

Selecting the appropriate size of the graph for self-diagnostic model with graph density

Sutat Gammanee¹, Sunantha Sodsee²

¹Department of Information Technology, Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

²Department of Data Communication and Networking, Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Article Info

Article history:

Received Jan 13, 2022

Revised Mar 21, 2022

Accepted Apr 1, 2022

Keywords:

Bipartite graph
Graph density
Machine learning

ABSTRACT

Self-diagnosis is the concept of self-diagnosing disease from symptoms. We have the idea to create self-diagnostic models from diagnostic data. The data to be analyzed were from a medium-sized hospital in Thailand. The model is divided by structured data and unstructured data. The first step is to process structured data with cluster algorithms. The second step is to evaluate the unstructured data to group symptoms into a bipartite graph. After the graph was created, the model was divided into 10 levels, according to the level of similarity. This research aims to apply the concept of density graph, the Kappas and multiple line graph to selecting the appropriate diagnosis model. The results of all three experiments showed that the appropriate model was at a level of similarity at 40%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sutat Gammanee

Department of Information Technology, Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand

Email: sutat@kru.ac.th

1. INTRODUCTION

The current situation regarding the Coronavirus 2019 has made us realize that public health problems are an important problem of the country. Especially the problem of doctors, those are not sufficient to support the demand of the number of patients. Thailand has a ratio of 1 registered nurse per population of 412 people [1]. Therefore, the government has a policy that emphasizes the participation of the people in taking care of their own health [2], which diagnose themselves about the risks of different diseases to reduce the problem of seeing a doctor.

Self-diagnosis is a concept of observing their own abnormalities, that reduce the risk of disease. Due to the current situation, social conditions, people work hard and they do not have time to go for a medical examination. People will go to the hospital only when they really get a serious health problem. The current situation makes the doctor's work is overload that there is no time to serve normal patients. Self-diagnosis is an option to reduce health problems and also choose to use public health services instead. Currently, the concept is applied machine learning to assist in self-diagnosis [3]-[6].

From the Figure 1, the researcher has the concept of self-diagnosis design. The concept has divided data into structured data and semi-structured data. Structured data include: preliminary values measured from body characteristics such as blood pressure values and heartbeat values, which is employed in screen patients

between normal group and surveillance group [7]. The unstructured data is a diagnostic based on the patient's medical history. It has been separate into a group of symptoms. We use this data to generate bipartite graph [8]-[11] consisting of two classes as in Figure 1.

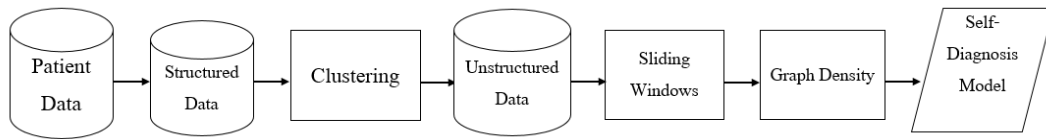


Figure 1. Self-diagnosis concept

From Figure 2, it consist of two classes, the patient class represented by P and the symptom class represented by S. The experiment showed that there are too many sets of class. The graph was complicated, and made inefficient processing. Therefore, The symptoms were grouped using the sliding window method [12], [13], and created a graph again. There are 10 graphs divided by similarity 0% to 100%. The challenge of this paper was selected appropriate model from 10 graphs.

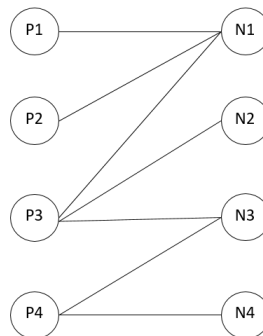


Figure 2. Bipartite graph create from unstructured data

The concept of graph density has been applied to this research include: the Kappas method, and the concept of multiple line graph to find the most appropriate graph model for further disease prediction. This paper proposes two graph density formulas that applied to bipartite graphs. Which is applied from the concept of finding the density of the graph and the Kappas method. In addition, Multiple line graph theory is added to assist confirm the results.

2. RESEARCH METHOD

In this section, we discuss the theory that applied to find the most appropriate graph model for further disease prediction, including graph density, the Kappas and multiple line graph. Sections 2.1 and 2.2 discuss the concept of graph density, but in different methods. Section 2.3 uses the concept of Multiple line graph to encourage the results. This research is divided into 3 stages of experiment design. The first stage is the graph density test. The second stage is to calculate the graph density as well, but different algorithms. The third stage applied the multiple line graph theory.

2.1. Graph density

Graph density is a mathematical concept to measure the density of a graph. It is a calculate of the ratio between the number of edges to the total number of possible edges. The concept of graph density is used to determine the communication and connection between nodes. Which can compare more and less densities. A large value represents the number of connections between the large number of nodes which means that nodes have multiple paths to connect to each other [14]-[17]. Smaller values mean that the less number of lines, making communication with less choice of routes. The (1) for calculating the density of graph in non-directional graph as:

$$D = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V|-1)} \quad (1)$$

- D refer to the density of the graph.
- |E| refer to the number of actual edges.
- |V| refer to the number of nodes.

The density value is a minimum of 0, which means that there are no edges in the graph at all. The highest density value is 1, i.e., the number of lines is equal to the total number of possible edges. Presently, there are many aspects of the graph density problem that are still being developed and divided into several theories. Both in terms of selecting a cluster and determining the importance of node relationships [18]-[20].

We applied the above theory to find the appropriate graph density. It calculates from the actual edge to all possible edges as shown in (2).

$$\text{Density} = \frac{\text{Number of Edges}}{\text{Number of possible edges}} \quad (2)$$

The number of Edges means the total number of edges in the graph and the number of possible edges means the maximum number of edges possible. The value of maximum of possible edges is obtained by multiplying the number of nodes of two sets in the bipartite graph. The maximum value of density is 1, meaning the number of edges equal to the number of possible edges, and the lowest value is 0, meaning no edge at all, and once the density is obtained, the best approximate value is 0.5, which is the middle value between 0 and 1. The concept of graph density determination allows to select an appropriate graph which selects from values close to 0.5.

2.2. The Kappas

The Kappas [21], [22] is a development in the concept of graph density to determine the strength of node segmentation to find a group of node divisions, the concept uses global connection determination and then compares it with local connection. The method is to find the ratio value called Intra, which is the mean of Local connections, and K inter for global averaging. The Kappas method is based on the general graph density method as the (3).

$$K = \frac{|E|}{0.5 \times N(N-1)} \quad (3)$$

- K Refer to the density of graph
- |E| refer to number of edges in the graph
- N refer to number of nodes

It has adjusted the formula for calculating local as a (4).

$$\bar{K}_{\text{intra}} = \frac{1}{l} \sum_{i=1}^l (k_i) = \frac{1}{l} \sum_{i=1}^l \frac{|E_{ii}|}{0.5 \times n_i(n_i-1)} \quad (4)$$

- K Refer to the density of graph of intra
- |E| refer to number of edges in the graph intra
- n refer to number of connected nodes

The formula for calculating global connection as (5).

$$\begin{aligned} \bar{K}_{\text{inter}} &= \frac{1}{0.5 \times l(l-1)} \sum_{i=1}^l \sum_{j=i+1}^l k_{ij} \\ &= \frac{1}{0.5 \times l(l-1)} \sum_{i=1}^l \sum_{j=i+1}^l \frac{|E_{ij}|}{0.5 \times ((n_i+n_j)(n_i+n_j-1) - n_i(n_i-1) - n_j(n_j-1))} \\ &= \frac{1}{0.5 \times l(l-1)} \sum_{i=1}^l \sum_{j=i+1}^l \frac{|E_{ij}|}{n_i \times n_j} \end{aligned} \quad (5)$$

- K Refer to the density of graph of inter
- |E| refer to number of edges in the graph
- n refer to number of connected nodes

The concept of calculating Kappas value by comparing the values between graph global connection and node local connection. In this research, the formula has been adjusted to our research as (6).

$$\text{Kappa Density} = \frac{\text{Number of possible edges}}{(0.5 \times N1 \times N2)} \tag{6}$$

- By N1 refers to Number of Patients
- N2 refers to Number of Symptom

The maximum value is 2, meaning the number of edges equal to the number of possible edges. The best value is 1, which means the middle value between 0 and 2. The point that differs from the density determination in Section 2.1 is that this formula is more elaborate because the density is calculated from the external and internal of the node. This formula is used in emphasizing with Section 2.1. If the experimental results are in the same direction, the results are considered reliable.

2.3. Multiple line graph

A line graph is a graph that contains two or more data points that connect by line. These explain the relationship between two axes of the sgraph. A line graph the capability to show data variable and trends, which lead to prediction results of data.

A multiple line graph is a two or more lines. It used for comparison between the lines, viewing trends, and relation between the lines [23]-[27]. A multiple line graph is a type of graph that shows two or more variables changing at the same time. Multiple line graph example as Figure 3.



Figure 3. Multiple graphs showing monthly sales

From Figure 3, it shows a multiple line graph of the sales of 3 products from January to May. It is able to analyze trends that the three products have trends in the same direction with each other, and the second product has sales that fall between the 1st and 3rd products. This research is a continuation of study [7] on applying sliding windows [12] to group symptoms. The results of the study were to create 10 graphs grouped by similarity, in order from the level of similarity at 10%-100%. The graph is characteristic of the Bipartite graph as Figure 2, on the left side are a patient sample class from P1-P36 and the other are the disease group from N1-N22.

From the Figure 4, The graph was created of the 100% similarity level of the symptoms group, the number of classes N has not been grouped. From the Figure 5, the graph was created of the 50% similarity level, the number of N classes is reduced from 22 to 12, because of the grouping of N classes. In Figure 5, the number of class N is one at 20% similarity.

From the Figures 4 to 6, the class n tends to decrease from 100% to 20% similarity. This research aims to select one graph that the proper model to apply self diagnosis. The appropriate graph that makes the model fit, it's not overly complicated and the number of classes is not too small to be unusable. Therefore, the concept of graph density was applied to select the appropriate graph.

This research applied the hypothesis multiple line graph, where the X-axis represents patient classes from P1-P36. The Y-axis is the number of edges exiting each node. The assessment method is based on the graph trend, which is a multiple line graph. The appropriate values from the graph where the optimal value is the line in the middle of every line. From Figure 7, it is shown that P2 is in the middle between the lines P1 and P3. The observation point is that P2 value is no more and no less than P1 and P3. So that means that the value of P2 is the appropriate value.

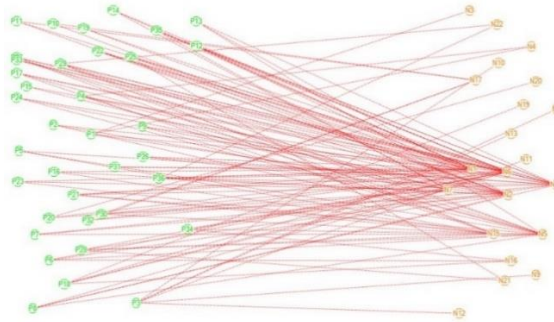


Figure 4. Graph at 100% similarity

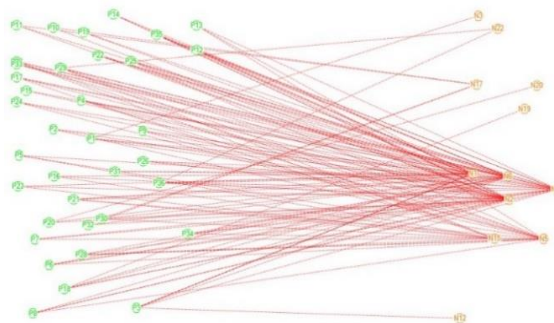


Figure 5. Graph at 50% similarity

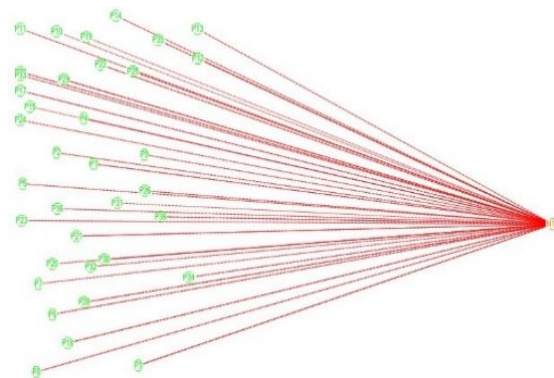


Figure 6. Graph at 20% similarity

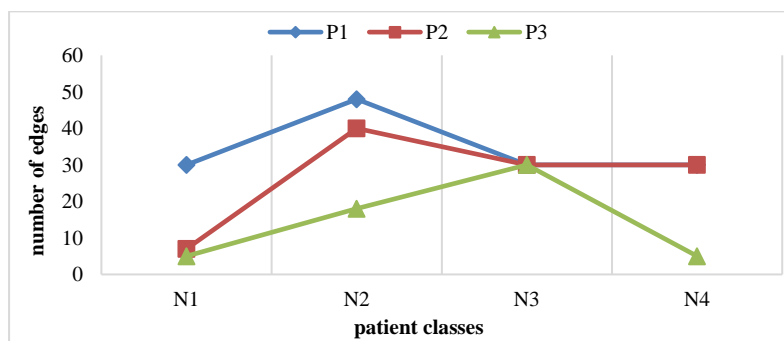


Figure 7. Multiple line graph showing optimal value example

3. RESULTS AND DISCUSSION

3.1. Experimental results from graph density and Kappa method

This section presents the experimental results of the three methods to find the appropriate model out of 10 different models. The Table 1 shows the graph density results and the Kappas calculations. The density result is the value calculated in (2), while the Kappas are the values calculated as shown in (6).

Table 1. Density graph and the Kappas experiment results

Percent	No class (N)	No class (P)	Edge	Density	distance to 0.5	The Kappas
100%	21	36	130	0.171958	0.328042	0.343915
90%	21	36	130	0.171958	0.328042	0.343915
80%	20	36	130	0.180556	0.319444	0.361111
70%	18	36	130	0.200617	0.299383	0.401235
60%	16	36	130	0.225694	0.274306	0.451389
50%	13	36	127	0.271368	0.228632	0.542735
40%	8	36	122	0.423611	0.076389	0.847222
30%	2	36	60	0.833333	0.333333	1.666667
20%	1	36	36	1	0.5	2
10%	1	36	36	1	0.5	2

From the Table 1 shown, the level of similarity number of classes N; the number of classes P, number of Edge, and density values. The 5th column shows the density value, and the 6th column shows the distance from 0.5, which is considers the middle value between 0 and 1. The result is shown that the density value of 40% similarity is the closest to the 0.5 with only 0.076389. The 7th column in Table 1 shows the Kappas value, and closest to 1 was 0.847222 at a level of 40% similarity.

3.2. Select model with multiple line graph

From the concept in section 3.3, the X-axis represents the patient class from P1-P36. The Y-axis is the number of edges each node. The graph shows the level of similarity model from level 100 to level 10, where some levels may be equal to others. For example, the degree of similarity at 100 is equal to 90 and thus is shown on the same line. The results are shown in the Figure 8.

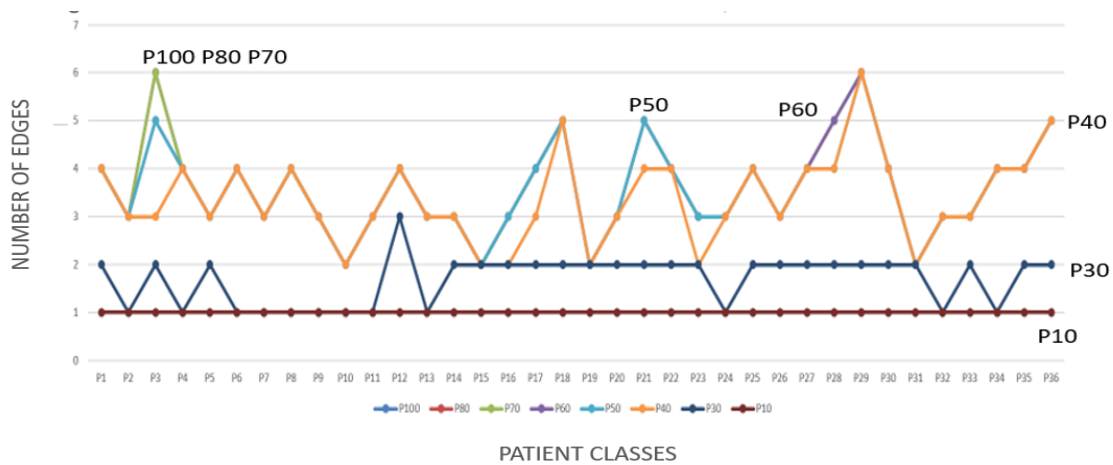


Figure 8. Multiple line of patient class and number of edge

Based on Figure 8, the blue line represents the P100 level as 100% similarity, the red at the P80 as 80% similarity level, the green at the P70 level as 70% similarity, the purple at the P60 level as 60% similarity, the blue at the P50 level 50% similarity, the orange at the 40 level as 40% similarity, the dark blue at the P30 level as 30% similarity, and the red at the P20 level as 20% similarity. The appropriate values are found on the orange line, which is the level. at 40% similarity, where the value lies between the remaining lines. Therefore, these results support the density and Kappa values in selecting the appropriate model at 40% similarity.

4. CONCLUSION

The self-diagnosis was designed by separating structure and unstructured data. The unstructured data is used to create medical terminology and lead to the creation of bipartite graphs, which have two classes: patient classes and symptoms classes. However, when creating the graph it is complicated to process. There are 10 graphs divided by similarity 0% to 100%.

In this paper, we proposed a concept of graph density; the Kappas method and support with the concept of graph trend, which selecting an appropriate graph for self-diagnosis. The result from three methods indicate, the appropriate at 40% similarity. First method, the value from graph density at 40% similarity is 0.423611, which has the shortest distance from 0.5, Kappas provide value 0.847222 at 40% similarity, which is close to the value 1, and when we observe from graph trend, we find 40% similarity is in the middle of all lines. Therefore, the graph at 40% similarity was chosen as a model for self-diagnosis. From the experiment results, Appropriate model results were obtained at the 40% level. The reason why the results are closer to 0% is that there is not much difference between the 80%-100% levels. Therefore, the result is in the appropriate in the 40% level.

The experiment was based on the assumption of selecting the most appropriate model from 10 models for further processing. This paper presents the use of two graph density methods to select the desired model and the use of multiple line graph theory to assist in the selection. This research requires a model that is neither too complicated nor too simple for efficient processing. The experimental methods presented in this paper are effective at dividing complexity. In other researches that require to select models with different characteristics It is better to use alternative methods of selection that can display the exacted characteristics. The future of this work, we bring the selected model to be examined in other disease groups for validity to further confirm the model.

ACKNOWLEDGEMENTS

The author would like to thank the Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, and related agencies for their support in the successful completion of this research.




REFERENCES

- [1] K. Sawaengdee, "Crisis of nursing shortage in health service facilities under office of permanent secretary, Ministry of Public Health: Policy recommendations," *Journal of Health Science*, vol. 26 no. 2, pp. 448-456, Mar-Apr. 2017.
- [2] R. Phalasuek, B. Thanomchayathawatch, and D. Songloed, "Participatory action research: Development of a participatory process for health promotion in the community," *The Southern College Network Journal of Nursing and Public Health*, vol. 5, no. 1, pp. 211-223, Jan-Apr. 2018.
- [3] M. Al-Zeyadi *et al.*, "Deep learning towards intelligent vehicle fault diagnosis," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1-7. doi: 10.1109/IJCNN48605.2020.9206972.
- [4] M. Elhadeif, "A machine learning approach for self-diagnosing multiprocessors systems under the generalized comparison model," in *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, Dec. 2014, pp. 417-424. doi: 10.1109/UIC-ATC-ScalCom.2014.5.
- [5] S. A. Kumar, C. V. Krishna, P. N. Reddy, B. R. K. Reddy, and I. J. Jacob, "Self-diagnosing health care chatbot using machine learning," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 9323-9330, May 2020.
- [6] A. Ćirković, "Evaluation of four artificial intelligence-assisted self-diagnosis apps on three diagnoses: two-year follow-up study," *Journal of Medical Internet Research*, vol. 22, no. 12, Dec. 2020, doi: 10.2196/18097.
- [7] S. Gammanee and S. Sodsee, "Applying clustering algorithm for primary screening of patients," in *The 14th National Conference on Computing and Information Technology*, 2018, pp. 462-467.
- [8] R. Diestel, *Graph theory*, 2nd ed., Springer 1997, doi: 10.1007/978-3-662-53622-3.
- [9] B. R. Arunkumar and R. Komala, "Applications of Bipartite Graph in diverse fields including cloud computing," *International Journal of Modern Engineering Research*, vol. 5, no. 7, pp. 1-7, Jan. 2015.
- [10] A. Singh, "Higher matching complexes of complete graphs and complete bipartite graphs," *Discrete Mathematics*, vol. 345, no. 4, Apr. 2022, doi: 10.1016/j.disc.2021.112761.
- [11] S. Jendroř, J. Miskuf, and R. Soták, "Total edge irregularity strength of complete graphs and complete bipartite graphs," *Discrete Mathematics*, vol. 310, no. 3, pp. 400-407, 2010, doi: 10.1016/j.disc.2009.03.006.
- [12] S. Gammanee and S. Sodsee, "Application of sliding windows to spelling error detection in medical diagnosis," in *Big Data Analytics, Data Mining and Computational Intelligence 2020*, 2020, pp. 149-156.
- [13] S. Köhler, S. Gulati, G. Cao, Q. Hart, and B. Ludascher, "Sliding window calculations on streaming data using the Kepler scientific workflow system," *Procedia Computer Science*, vol. 9, pp. 1639-1646, Dec. 2012, doi: 10.1016/j.procs.2012.04.181.
- [14] J. Darlay, N. Brauner, and J. Moncel, "Dense and sparse graph partition," *Discrete Applied Mathematics*, vol. 160, no. 16-17, pp. 2389-2396, Nov. 2012, doi: 10.1016/j.dam.2012.06.004.
- [15] P. Miasnikof, A. Y. Shestopaloff, A. J. Bonner, Y. Lawryshyn, and P. M. Pardalos, "A statistical density-based analysis of graph clustering algorithm performance," *Social and Information Networks*, Mar. 2020, doi: 10.48550/arXiv.1906.02366.
- [16] T. von Landesberger *et al.*, "Visual analysis of large graphs: state-of-the-art and future research challenges," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1719-1749, 2011, doi: 10.1111/j.1467-8659.2011.01898.x.




- [17] M. Charikar, "Greedy approximation algorithms for finding dense components in a graph," in: *Approximation Algorithms for Combinatorial Optimization*, 2000, pp. 84–95, doi: 10.1007/3-540-44436-X_10.
- [18] A. V. Goldberg, "Finding a maximum density subgraph," Technical report, California, USA, 1984.
- [19] B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory," *PLOS ONE*, vol. 5, no. 10, Oct. 2010, doi: 10.1371/journal.pone.0013701.
- [20] B. S. Anderson, C. Butts, and K. Carley, "The interaction of size and density with graph-level indices," *Social Networks*, vol. 21, no. 3, pp. 239-267, Jul. 1999, doi: 10.1016/S0378-8733(99)00011-8.
- [21] H. Brenner and U. Kliebsch, "Dependence of weighted kappa coefficients on the number of categories," *Epidemiology*, vol. 7, no. 2, pp. 199-202, 1996, doi: 10.1097/00001648-199603000-00016.
- [22] S. Kriglstein, M. Pohl, and M. Smuc, "Pep up your time machine: Recommendations for the design of information visualizations of time-dependent data," in *Handbook of Human Centric Visualization*, pp 203-225, 2013. doi: 10.1007/978-1-4614-7485-2_8.F
- [23] J. F. Herrmann, "Reading multiple-line graphs and graph interpretation.," Ph.D., University of Delaware, United States. Accessed: Sep. 02, 2021. [Online]. Available: <https://www.proquest.com/docview/302765558/citation/E9F02D397C0848C2PQ/1>.
- [24] G. Kim and C. Lui, "Impacts of luminance contrast on effectiveness of multiple line graphs," *Journal of Communications and Information Sciences*, vol. 2, no. 2, pp. 97-107, Apr. 2011, doi: 10.4156/ijipm.vol2.issue2.11.
- [25] S. Akbari, A. Alazemi, M. Andelić, and M. Hosseinzadeh, "On the energy of line graphs," *Linear Algebra and its Applications*, vol. 636, pp. 143-153, 2022, doi: 10.1016/j.laa.2021.11.022.
- [26] J. Asensio-Cubero, J. Gan, and R. Palaniappan, "Multiresolution analysis over simple graphs for brain computer interfaces," *Journal of Neural Engineering*, vol. 10, no. 4, Jul. 2013, doi: 10.1088/1741-2560/10/4/046014.
- [27] S. K Narang and A. Ortega, "Lifting based wavelet transforms on graphs," *In Conference of Asia-Pacific Signal and Information Processing Association*, 2009, pp. 441-444.

BIOGRAPHIES OF AUTHORS



Sutat Gammanee    received the B.Sc. and M.Sc. degrees in computer science and study in Ph.D. in information technology at King Mongkut's University of Technology North Bangkok. He became an Assistant Professor of information technology in 2019. He is lecturer at Kanchanaburi Rajabhat University, Thailand. His current research interests include Natural language processing, Machine Learning. He can be contacted at email: sutat@kru.ac.th.



Sunantha Sodsee    received the Doctoral Degree in Engineering, Chair of Communication Network, Faculty of Mathematics and Computer Science, FernUniversität in Hagen, Germany. She has been an Assistant Professor with King Mongkut's University of Technology North Bangkok. She is currently the Dean of Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Thailand. Her current research interests include Complex Network Routing, Data Communication, Recommender Systems. She can be contacted at email: sunantha.s@itd.kmutnb.ac.th.