

Local Interpolated Compressive Sampling for Internet Traffic Reconstruction

¹Indrarini Dyah Irawati, ²Andriyan Bayu Suksmono, Senior Member, IEEE, ³Ian Yosef Matheus Edward

School of Electrical and Informatics, Institut Teknologi Bandung

Bandung, Indonesia

¹indrarini@telkomuniversity.ac.id, ²suksmono@stei.itb.ac.id, ³ian@stei.itb.ac.id

Abstract— In this paper, we propose the integration of local interpolation on compressive sampling (CS) for application on internet traffic reconstruction. CS managed to solve the problem of missing Internet traffic but not in the case of extreme missing like Missing Elements at Random (MER) and Combine Missing Patterns (CMP). To overcome the problem is done by merging between local interpolations with CS. In this study conducted a comparison of local interpolation using correlation method and Euclidean norm and tested the effect of similarity parameters between rows and similarity between columns. The simulation results show that the addition of local interpolation can improve accuracy.

Keywords—local interpolation; compressive sampling; internet traffic matrix; correlation; Euclidean norm

I. INTRODUCTION

The problem of missing internet traffic becomes area that has been researched since 2006 by Roughan et al [1]. Roughan et al. explored the impact of six loss models, such as PureRandLoss, xxTimeRandLoss, xxElemRandLoss, xxElemSyncLoss, RowRandLoss, ColRandLoss on the performance of interpolation algorithms. They solved the problem by incorporating between spatio temporal CS using Sparsity Regularized Matrix Factorization (SRMF) and K-Nearest Neighbor (KNN) that performs superior for all loss models. In [2], Huibin et al. presented Self-Similarity and Temporal Compressive Sensing (SSTCS) algorithm for reconstructing lost traffic data consisting of Frequent Loss in Row (FLR), Successive Loss in Row (SLR), Frequent Loss in Column (FLC), Successive Loss in Column (SLC), Row Random Loss (RRL), Column Random Loss (CRL). This algorithm can retrieve the lost data with error less than 32% when data lost as much as 98%. In another research [3], the authors evaluated the effect of six missing patterns like as the missing problems on the actual network. The paper reported that CS reconstruction algorithms failed to recover the missing values for high loss, especially in the Missing Elements at Random (MER) and Combine Missing Patterns (CMP).

One of the simplest method for correcting the missing data is interpolation [1, 4, 5, 6]. It can solve the missing problems with small probability. The most recent work, since 2006, has

introduced a new method known as Compressive Sensing (CS), which can improve the missing value in presence the available sample data [7, 8]. CS works with the following characteristics, ie sparse signal representation and Restricted Isometric Property (RIP) [9]. Sparse means signal with a few non-zero elements. Signals that are not sparse can be converted into sparse signals using the proper base transformation. Term that must be fulfilled in performing sparse signal recovery is RIP [10]. This property is almost orthonormal that acts as measurement matrix and applied to sparse vector.

In this paper, we proposed local interpolated compressive sampling to overcome the missing internet traffic. We first applied local interpolation which is focused on similarity property between rows and columns. We used two technique, such as correlation and Euclidean norm. We then combine its with CS reconstruction algorithms, such as Sparsity Regularized Singular Value Decomposition (SRSVD) [1], l_1 -norm optimization [11], Iteratively Reweighted Least Square (IRLS) [12], Orthogonal Matching Pursuit (OMP) [13].

II. THE PROPOSED METHODOLOGY

This section explains the proposed method of local interpolated compressive sampling for internet traffic reconstruction. The process of traffic matrix reconstruction in this research is shown in the Figure 1. This proposed method is an extension of previous research in [3]. Traffic matrix data that used in our simulation is from Abilene network [14]. The matrix represents that the row as a link between nodes and column as time measurements. The missing value is executed on TM matrix with probability p . After the missing process, were estimated based approach similarity between rows and columns using correlation and Euclidean norm. The CS process begins with a low-rank representation using SVD [15]. The next step is CS procedure of the low-rank matrix and measurement matrix A . The compression result is returned as the original TM using CS reconstruction algorithms, consisting of SRSVD, SVDL1, IRLS, and OMP. The scaling process is purpose for obtaining an amplitude value proportional to the original value.

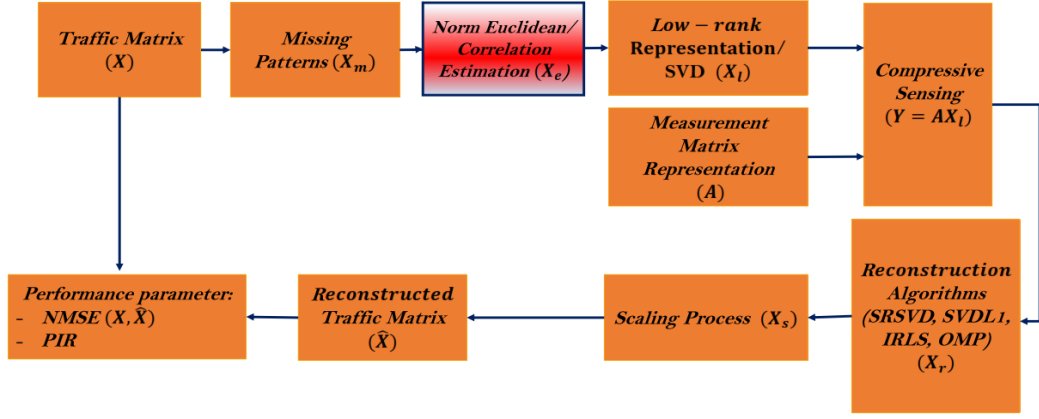


Fig. 1. The proposed method of local interpolated compressive sampling for internet traffic reconstruction

III. LOCAL INTERPOLATION

The interpolation solution is solved through a similarity approach between rows and columns to improve accuracy. If two rows/ columns show the similarity, then it can be assumed that one could be a great interpolant to another. We consider that the similarity of elements is influenced by the effect of the similarity between rows represented as α and the influence of similarities between columns expressed in β . We compare two methods to obtain the similarity, ie: correlation and Euclidean norm.

A. Correlation

Correlation uses the principle that the greater value of the correlation coefficient between rows / columns then the two rows / columns are similar. A traffic matrix of \mathbf{X} sized $(i \times j)$ and missing elements at position $\mathbf{X}(m, n)$, then the correlation procedure on traffic matrix \mathbf{X} are as follows:

Step 1) For the missing value in $\mathbf{X}(m, n)$, calculate the correlation coefficient between the row in the missing element \mathbf{X}_m and the other row \mathbf{X}_i with $i = 1, 2, 3, \dots, i$ according to equation (1) below:

$$\rho(\mathbf{X}_m, \mathbf{X}_i) = \frac{\text{cov}(\mathbf{X}_m, \mathbf{X}_i)}{\sqrt{\text{cov}(\mathbf{X}_m, \mathbf{X}_m)\text{cov}(\mathbf{X}_i, \mathbf{X}_i)}} \quad (1)$$

where $m \neq i$. Save the result in C_r

$$\text{cov}(\mathbf{X}_m, \mathbf{X}_i) = \frac{1}{J-1} \sum_{j=1}^J ((\mathbf{X}_{m_j} - \mu_{x_m})(\mathbf{X}_{i_j} - \mu_{x_i})) \quad (2)$$

$$\mu_x = \frac{1}{J} \sum_{j=1}^J \mathbf{X}_j \quad (3)$$

Step 2) Calculate the correlation between column of \mathbf{X}_n with the other column of \mathbf{X}_j with $j = 1, 2, 3, \dots, j$ and $n \neq j$ as in equation (1) and save the result in C_c .

Step 3) Find the maximum correlation coefficient in step 1 and 2

Step 4) Replace missing value $\mathbf{X}(m, n)$ using equation (4)

$$\mathbf{X}(m, n) = \alpha \arg \min_{v_i, m \neq i} \rho(\mathbf{X}_m, \mathbf{X}_i) + \beta \arg \min_{v_j, n \neq j} \rho(\mathbf{X}_n, \mathbf{X}_j) \quad (4)$$

where $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, \alpha + \beta = 1$

B. Euclidean Norm

The Euclidean norm approach states that the closer distance of the rows/ columns are, the two rows/ columns are similar. A traffic matrix size \mathbf{X} $(i \times j)$ and missing elements at $\mathbf{X}(m, n)$, then the Euclidean norm steps on traffic matrix \mathbf{X} are as follows:

Step 1) For the missing value in $\mathbf{X}(m, n)$, calculate the Euclidean norm between the row in the missing element \mathbf{X}_m and the other row \mathbf{X}_i with $i = 1, 2, 3, \dots, i$. The equation becomes:

$$d(\mathbf{X}_m, \mathbf{X}_i) = \sqrt{\sum_{j=1}^j (\mathbf{X}_{m_j} - \mathbf{X}_{i_j})^2} \quad (5)$$

where $m \neq i$. Save the result to D_r

Step 2) Calculate the Euclidean norm between column of \mathbf{X}_n with the other column of \mathbf{X}_j with $j = 1, 2, 3, \dots, j$ and $n \neq j$ as in equation (6) and save the result in D_c .

$$d(\mathbf{X}_n, \mathbf{X}_j) = \sqrt{\sum_{i=1}^i (\mathbf{X}_{n_i} - \mathbf{X}_{j_i})^2} \quad (6)$$

Step 3) Find the minimum norm from step 1 and 2

Step 4) Calculate the missing value $\mathbf{X}(m, n)$ using the following equation (7)

$$\mathbf{X}(m, n) = \alpha \arg \min_{v_i, m \neq i} d(\mathbf{X}_m, \mathbf{X}_i) + \beta \arg \min_{v_j, n \neq j} d(\mathbf{X}_n, \mathbf{X}_j) \quad (7)$$

where $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, \alpha + \beta = 1$

IV. EXPERIMENTAL AND RESULTS

A. Data Set

We use Abilene traffic data to test our proposed methods. This data set is used previously in our research [3] [4]. The Abilene network is composed of 12 nodes so there are 12×12 traffic flows connecting between nodes [14]. We assume that flow from the same source and destination is zero. Traffic flow is measured every 5 minutes. In one day, there are 288 measurements. In this study, TM rows represent traffic flow between nodes and TM columns represent time measurements.

B. Missing Patterns

We use six missing patterns such as Missing Row Elements (MRE), Missing Column Elements (MCE), Missing Rows at Random (MRR), Missing Columns at Random (MCR), Missing Elements at Random (MER), and Combine Missing Patterns (CMP) [3]. The missing is done by making a zero value on TM. This process is create randomly with the probability of missing (ρ).

MRE is a missing pattern by selecting one row and eliminating some elements in the selected row. Whereas MCE is a missing that chooses single column in TM and omits some elements in the column chosen. The missing model that deletes rows randomly is MRR, whereas the column is MCR. MER is missing by removing random elements. CMP is a combination of all previous missing models.

C. Performance Parameter

The performance parameter used to calculate the accuracy of TM reconstructed is Normalized Mean Square Error (NMSE). The NMSE is Mean Square Error (MSE) between the original TM $\mathbf{X}(i, j)$ and the reconstructed TM $\hat{\mathbf{X}}(i, j)$ normalized by MSE of original TM, which is mathematically expressed in the following equations [16]:

$$\begin{aligned} NMSE(\mathbf{X}(i, j), \hat{\mathbf{X}}(i, j)) &= \frac{MSE(\mathbf{X}(i, j), \hat{\mathbf{X}}(i, j))}{MSE(\mathbf{X}(i, j), 0)} \\ &= \frac{\|\mathbf{X}(i, j) - \hat{\mathbf{X}}(i, j)\|_2^2}{\|\mathbf{X}(i, j)\|_2^2} \end{aligned} \quad (8)$$

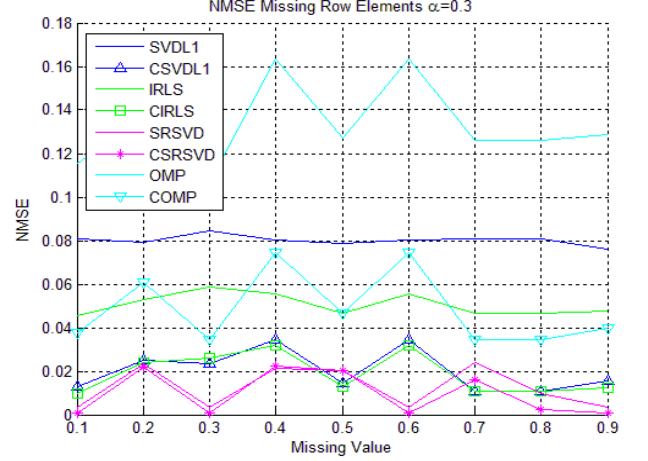
The other metric is Performance Improvement Ratio (PIR). The PIR denotes an increasing in a new approach to the old method. In this study, we used NMSE to calculate PIR, which is defined as follows [17]:

$$PIR = \frac{NMSE_l - NMSE_n}{NMSE_l} \quad (9)$$

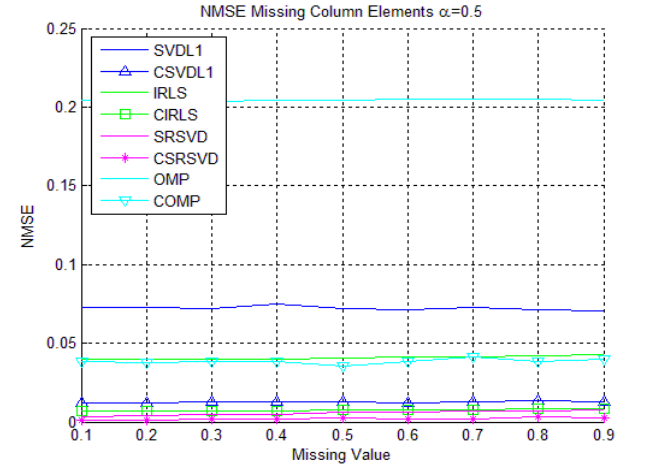
where $NMSE_l$ denotes performance parameter from old algorithm, while the $NMSE_n$ states the performance of the proposed algorithm.

D. Combination of Correlation

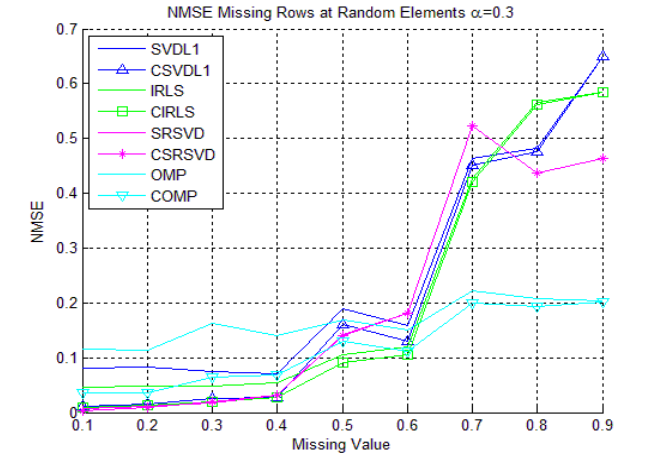
In this study, a combination of correlation with CS reconstruction algorithms (SRSVD, SVDL1, IRLS, and OMP) was proposed to improve the reconstruction performance. This combination produces new methods called Combine SRSVD (CSRSVD), Combine SVDL1 (CSVDL1), Combine IRLS (CIRLS), and Combine OMP (COMP).



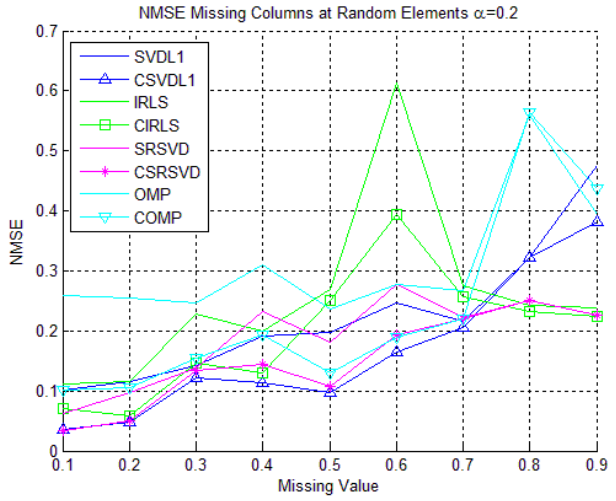
(a) NMSE in MRE



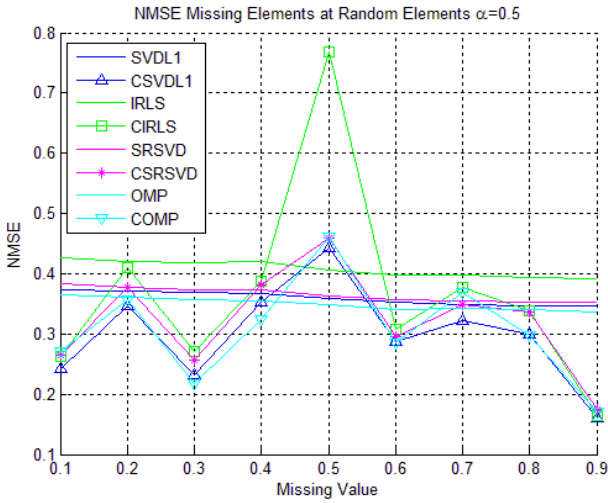
(b) NMSE in MCE



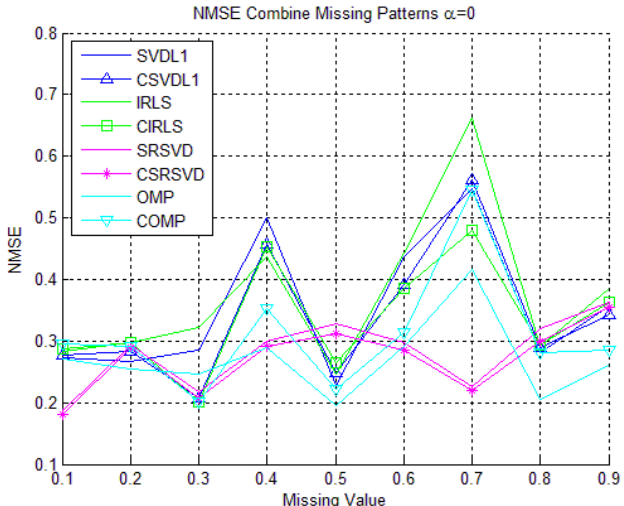
(c) NMSE in MRR



(d) NMSE in MCR



(e) NMSE in MER



(f) NMSE in CMP

Fig. 2. NMSE and missing value relationships in different missing patterns (a) MRE, (b) MCE, (c) MRR, (d) MCR, (e) MER, (f). CMP

Figure 2 illustrates the NMSE in the missing patterns. The X-axis represents missing value and the Y-axis denotes the value of NMSE. The figure also shows in different α parameter. NMSE testing performed on different missing value, this describes the effect of missing value on the reconstruction result

Figure 2. (a) shows NMSE on MRE pattern with α parameter 0.3. All proposed algorithms can improve accuracy with NMSE results <0.08 . The increasing of missing probabilities has no effect on NMSE. Because missing elements only occur on one row so that the number of available samples is still quite a lot. In missing probability 0.9 on one row is equal with 0.6% missing from the total number of matrix elements.

Figure 2. (b) describes NMSE on MCE pattern with α parameter 0.5. The simulation results show the NMSE value less than 0.05 on all the proposed algorithms. The amount of missing probability has no effect on NMSE, this is because the missing probability of 0.9 on one column is identical to 0.3% missing of whole elements matrix.

Figure 2. (c) shows NMSE on MRR patterns with α parameter 0.3. The simulation results show that the greater the probability of missing, the greater the NMSE. In this case, the proposed algorithm can not significantly decrease the NMSE value especially in CSRSVD. In the CSVDL1 and CIRLS algorithms, the NMSE value decrease still occurs until the probability value is lost by 0.8. While COMP showed the best performance

Figure 2. (d) describes NMSE on MCR pattern with α parameter 0.5. In this model, the NMSE value is proportional to the probability of missing. In CSRSVD there is a decrease in the value of NMSE, but on the probability of traffic lost 0.7, the NMSE decline is very slow. In the CSVDL1 and CIRLS algorithms, there is a decrease in the NMSE value on all lost traffic probability values. COMP produces poor performance due to an increase in NMSE value when the probability of lost traffic starts from 0.8.

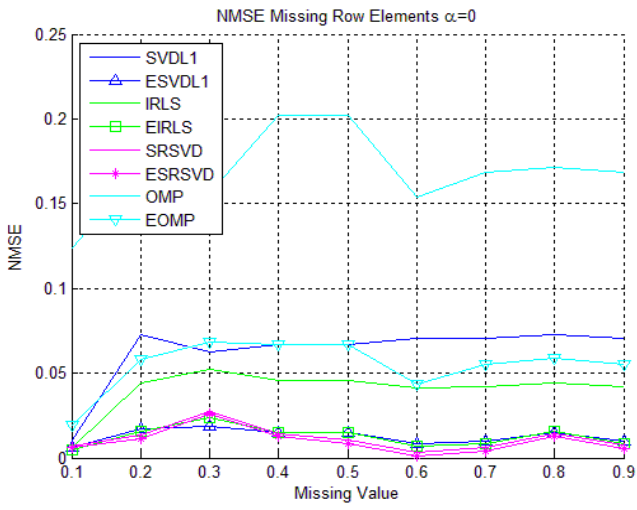
Figure 2. (e) describes NMSE on MER pattern with α parameter 0.5. Some conditions suggest that the proposed algorithm decreases the performance of the accuracy results, as shown in the probability of missing 0.5. This is due to the random nature of the missing traffic elements so that existing traffic does not have a high correlation with each other to predict the values of missing elements.

Figure 2. (f) shows NMSE on CMP pattern with α parameter 0. Combine CS does not significantly decrease the value of NMSE especially in CSRSVD. Significant decrease occurs in CIRLS, whereas in the CSVDL1 algorithm there is a decrease in NMSE value only on some probability value of lost traffic. COMP is not suitable for CMP missing model.

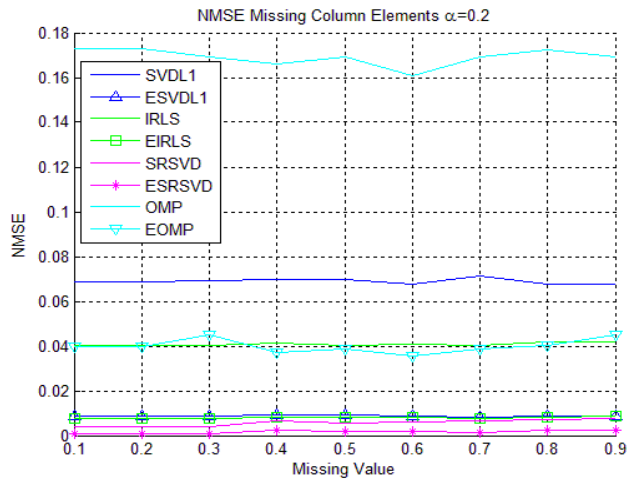
E. Enhance Euclidean Norm

Enhance Euclidean norms aims to improve the performance of CS reconstruction algorithms. The proposed algorithm is named Euclidean SRSVD (ESRSVD), Euclidean SVDL1 (ESVDL1), Euclidean IRLS (EIRLS), and Euclidean OMP (EOMP).

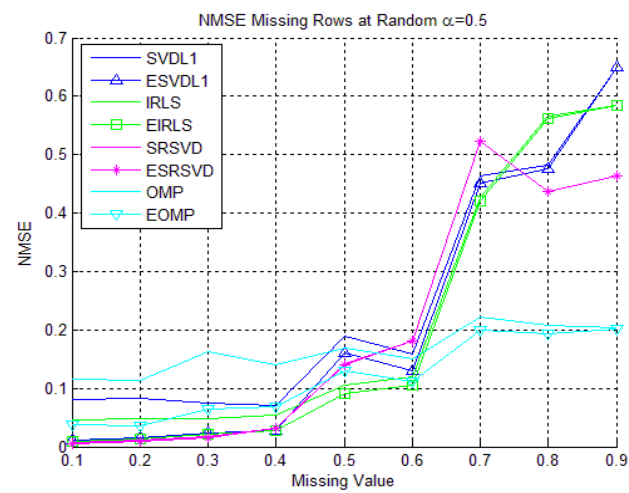
Figure 3 shows the relationship between NMSE and the probability of missing on different missing types and α parameter. The X-axis is missing value and the Y-axis is the NMSE.



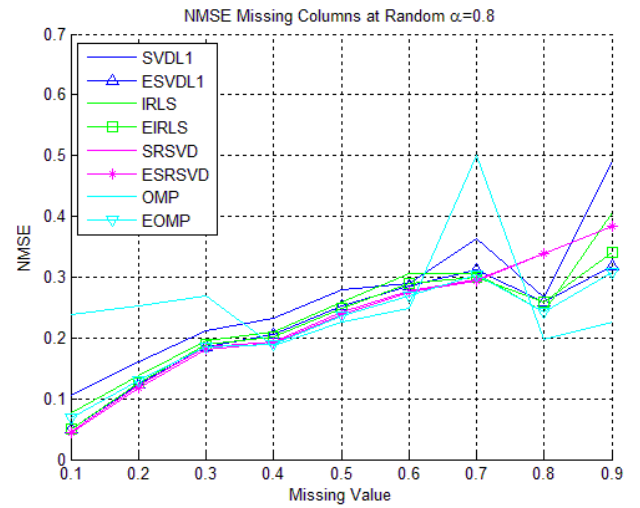
(a) NMSE in MRE



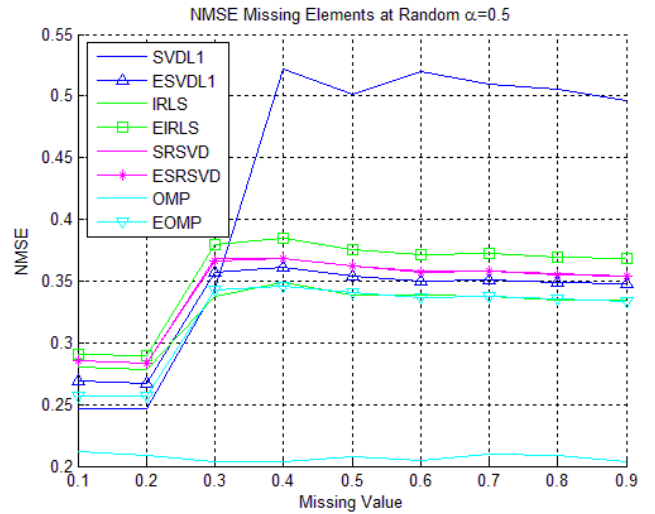
(b) NMSE in MCE



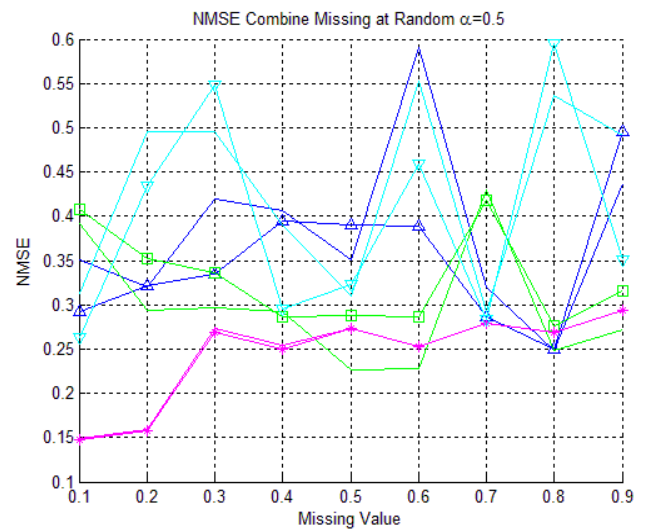
(c) NMSE in MRR



(d) NMSE in MCR



(e) NMSE in MER



(f) NMSE in CMP

Fig. 3. NMSE and missing value relationships in different missing patterns (a) MRE, (b) MCE, (c) MRR, (d) MCR, (e) MER, (f) CMP

Figure 3. (a) expresses NMSE on MRE missing pattern with α parameter 0. All proposed algorithms can improve accuracy with NMSE results <0.07 . EOMP shows the largest NMSE decline, followed by ESVDL1 and EIRLS. While the ERSVD decrease in NMSE is not significant.

Figure 3. (b) illustrates NMSE on MCE missing type with α parameter 0.2. The reconstruction algorithm applied to the missing MCE pattern always yields the best accuracy compared to the other missing techniques. Enhance Euclidean norm able to lower NMSE less than 0.05.

Figure 3. (c) describes NMSE on MRR missing model with α parameter 0.5. The NMSE value increases with increasing the probability of missing. The proposed algorithm can slightly lower NMSE.

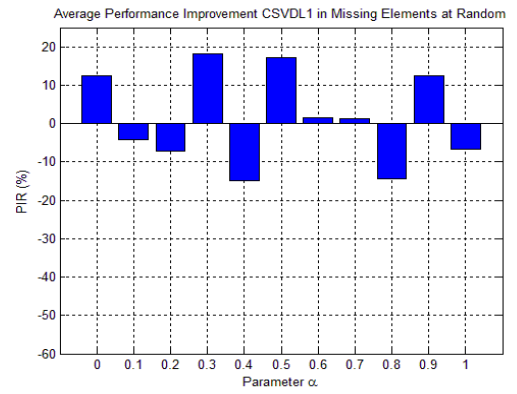
Figure 3. (d) shows NMSE on MCR missing pattern with α parameter 0.8. As in MRR, NMSE value increases with increased probability of missing. All proposed algorithms succeeded in decreasing the NMSE, except on EOMP. In the probability missing above 0.8, EOMP actually increases the NMSE.

Figure 3. (e) describes NMSE on MER pattern with α parameter 0.5. The simulation results shows that only ERSVD decreases NMSE even though the decrease is very low. The other proposed algorithms can not work well. This is greatly influenced by the random way of missing elements so that the present elements in the matrix may have bit of similarity. Therefore, the new algorithm is difficult to get the closest distance between matrix elements.

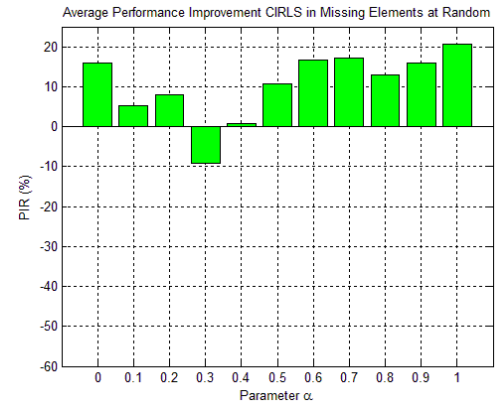
Figure 3. (f) illustrates NMSE on CMP pattern with α parameter 0.5. ERSVD is able to decrease NMSE even though its value is very small. While other proposed algorithms do not work well where some tests are able to decrease NMSE, and some actually increase NMSE. Because the CMP is a combination of some missing randomly chosen, such as missing rows, missing columns, and missing elements, then the probability of a certain missing can result in intersection between missing processes so that the amount of missing becomes less or even vice versa.

F. Similarity Parameter

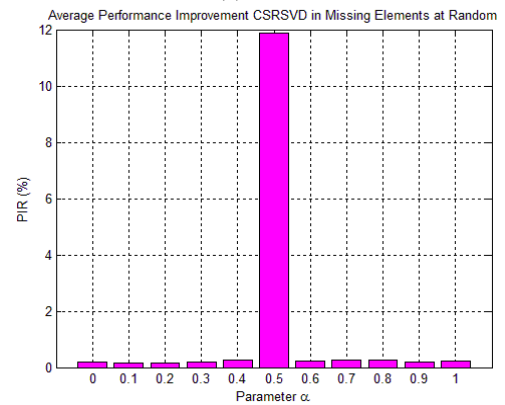
The similarity parameter that used are two, namely the similarity between rows (α) and similarities between columns (β). The similarity between rows implies the relationships that occur between links, while the similarity between columns refers to the relationship between time. The experiments were performed 10 times and the average result shown in figure 4. The results are presented only in the case of missing MER.



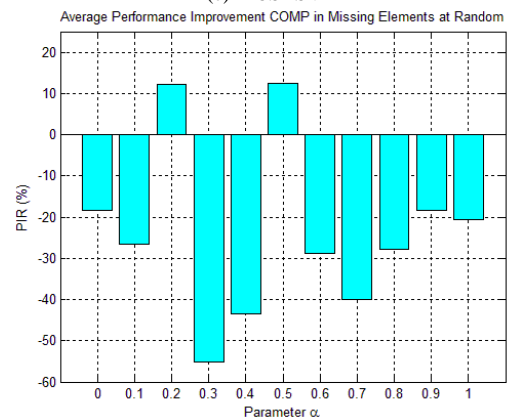
(a) CSVDL1



(b) CIRLS



(c) CSRSVD



(d) COMP

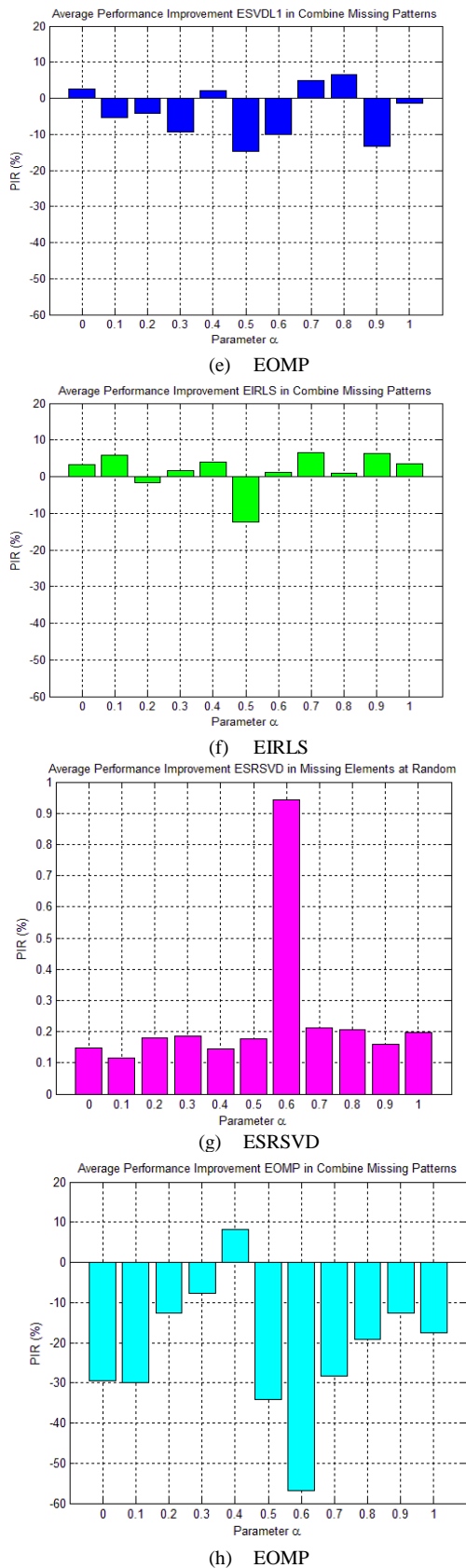


Fig. 4. The influence of parameter α on the MER missing pattern in the proposed algorithms (a) CSVDL1, (b) CIRLS, (c) CSRSVD, (d) COMP, (e) ESVDL1, (f) EIRLS, (g) ERSVD, (h) EOMP

Performance improvements at all parameter α occur in CSRSVD and ESRSVD, although the increase is very small, ie less than 12% in CSRSVD and less than 1% in ESRSVD. While on the other algorithms, performance improvement is strongly influenced by correlation factor between row and column. In CIRLS, 90% of the experiments show performance improvement, and the highest performance occurs at $\alpha = 1$ and $\beta = 0$, it indicates that accuracy is only affected by correlation between rows. In CSVDL1, 60% of the experiments showed the highest performance improvement with $\alpha = 0.3$ and $\beta = 0.7$, this illustrates that the correlation between columns is more important than the correlation between rows at the time of reconstruction. At COMP, 20% of the experiments increased, and the maximum performance occurred at $\alpha = \beta = 0.5$.

In EIRLS, 90% of the experiments showed improved performance, and the highest performance occurred at $\alpha = 0.7$ and $\beta = 0.3$. In ESVDL1, 60% of the experiments showed the highest performance improvement occurred at $\alpha = 0.8$ and $\beta = 0.2$. While EOMP, 10% experiments increased, and maximum performance occurred at $\alpha = 0.4$ and $\beta = 0.6$.

V. CONCLUSIONS

Enhance norm Euclidean and Combination of Correlation on CS reconstruction algorithm (SVDL1, IRLS, SRSVD, OMP) can improve accuracy in case of lost traffic MRE and MCE. ESRSVD and CSRSVD do not provide significant performance improvements since SRSVD has been able to work well in TM reconstruction. EIRLS and CIRLS can provide significant performance improvements even though NMSE values still need to be fixed. EOMP and COMP are not suitable for performance improvements, especially in lost MCR, MER, and CMP traffic patterns.

ACKNOWLEDGMENT

The authors would like to thank Telkom Foundation, Indonesia Ministry of Higher Education, and LPPM ITB for financial support in this research.

REFERENCES

- [1] M. Roughan, Y. Zhang, W. Willinger and L. Qiu, "Spatio-Temporal Compressive Sensing and Internet Traffic Matrices (Extended Version)," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 20, no. 3, pp. 662-676, 2012.
- [2] Zhou Huibin; Zhang Dafang; Xie Kun; Wang Xiaoyan, "Data Reconstruction in Internet Traffic Matrix," *IEEE Journal and Magazine, China Communication*, vol. 11, pp. 1-12, 2012.
- [3] I. D. Irawati, A. B. Suksmono and I. J. M. Edward, "Missing Internet Traffic Reconstruction using," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 9, no. 1, pp. 57-66, 2017.
- [4] I. D. Irawati, A. B. Suksmono and I. J. M. Edward, "Low-Rank Internet Traffic Matrix Estimation based on Compressive Sampling," *Advanced Science Letters*, vol. 23, no. 5, 2017.

- [5] A. B. Suksmono, "Interpolation of PSF based on compressive sampling and its application in weak lensing survey," *Monthly Notices of the Royal Astronomical Society (MNRAS)*, vol. 443, pp. 919-926, 2014.
- [6] D. Guo, X. Qu and L. Huang, "Sparsity-Based Spatial Interpolation in Wireless Sensor Networks," *Sensors*, vol. 11, pp. 2386-2407, 2011.
- [7] D. Donoho, "Compressed Sensing," *IEEE Transaction Information Theory*, vol. 52, pp. 1289-1306, 2006.
- [8] E. Candes, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.," *IEEE Transaction Information Theory*, vol. 52, pp. 489-506, 2006.
- [9] E. Candes, "The Restricted Isometry Property and Its Implications for Compressed Sensing," *Compte Rendus de l'Academie des Sciences*, vol. 1, pp. 589-592, 2008.
- [10] E. J. Candes and T. Tao, "Decoding by Linear Programming," *IEEE Transaction Inf.*, vol. 51, no. 12, pp. 4203-4215, 2005.
- [11] E. Candes, J. Romberg and Caltech, "L1magic: Recovery of Sparse Signals via Convex Programming," 2005.
- [12] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sampling," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 3869-3872, 2008.
- [13] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurement Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655-4666, 2007.
- [14] "The Abilene Research Network," <http://abilene.internet2.edu/s..>
- [15] A. Lakhina; K. Papagiannaki; M. Crovella; C. Diot; E. Kolaczyk; N. Taft, "Structural Analysis of Network Traffic Flows," in *The Joint International Conference on Measurement and Modeling of Computer Systems*, New York, 2004.
- [16] D. Brunet, E. R. Vrscay and Z. Wang, "On the Mathematical Properties of the Structural Similarity Index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488-1499, 2012.
- [17] L. Nie, D. Jiang and L. Guo, "End-to-End Network Traffic Reconstruction Via Network Tomography Based on Compressive Sensing," *Journal of Network and Systems Management*, vol. 23, no. 3, p. 709-730, 2015.



Indrarini Dyah Irawati received B.S. and M.S. degree in electrical engineering from Sekolah Tinggi Teknologi Telkom (STTT), Bandung, Indonesia. She is a doctoral candidate in School of Electrical Engineering and Informatics, Institut Teknologi Bandung (ITB), Bandung, Indonesia. She joined as a Lecturer at Telkom Applied Science School, Telkom University, since 2007.

She is currently a member of the Association for Computing Machinery (ACM). Her research interests are in the areas of computer network, signal processing, and compressive sensing.



Andriyan Bayu Suksmono (M'02-SM'08) received the B.S. degree in physics and the M.S. degree in electrical engineering from Institut Teknologi Bandung (ITB), Indonesia, and the Ph.D. degree in engineering from the University of Tokyo, Japan, in 1990, 1996 and 2002, respectively. He joined ITB as an Instructor (1996-2005), Associate Professor (2005-2009), and Professor (2009-present) at the School of Electrical Engineering and Informatics, ITB.

His main research interests are compressive sensing, signal processing and imaging, and radar.

He is a Senior member of IEEE, and IEICE. He received the 2013 Multimedia Information Technology and Applications Best Paper Award from Korea Multimedia Society (KMMS), the 2014 Best Presenter Award from Republic of Indonesia Ministry of Research, Technology and Higher Education.



Ian Joseph Matheus Edward received the B.S. degree and the M.S. degree in electrical engineering from Institut Teknologi Bandung (ITB), Indonesia, and Ph.D. degree in telecommunication management from Universitas Indonesia (UI), in 1992, 1996, and 2007, respectively. He is an Associate Professor at the School of Electrical Engineering and Informatics, ITB.

He is currently an expert team of defense device safety at Republic of Indonesia Ministry of Communication and Informatics. His main research interests are Telecommunication Management, Capacity Planning, Fault Management, Optical Fiber Communication, and Wireless Communication.