A Quantitative Analysis-based Algorithm for Optimal Data Signature Construction of Traffic Data Sets

Department of Computer Science (Algorithms and Complexity Laboratory), University of the Philippines Velasquez Ave., Diliman, Quezon City 1101, Metro Manila, PHILIPPINES
Email: {jamalinao, hnadorna}@dcs.upd.edu.ph

Abstract—In this paper, a new set of data signatures is derived to obtain better Vector Fusion 2D visualizations of a time series and periodic nD traffic data set as compared with previous work. The latter had used the entire Power Spectrum components for visualization purposes to produce 2D representations of each subset of the data. With the feasibility of obtaining a smaller representation of the data set in obtaining better cluster models compared to using the original n-dimensions, we now explore this feasibility for visualization purposes. We propose an algorithm that determines, in quantitative terms, how good the selected set of signatures represents the nD data set in 2 dimensions. We use the Vector Fusion visualization algorithm in transforming each signature from its n dimensions into 2 dimensions. An improved set of qualitative criterion is drawn to measure the goodness of the 2D data signature-based visual representation of the original nD data set. Finally, we provide empirical testing and discuss the results.

Index Terms—Data Signatures, Vector Fusion, iDIRBrG, Power Spectrum

I. INTRODUCTION

A data signature, as defined in [1], is a mathematical data vector that captures the essence of a large data set in a small fraction of its original size. It had been shown in previous studies[2][3] that Fourier-based data signatures employed on time series traffic data sets provide better characterization on sets of traffic flow behavior and unravel previously unknown information from the data set. In particular, these studies show the effectiveness of using such type of signatures to produce an optimal cluster model from the 2006 North Luzon Expressway (NLEX) Balintawak-Northbound (BLK-NB) traffic volume data set. The data set had to be preprocessed, partitioned, and projected as discrete input time domain signals. Each signal, representing a week with 168 hourly traffic volume entries, is then decomposed through the Fast Fourier Transform and its corresponding set of Power Spectrum components computed. A data signature is obtained by selecting the first 85 components to represent each week in its frequency domain. The X-Means[4] clustering algorithm was used to group all similar weeks and extract a number of outliers by using these data signatures.

Shown in Figure 1 is the time domain visualization of the data set with rows (representing weeks of 2006) structured contiguously conforming to the cluster model produced through X-Means on the data signatures, denoted as XMeans(F85). The horizontal axis reflects the 168 hourly total traffic volume of each week from Sunday to Saturday. Each pixel is colored based on the current traffic volume of a time step in a week. This image is produced using the Iterative Data Image Rotated Bar Graph (iDIRBrG)-based approach in [5].

Figure 1. iDIRBrG-based[5] visualization of the time-domain data set. The y-axis is composed of weeks rearranged according to the results of frequency-domain clustering (using the 85-dimensional data signatures) through X-means. The lines separate the clusters from one another.

Using Figure 1 of the time domain data set, it is used to determine inter-cluster and intra-cluster similarities and differences. Outliers are also easily identified with the significant changes highlighted in various sections in their rows. In addition to these results, analysts are also capable of mining out weeks which belong to a cluster that possess peculiar behavior among its cluster co-members. These weeks are referred to as potential outliers in [3]. Cluster and outlier analysis using this
cluster model is also detailed in the same paper. However, using a 2D Vector Fusion (VF) visualization technique [7][8], a similar cluster, outlier, and potential outlier analysis can also be accomplished in a more simplified and straightforward manner than the iDIRBrG-based visualizations. We initially used the 168-dimensional Power Spectrum components of each week and use VF to obtain a scattergram visualization of the data set. Colors used on the projected weeks in the scattergram are associated to specific clusters produced from the X-Means(F,85).

Shown in Figure 2 shows the Vector Fusion Visualization [7] VF(85, 168) where each point (representing a week) are colored using the results of XMeans(F,85). The notation VF(X,Y) means that we have used the first X and Y components of the Power Spectrum as inputs to the X-Means clustering algorithm and the VF visualization algorithm, respectively. In the latter part of the paper, the second component, i.e. Y may also be a set of elements \( i \) where \( i \in \{1,2,\ldots,168\} \) associated to a Power Spectrum component \( A_i \).

With the use of the entire set of Power Spectrum components of each week, a “good enough” VF visualization is obtained from the data set. However, a previous work in [3] has shown the possibility of obtaining a signature from a small subset of the nD Power Spectrum components yet obtaining a good cluster model of the data set was achievable. This model, in fact, was concluded as a better one when compared to the model obtained by using the original nD Power Spectrum of the data set. Thus, this study aims to produce a similar set of signatures for visualization purposes. This study also aims to improve the VF visualizations in terms of the qualitative and quantitative goodness-of-visualization criteria by using the signatures themselves as input to the VF algorithm. Furthermore, there are also some potential outliers which are readily identifiable in Figure 1 that are not projected relatively far enough from their co-members. Even though these points were not considered as outright outliers by X-Means due to a significant similarity with their co-members yet they are still worthy of further investigation. They may be telling of noteworthy actual traffic incidents in the expressway providing more novel information regarding the data set.

The initial investigation performed clearly shows that the results of the X-Means clustering algorithm and the time domain visualizations are highly needed to establish quality traffic data analysis to verify the observations derived from the VF visualizations. The goal of this paper now is to determine a way to improve the reliability of the VF visualizations 2D representation of the original nD data set. By doing so, we provide analysts visualizations that are simplistic, intuitive, abstract, yet readily-interpretable representations of large nD traffic data sets. In contrast to the iDIRBrG-based visualizations, point-to-point, intercluster, and intracluster relationships inherent in the data set are easily established.

In Section II, we give definitions and notations to the concepts building the theoretical backbone of this study. Provided in Section III are the details on how “good” visualizations are to be obtained through qualitative and quantitative approaches. We further strengthen our qualitative results by introducing an algorithm to obtain a quantitative value measuring the consistency of the visualization of the vector-fused data signatures with respect to their actual Euclidean distances. In Section IV, we give insights and characterizations of Power Spectrum values so as to select reasonable components for data signature construction. Finally, we show empirical tests on various data signatures and detailed discussions on their results in Section V and formulate conclusions in Section VI.

II. BASIC DEFINITIONS AND NOTATIONS

A. The Data Set

The data sets used in this work have the following characteristics: time-series, periodic and multidimensional. In particular, we used records in the 2006 NLEX Balintawak Northbound (BLK-NB) data set provided by the National Center of Transportation Studies (NCTS). A record contains hourly entries accumulated via an automatic detector embedded in every lane of NLEX’s segments. The detector inserts the mean spot volume in its record for each lane per hour, thus, 168 data points are collected in each week. We totaled these values in all four lanes to summarize the data set as a 52 \( \times \) 168 data matrix, i.e. 52 weeks with 168 data points each.

B. Power Spectrums

Fourier descriptors such as Power Spectrums rely on the fact that any signal can be decomposed into a series of frequency components via Fourier Transforms. By treating each nD weekly partitions in the NLEX BLK-NB time-series traffic data set as discrete signals, we obtain their Power Spectrums through the Discrete Fourier Transform (DFT) decomposition as shown below,

\[
\theta(t) = \mu_0 + \sum_{k=1}^{n-1} (a_k \cos \frac{2\pi k}{n} - i b_k \sin \frac{2\pi k}{n}),
\]

\( \theta(t) \) is the signal, \( \mu_0 \) is the signal mean, \( a_k \) and \( b_k \) are the real and imaginary parts of the real Fourier coefficients respectively.
where \( \mu_0 \) is the component referred to as the offset of the signal translated from the horizontal axis. Using DFT, a vector of real numbers can produce a vector \( F \) of frequency components of the same length, where \( F = (a_0 \pm b_0 i, a_1 \pm b_1 i, a_2 \pm b_2 i, \ldots, a_n - 1 \pm b_n i, -1) \), \( \mu_0 = a_0 \pm b_0 i \) and \( i^2 = -1 \). For the resulting \( n \)D vector, we produce distinct values \( a_0 \pm b_0 i, a_1 \pm b_1 i, a_2 \pm b_2 i, \ldots, a_n/2 \pm b_n/2 i \) and the succeeding values are their complex conjugates. If \( n \) is even, the \( \left( \frac{(n/2)}{2} \right) \)th component will be real.

Power Spectrum is the distribution of power values as a function of frequency. For every frequency component, power can be measured by summing the squares of the coefficients of the corresponding sine-cosine pair and then getting its square root. The Power Spectrum \( A_k \) of the signal, \( k = 0, 1, \ldots, n - 1 \) is given by,

\[
A_k = \sqrt{a_k^2 + b_k^2}.
\]

C. Vector Fusion Visualization

In his work[7], Robert Johnson introduces a method of providing a 3D perspective of any given \( n \)D data vector. It utilizes the Single-point Broken-line Parallel coordinates (SBP) algorithm to achieve such a representation. Each instance in a given multidimensional data set is projected in 3D as a vector resultant of the components of that instance. The paper[8] simplifies this visual representation into a 2D visualization of the points. In the paper, an \( n \)D data point \( w = [w_1, w_2, \ldots, w_n] \) is represented as a 2D resultant point in a scattergram projected by summing all of data point’s component \( w_j \) projected using a precomputed angle \( \theta_j \) with \( w_{j-1}, j = 1, 2, \ldots, n \). Shown below is the formula to compute the 2D coordinates \((SXPB_x, SYPB_y)\) for an \( n \)D data point \( w \).

\[
w = w_1 e^{i \theta_1} + w_2 e^{i \theta_2} + \ldots + w_n e^{i \theta_n}
\]

\[= \sum_{j=1}^{n} w_j \cos(\theta_j) + i \sum_{j=1}^{n} w_j \sin(\theta_j)
\]

\[= (w_{sumX}, w_{sumY}) = (SXPB_x, SYPB_y)
\]

where \( \theta_j = (j-1)180^\circ/n \), \( n \) is the dimension of the vector, and \( w_j \) is the value of the \( j \)th dimension.

III. METHODOLOGY

A. A Qualitative Goodness Measure of the VF Visualization

After obtaining the set of Power Spectrum components for each \( n \)D week in the data set, these components shall be selected and used as its data signature for VF visualization purposes. Note that \( m \leq n \). To obtain relationships of the weeks using the scattergram visualization, we shall use the cluster model XMeans(F,85) to color each scatter point. Then, we determine how well the 2D scattergram represents the pre-computed point-to-point, intercluster, and intracluster relationships obtained from XMeans(F,85). Shown below is the criteria[3] that we have improved in this work.

1) Closeness of co-members A good visualization should show reasonable visual proximity of points belonging to a common cluster. Regions occupied by clusters should have minimal overlaps in the visualization.

2) Visibility of all points No total occlusion should exist for any given set of points.

3) Outlier detection Outliers seen after using X-Means clustering algorithm should have a significant distance from all other weeks such that they can easily be pinpointed in the visualization.

4) Detection of potential outliers Potential outliers, i.e. weeks within a cluster that show “interesting” behavior as seen in the iDIRBrG-based visualizations, should be found near or at the periphery of the region occupied by a cluster. These should be projected far from their co-members in the VF visualization. The detection of potential of outliers by use of projection of convex hulls along the periphery of its cluster region and determining whether they are “far enough” from their co-members is a highly subjective process. Candidate potential outliers may exist along the convex hull but may be significantly spatially near from its cluster centroids compared to other points that may have a larger spatial distance but are not along with the hull. Thus, a previous work [6] formalized the definition of potential outliers using regression curves, confidence bands, and confidence ellipses. We shall use this method to provide us a list of these points in our empirical tests.

5) Characterizing relationship across clusters A good visualization should aid users in efficiently determining on what characteristics one cluster differ from other clusters in the data set.

6) Consistency of the 2D geometric representation of the data points Suppose a point \( R \) has a smaller data signature Euclidean distance from a point \( S \) compared to a point \( Q \). Then, the (SBPx,SBPy) Euclidean distance of \( R \) to \( S \) must also be smaller compared with \( R \)'s distance to \( Q \).

Exploring the first five criterions is easily done. Figure 2 in Section I was shown in the previous work[3] to be the best VF visualization of the 2006 NLEX BLK-NB traffic volume data set based on these first 5 criterions. In this paper, we propose an algorithm in the succeeding subsection that checks the consistency stated in the last criterion and state our results.

B. Exploring Criterion 6 via Benchmark Model \( M_1 \) and VF Test Model \( M_2 \)

Criterion 6 requires us to initially model how points, represented by their data signatures, relate to one another in terms of their Euclidean distances. This is the benchmark model \( M_1 \).

To build the benchmark model \( M_1 \), let \( R = [r_0, r_1, \ldots, r_m] \) and \( S = [s_0, s_1, \ldots, s_m] \) be data signatures from two arbitrary data points from the data set \( D \). Let \( \delta(R, S) \) be the Euclidean Distance from \( R \) to \( S \), i.e. \( \delta(R, S) = \sqrt{n \sum_{i=1}^{m} (r_i - s_i)^2} \).

By computing all the Euclidean distances of all data signatures in \( D \), \( M_1 \) shall have a distance matrix for all the points of \( D \).
We build the test model $M_2$ by initially using the VF visualization algorithm to generate the $(SBPx, SBPy)$ 2D representation of the $mD$ data signature in $D$. Let $R' = (SBPx_1, SBPy_1)$ and $S' = (SBPy_2, SBPy_2)$ be the vector-fused data signature $R$ and $S$, respectively. Let $\delta(R', S')$ be the Euclidean Distance from $R'$ to $S'$. Then, compute all the Euclidean distances of all vector-fused data signatures in $D$ to obtain a distance matrix of $D$ for $M_2$. Finally, using the distance matrices of $M_1$ and $M_2$, we compare how consistent is $M_2$’s 2D representation of the original $mD$ signatures of $M_1$ by using the algorithm below:

1) Let $N$ be the number of weeks in the data set $D$. For each week $R$ in $D$, create a list $L_R$ containing every other week $S \in D$ arranged from the smallest to the largest Euclidean Distance $\delta(R, S)$ from $R$ of $M_1$.

2) Get the maximum distance $MaxD$ which is equal to the distance from $R$ to the last week in the list $L_R$. Let $d = MaxD/N$. Let $P(i)$ be the set of weeks in $D$ in the $i^{th}$ partition in the list $L_R$, i.e. the set of weeks found at the distance $q, q \in [i * d, (i + 1) * d]$, where $i = 0, 1, \ldots, N - 1$. Let $Count(i)$ be equal to the number of weeks in $P(i)$, where $i = 0, 1, \ldots, N - 1$.

3) Create a list $L'_R$ containing every other week $S \in D$ ranked from the one with the closest Vector-Fusion Visualization Euclidean distance $\delta(R', S')$ in $M_2$ to the farthest (with respect to $R$).

4) Let $P'(i)$ of $M_2$ be the $i^{th}$ partition in the list $L'_R$ containing the set of weeks ranked from $r$, where $\forall r$,

\[
r \in \left(\sum_{j=0}^{i-1} |P(j)| + 1, \sum_{j=0}^{i} |P(j)| \right).
\]

5) For each partition $i$, $i = 0, 1, \ldots, N - 1$, compute for the number of matches at the $i^{th}$ partition of $M_1$ using $R$, denoted as $M(R, i)$, where $M(R, i) = P(i) \cap P'(i)$. Then, compute for the errors in $M_2$ with respect to $M_1$ at the $i^{th}$ partition using $R$, denoted as $A(R, i)$,

\[
A(R, i) = \frac{\sum_{S : S \in P(i')} |i - i'|}{N},
\]

where $S \in P(i')$ of $M_2$. Finally, compute for the consistency of $M_2$ with $M_1$ using $R$ on the $i^{th}$ partition, denoted as $Cons(R, i)$, where

\[
Cons(R, i) = M(R, i) - A(R, i).
\]

6) Compute for the overall consistency of $M_2$ with $M_1$ using $R$ as $OC(R)$,

\[
OC(R) = \sum_{i=0}^{N-1} Cons(R, i).
\]

Finally, compute for the model consistency of $M_2$ with the benchmark model $M_1$ using all data points in the data set $D$, denoted as $MC$, were

\[
MC = \sum_{R \in D} OC(R).
\]

In this paper, the algorithm builds the model for which we use the original $mD$ data signatures of all weeks in the data set as illustrated in Figure 3. The algorithm determines and stores information regarding the relationships of the weeks of the data set expressed in their data signatures. Then, it builds $M_2$ that checks this relationship but in the 2D space produced during the application of the VF algorithm on these signatures.

![Figure 3. Benchmark and Test Models M1 and M2, respectively.](image)

As illustrated in Figure 3, the relationship of week $R$ among all other points in the data set in $M_1$ may become distorted as the 2D projections of the signatures in $M_2$ are determined.

In order to determine the aforementioned distortion of the relational values of the signatures of all the weeks, the algorithm measures how accurately each week $R$ is modeled in $M_2$ against $M_1$’s. Then, the algorithm projects $M_1$’s space where $R$ is projected as its centroid and all other points scattered around $R$ based on their Euclidean distance in the data signatures’ $mD$ space. Then, partitions are determined so that ranking is established among the weeks with respect to their relationship with $R$. Note that this partitioning results to concentric circles formed from $R$ with radius increasing by a factor $d$ for $M_1$. Then, $M_2$’s space is built, again projecting $R$ as its centroid and all other weeks scattered around $R$ by taking into consideration their Euclidean distance in the 2-dimensions $(SBPx, SBPy)$. Using the information derivable from the partitioning of $M_1$, we build partitions in $M_2$ and ranking of the weeks with respect to $R$ in this space. From the partitions built for both models, the algorithm will determine the consistency and error of the projections of all other weeks with respect to $R$ in $O(N^2)$ time and space complexity, where $N$ is the number of weeks in the data set analyzed.

A perfect representation of $M_2$ of $M_1$ would have every $R$ and all other weeks placed in their correct partitions for both models. In such case, the algorithm outputs the maximum quantitative analysis value of $N^2$.

IV. OBTAINING AN OPTIMAL DATA SIGNATURE FOR VECTOR FUSION VISUALIZATION

In previous literature[2], [3], Power Spectrum-based data signatures of each week in the 2006 NLEX BLK-NB traffic...
volume data set were used in obtaining optimal cluster models through the X-Means clustering algorithm. An optimal cluster model can be obtained by using the first 85 components of the Power Spectrums of each input rather than using all of them. Shown in Figure 4 are the values of the first 85 components of the Spectrum. When projecting the last few components, we just obtain a mirror image of the visualization below.

As seen in Figure 4, almost all weeks had the 7th dominant Power Spectrum component (also known as harmonic). The first 7 harmonics show significant variations compared to the succeeding values. Note the 42nd harmonic shows a significant increase of the value from the normally-decreasing values of previous values of Harmonics 7, 14, 21, 28, and 35. Most weeks already had low values harmonics 21, 28, 35 and increases at the 42nd harmonic. The succeeding harmonics already shows values converging to zero. Thus, it is reasonable to use harmonics 7, 14, . . . , 42 and the offset as candidate components of the data signatures of the weeks in the N'X data set for visualization purposes.

A closer look at the candidate harmonics in Figure 5 using a set of sampled weeks from different clusters of the data (inclusive of outliers), it can be seen that the most variation of the Power Spectrum values are in \( A_7, A_{14}, \) and \( A_{21} \). Thus it is also interesting to obtain a visualization of the data set using the data signature composed of the Power Spectrum components \( A_0, A_7, A_{14}, \) and \( A_{21} \). Finally, note that Week 15 has its first harmonic to be dominant. A relatively large and peculiar set values for the first few harmonics of Week 44 can also be seen in Figures 4 and Figure 5. By taking advantage of Weeks 15 and 44 non-conforming dominant harmonics, we can choose to use components of factor 7 as data signature, thus highlighting Weeks 15 and 44 as apparent outliers in the entire data set.

V. RESULTS AND DISCUSSION

By implementing the algorithm on the benchmark and test models \( M_1 \) and \( M_2 \) using varied data signatures of each data point in \( D \), the following values for Model Consistency (MC) are obtained as shown in Table I.

The known-best visualization \( VF(85,168) \) [3] clearly is being defeated by the proposed set of Power Spectrum components \( \{0, 7, 14, 21\} \). Shown in Figure 6 is the vector fusion visualization \( VF(85,\{0,7,14,21\}) \).

To support the qualitative results, we check whether \( VF(85,\{0,7,14,21\}) \) is better compared with \( VF(85,168) \) in the qualitative criteria.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Consistency of VF Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( VF(85,168) )</td>
<td>104.00</td>
</tr>
<tr>
<td>( VF(85,85) )</td>
<td>104.00</td>
</tr>
<tr>
<td>( VF(85,43) )</td>
<td>94.00</td>
</tr>
<tr>
<td>( VF(85,{0,7,14,21}) )</td>
<td>300.00</td>
</tr>
<tr>
<td>( VF(85,{0,7,14,21,28,35,42}) )</td>
<td>207.00</td>
</tr>
</tbody>
</table>

In terms of Criterion 1, \( VF(85,168) \) slightly surpasses the quality of \( VF(85,\{0,7,14,21\}) \) by an outlier Week 18 (i.e. Cluster 5). Nevertheless, both visualizations had minimal overlaps in some of their clusters such as Clusters 1 and 0. In general, closeness of co-members are maintained for both VF models. Both models also satisfy Criterion 2. In terms of Criterion 3, known outliers, i.e. Weeks 15, 51, and 44 are easily seen in both visualizations. It is notable that both Weeks 44 and 18 seem to have similarities with Cluster 4 in both VF visualizations. However, \( VF(85,168) \) had clearly projected Week 18 far enough from this cluster as compared to \( VF(85,\{0,7,14,21\}) \). The latter had actually placed this outlier within Cluster 4’s region while the former projected it along the edge of the confidence ellipse of the cluster.

In the detection of potential outliers, we had to refer back to the time domain iDIFRBrG visualization in Figure 1 to obtain a set of this points by intra-cluster analysis which we identified as Weeks 1, 30, 31, 43, and 48. It should be noted that a previous result in [2] has shown that Week 30 has the
set of the smallest traffic volume values for year 2006. By using definitions of potential outliers and categories thereof as defined in [6], it can be observed that VF(85, \{0,7,14,21\}) slightly outperforms VF(85,168) by the former’s capability of detecting Week 31 as a potential outlier. However, for both models, Week 43 has not been detected as this type. As for the fifth criterion, it can be seen that both models clearly projects clusters from the leftmost to the rightmost parts of the visualization in terms of ascending magnitudes of traffic volume values. In summary, for the first five qualitative goodness-of-representation, the two VF visualization models clearly has a competitive quality in terms of representing the relationships of the data signatures in their original mD space in the transformed 2D \((SBP_x, SBP_y)\) space.

VI. CONCLUSIONS

In this paper, we were able to obtain an optimal data signature more effective in representing points in the data set for Vector Fusion visualization purposes compared with using the entire set of Power Spectrum components. We added another qualitative criterion, i.e. Criterion 6, to further check the goodness of the data set visualization. An algorithm was formulated to check how each proposed vector-fused data signature visualization fairs with this criterion. Different data signature constructions were formulated and checked through the algorithm with results showing that the Power Spectrum components \(A_0, A_7, A_{14}, \text{ and } A_{21}\) provides the best VF visualization quantitative value among the evaluated models in this paper for the 2006 NLEX BLK-NB traffic volume data set. Further validation of this model and the known best VF(85,168) has also shown competitiveness of the model in terms of the first five criterions. Thus, the quantitative analysis-based algorithm had shown that optimal models for 2D representations of high dimensional data sets (or signatures) can still be built without sacrificing the amount of information analysts can derive from the visualizations. In fact, this study has shown that we had reduced the original 168-dimensional frequency domain representation of the data set into a 4-dimensional data signature being Vector Fused into a 2D point. This first reduction accounts for a total of \(\approx 97.6\%\) Power Spectrum values being dropped from the representation of weeks in the original data set. This reduction is crucial when analysts apply any additional exploratory data mining techniques such as clustering. Finally, with reliable, yet simple, 2D visuals produced by use of Vector Fusion and data signatures, analysts are now capable of pinpointing weeks that may exhibit “interestingness” due to their spatial distance from all other points (or co-members) in the visualization. Since Vector Fusion independently computes for a 2D representation of an input vector, new sets of data points (weeks) are easily projected into the visualization after signatures were computed from them. Finally, analysts would then be able to determine whether this new points are worthy of further scrutiny.

ACKNOWLEDGMENT

This work is funded by an ERDT-DOST project entitled “Information Visualization via Data Signatures”. The authors would like to thank Roberto V. Bontia, VP-Toll Operations, of Manila North Tollways Corporation (MNTC), and Nicolas Manalo, VP Traffic Operations, of Tollways Management Corporation (TMC) for providing the research team the NLEX data set. E.R. Oquendo would like to thank the University of the Philippines Visayas for his Fellowship Grant. J. Clemente, R.A. Juayong, and R.M. Tadlas are being assisted in their graduate studies by the ERDT Scholarship. J.B. Lee would like to thank the University of the Philippines Information Technology Training Center (UP-ITTC) for his graduate studies scholarship. J. Malino and H. Adorna is supported partially by DOST- PCASTRD through the ERDT project entitled Information Visualization via Data Signatures.

REFERENCES