphs. We extend classical quantitative vertex measures with attributed-based ones, proposing a unified vertex ranking for detecting abnormal vertices in weighted graphs. An empirical study using CDRs from a real mobile telecom company showed us the feasibility of the proposed strategy, while recovering most of the potential fraud situations.

References

1. Alves, R., Ferreira, P.G., Belo, O., Lopes, J., Ribeiro, J., Cortesao, L., Martins, F.: Discovering telecom fraud situations through mining anomalous behavior patterns. In: Proceedings of the DMBA Workshop, on the 12th ACM SIGKDD (2006)
2. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Comput. Surv. 38(1), article 2 (June 2006)

3. Cortes, C., Pregibon, D., Volinsky, C.: Communities of interest. Intelligent Data Analysis 6, 211–219 (2002)
4. Cortes, C., Pregibon, D., Volinsky, C.: Computational methods for dynamic graphs. Journal of Computational and Graphical Statistics 12, 950–970 (2003)
5. Euler, T.: Churn Prediction in Telecommunications Using Mining Mart. In: Proceedings of the DMBiz Workshop, on the 9th European Conference on Principles and Practice in Knowledge Discovery in Databases, PKDD 2005 (2005)
6. Ferreira, P., Alves, R., Belo, O., Cortesão, L.: Establishing Fraud Detection Patterns Based on Signatures. In: Perner, P. (ed.) ICDM 2006. LNCS (LNAI), vol. 4065, pp. 526–538. Springer, Heidelberg (2006)
7. Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., Joshi, A.: On the structural properties of massive telecom call graphs: findings and implications. In: Proceedings of CIKM 2006, pp. 435–444 (2006)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab. (1999)
9. Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., Giles, C.L.: Winners don't take all: Characterizing the competition for links on the web. In: Proceedings of Proc. Natl. Acad. Sci. USA, pp. 5207–5211 (2002)
10. Akoglu, L., McGlohon, M., Faloutsos, C.: `oddball`: Spotting Anomalies in Weighted Graphs. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS, vol. 6119, pp. 410–421. Springer, Heidelberg (2010)
11. Aggarwal, C.C., Zhao, Y., Yu, P.S.: Outlier detection in graph streams. In: Proceedings of ICDE 2011, pp. 399–409 (2011)