Research *in* Statistical Sciences, pp. 107-117.

Chapter 12

A STUDY ON FACTORS AFFECTING MOBILE NETWORK PREFERENCES IN PENANG ISLAND USING BAYESIAN NETWORK

HONG CHOON ONG and KOK BAN TEOH

School of Mathematical Sciences, Universiti Sains Malaysia
11800 USM Pulau Pinang, Malaysia
hcong@cs.usm.my, derickteoh_1128@hotmail.com

Malaysia's mobile network services exhibit a very tense competition among all service providers. Hence, it is crucial to identify the factors which influence the business of mobile network company. Eight different structural learning algorithms from the bnlearn package in the R programming language are used to run all the eight structural learning algorithms. Network scores are used from the package are used to identify the best fitted network. The arc strength is determined in the final network to know the strength of each relationship. From our result, the network using both the Hill Climbing and Tabu Search algorithm is found to be the best network in this study. Furthermore, race is identified as the main factor which affects the Penang citizen mobile phone users towards their choices on mobile network services.

Keywords: Structural learning; mobile network preferences; network scores; Bayesian network.

1. Introduction

The mobile network industry in Malaysia is a fast growing sector. Therefore, the nation's economic development has very much contribution from this sector. Sany *et al*. (2011) had conducted a few studies focusing on the effects of customer satisfaction on customer loyalty in Malaysian mobile network service providers. It is concluded that customer satisfaction possesses positive relationship with customer loyalty. However, it will be an interesting issue to investigate on the factors which affect the choices of Malaysian mobile network service providers because this will not only increase the customer loyalty but also help the marketers of service providers to design the better marketing programs which deliver better customer value and increase the business of a particular mobile network company.

Bayesian network (BN) is a multivariate statistical model which belongs to the family of probabilistic graphical models. These graphical structures are particularly useful to represent the relationships about an unknown domain. Therefore, the random variables are represented by every node in the graph while the edges that link the nodes represent the probabilistic dependencies among the random variables.

Nodes represent the random variables of interest while the directed edges represent direct dependence among the random variables that are linked with arrows. The absence of an edge between two random variables might indicate that a conditional independence relationship exists between two random variables.

Bayesian networks are widely used in many fields like gene expression analysis (Friedman *et al*. 2000), financial risk management (Neil *et al*., 2005), environment protection (Henriksen & Barlebo, 2007) and breast cancer prognosis and epidemiology (Holmes & Jain, 2008). Thus, Bayesian network is shown to be a powerful tool for casual relationship modeling and probabilistic reasoning as indicated by Tang *et al*. (2010). This is thus a motivation to study on the factors affecting the mobile network preferences.

In this paper, the structural learning algorithms are used to determine the direct and indirect factors that affect the choices on mobile network service provider in Penang Island among the mobile phone users. Network scores are used as comparison criterion to decide the best algorithm as the final network while the arc strength is utilized to determine the strength of relationships between two linked variables in the final network.

2. Basic Concepts

2.1 *The data set*

The sample for this study is randomly selected with stratified sampling method from Penang island mobile phone users aged from 15 to 55 through questionnaire. There are a total of 1000 samples. 14 variables are used and the details are in Table 1.

Table 1. 14 variables in the data.

Variable	Possible values	Description
Gender	2	2 types of gender: male and female
Race	4	4 types of race: Malay, Chinese, Indian and others
Age	4	4 groups of age: 15-25, 26-35,36-45 and 46-55
Education	4	no formal education, UPSR / PMR / SPM, Diploma / STPM / Matriculation / A-levels / Foundation and Bachelor's Degree / Professional qualifications / Postgraduate
Occupation	5	government sector, private sector, self employed, student and not working
Salary	4	no salary, less than RM2000, RM2000-RM3000 and more than RM3000
Phone	6	Nokia, Samsung, Motorola, Sony Ericsson, iPhone and others
Network	4	Celcom and others
FamilyFriends	5	5 rating scales on the degree of agreement in terms of the influence by family and friends
Coverage	5	5 rating scales on the degree of agreement in terms of the influence by network coverage
Cost	5	5 rating scales on the degree of agreement in terms of the influence by cost of charges
Service	5	5 rating scales on the degree of agreement in terms of the influence by customer service
Convenience	5	5 rating scales on the degree of agreement in terms of the influence by friendliness of services
Marketing	5	5 rating scales on the degree of agreement in terms of the influence by marketing's promotion

2.2 *Structure learning algorithms*

To reduce the complexity of data while still learning the correct network, several related learning algorithms were developed. These algorithms are divided into three categories:-

2.2.1 Constraint-based structure algorithms

Constraint-based structure learning algorithms are the ones which learn the network structure via probabilistic relations implied by the Markov property of Bayesian networks with conditional independence tests and then developing a graph (Scutari, 2010).

- Grow-Shrink (GS)
  It is the simplest Markov blanket detection algorithm (Margaritis, 2003) used in a structure learning algorithm. Grow-Shrink algorithm possesses a grow and a shrink phase.
- Incremental Association Markov Blanket (IAMB)
  This is based on a two-phase selection scheme that is a forward selection followed by an effort to remove false positive (Tsamardinos *et al*., 2003). It follows two-phase structure as in GS and uses one dynamic heuristic in the growing phase so that the static and inefficient heuristic of GS is improved.
- Fast Incremental Association (Fast-IAMB)
  It is a type of IAMB which reduces the number of conditional independence tests by using speculative stepwise forward selection (Yaramakala & Margaritis, 2005).
- Interleaved Incremental Association (Inter-IAMB)
  Another type of IAMB which avoids false positives in the Markov blanket selection phase by using forward stepwise selection (Tsamardinos *et al*., 2003; Ge *et al*., 2010).

2.2.2 Score-based structure algorithms

Score-based structure algorithms are the approach which assigns a score to the entire Bayesian network and maximize it with some heuristic search algorithms. Greedy search algorithms such as hill-climbing and tabu search are the common choice

- Hill-Climbing (HC)
  Kojima *et al*. (2010) indicated that HC algorithm is utilized to find the local optima and upgraded versions of this algorithms leads to improving the score and structures of the results.
- Tabu Search (Tabu)
  This algorithm is a modified HC which is able to escape local optima via choosing a network that decreases the score function minimally (Scutari, 2010).

2.2.3 *Hybrid structure algorithms*

Hybrid structure algorithms are a combination of constraint-based and scoring-based approaches which use conditional independent tests to reduce the search space and also network scores to identify the optimal network in the reduced space simultaneously.

- Max-Min Hill Climbing (MMHC)
  This algorithm is a combination of the Max-Min Parents and Children (MMPC) algorithm which restricts the search space and the IIC algorithm which finds the optimal network structure in the restricted search space (Scutari, 2010).
- General 2-Phase Restricted Maximization (RSMAX2)
  This is a more general algorithm of the Max-Min Hill Climbing (MMHC) in terms of implementation. It can use any combination of constraint-based and score-based algorithms (Scutari, 2010).

2.3 *Network Scores*

Score functions predict the fitting of the network and the following are used.

Firstly, log-likelihood (loglik) score is used. Secondly, Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are adopted for learning Bayesian network in this study as well. Both these network scores are independent of data and only depend on the number of random variables and the structure of Bayesian network. The network score for AIC is defined as follows:

$$AIC = -2\log p\left(L\right) + 2p \tag{1}$$

where L represents the likelihood under the fitted model and p refers to the number of parameters in the model. AIC is used to select the model which minimizes the negative likelihood penalized by the number of parameters as stated in the equation (1). On the other hand, the network score for BIC is defined in the following equation:

$$BIC = -2\log p\left(L\right) + \log n \tag{2}$$

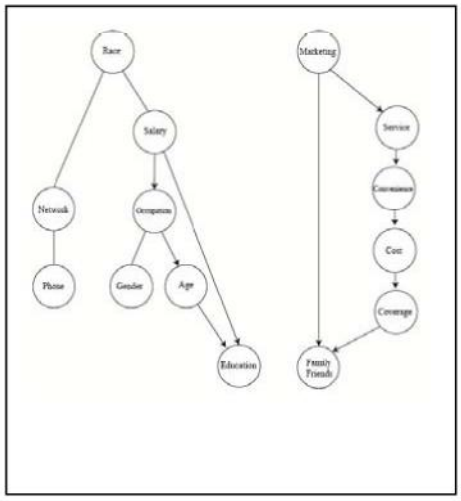
where n represents the sample size. BIC is derived within a Bayesian framework as an estimate of the Bayes factor for two competing models (Schwarz, 1978).

Bayesian Dirichlet Equivalent (BDE) is based on the Bayesian Dirichlet (BD) metric (Cooper and Herskovits, 1992) to estimate and evaluate a given dataset network.

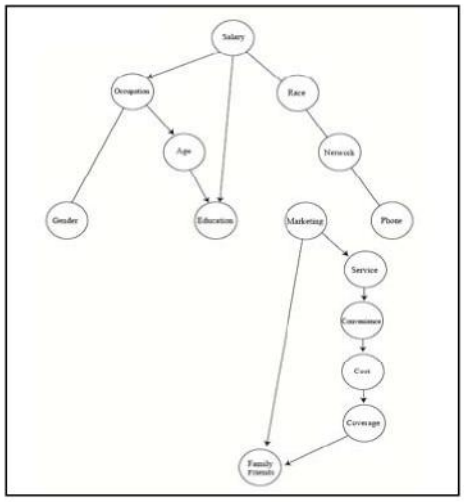
The logarithm of the K2 scores (K2) which is another Dirichlet posterior density proposed by Cooper and Herskovits (1992) is adopted in this study as well.

3. Results and Discussion

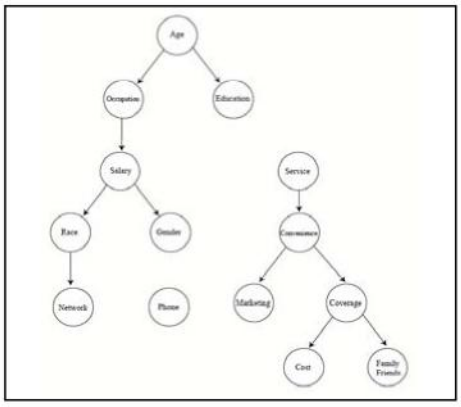
Eight different networks obtained by using the bnlearn package are shown in Figure 2. The existence of arc between two variables means that there exists a direct dependence relationship between the two variables. Also, Ge *et al*. (2010) proposed that the absence of arcs indicates the existence of conditional independence relationships.



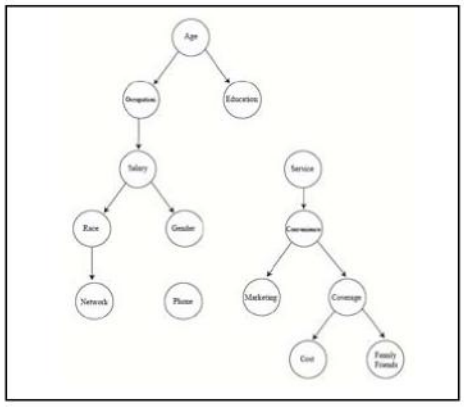
(a)



(b)



(c)



(d)

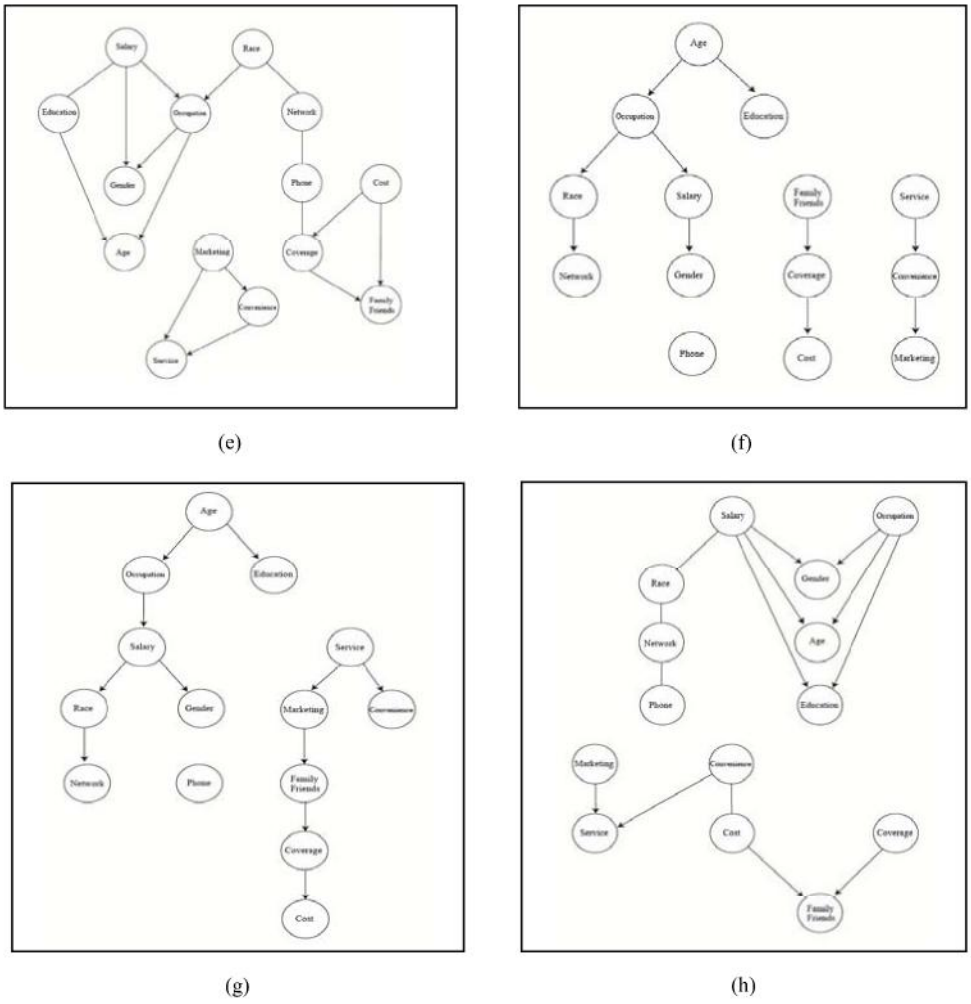


Figure 2. Networks structures learned by selected algorithms. (a) Incremental Association Markov Blanket, (b) Interleaved Incremental Association, (c) Hill Climbing, (d) Tabu Search, (e) Grow-Shrink, (f) General 2-Phase Restricted Maximization, (g) Max-Min Hill Climbing, (h) Fast Incremental Association.

The number of common edges or links and directed arcs among all learning algorithms are given in Table 3. The number of "edges" indicates the number of common links that exists regardless of its direction. Besides, the common number for links with direction that exists in between the learned networks is indicated by the number of "arcs". The main diagonal in Table 3 shows the common number of edges/arcs that exists in the network from each algorithm.

Table 3. Number of edges/arcs between each pairs of the learned networks.

	iamb	inter.iamb	hc	tabu	gs	rsmax2	mmhc	fast.iamb
iamb	14/10	14/10	8/3	8/3	11/5	7/2	10/3	10/4
inter.iamb	-	14/10	8/3	8/3	11/5	7/2	10/3	10/4
hc	-	-	11/11	11/11	9/2	9/8	9/8	6/2
tabu	-	-	-	11/11	9/2	9/8	9/8	6/2
gs	-	-	-	-	16/12	10/1	9/1	10/7
rsmax2	-	-	-	-	-	10/10	8/8	5/1
mmhc	-	-	-	-	-	-	11/11	7/1
fast.iamb	-	-	-	-	-	-	-	14/10

As a result from Table 3 that there are a number of common edges and arcs contained in every pair of learned networks. Hence, these common edges that exist in all networks suggest that there are strong relationships between the variables linked with these common edges. Therefore, these edges are between age and occupation, race and network, service and convenience, coverage and family & friends. Besides, the common edges that can be found in seven models denoting also strong relationships between the variables linked with these edges. These edges are between age and education, occupation and salary, coverage and cost. In addition, the edges between salary and gender, salary and race are only found as common edges in six models. Hence, all these nine edges are taken as common edges and they are shown in Figure 4.

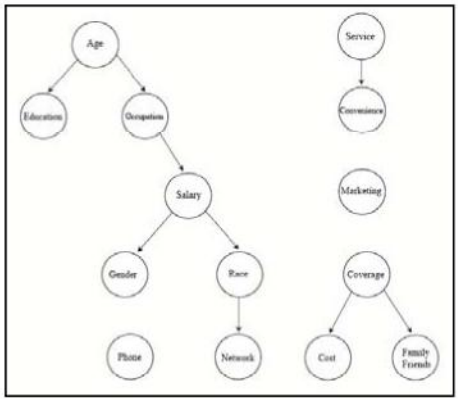


Figure 4. Common edges of all the learned networks.

In Figure 4, the nine common edges with directions can also be identified because of the logical relationships of these variables. For example, it is logical and reasonable to know that the age of a mobile phone user in Penang Island is related to the person's level of education and types of his occupation. It is usually true that a fifteen year old individual will usually be in the education level of secondary school and his or her occupation will usually be a student. Hence, the structure of the relationship between

these networks can be constructed by using the direct dependence relationship among the variables that are connected with these nine common edges.

Then, all the learning algorithms are run again and the common edges in Figure 4 are set to be the white list while the direction of the undirected edges is determined.

Consecutively, the values of the different scores for all the networks considered are obtained and shown in Table 5. The highest scores of these algorithms are represented by the bold and italic numbers.

Table 5. The results of scores of all learned networks for each algorithm.

	bde	k2	loglik	aic	bic
iamb	-16311.51	-16172.54	-15516.07	-15896.07	-16828.54
inter.iamb	-16311.51	-16172.54	-15516.07	-15896.07	-16828.54
hc	***-16151.08***	***-16136.40***	-15729.67	-15905.67	***-16337.55***
tabu	***-16151.08***	***-16136.40***	-15729.67	-15905.67	***-16337.55***
gs	-16386.30	-16249.65	-15623.92	-15978.92	-16850.05
rsmax2	-16218.25	-16208.07	-15829.23	-15989.23	-16381.85
mmhc	-16255.62	-16244.22	-15866.15	-16026.15	-16418.77
fast.iamb	-16350.32	-16200.38	***-15475.15***	***-15891.15***	-16911.96

It is seen from Table 5 that the Hill Climbing and Tabu Search algorithms show a better performance in terms of all five network scores except the log-likelihood and Akaike score, which is also quite close to the highest scores of the Fast Incremental Association algorithm. Hence, the networks obtained using Hill Climbing and Tabu Search algorithms are chosen as the final network.

Before constructing the final network diagram, the arc strengths in all the edges in the final network are considered so that the strength of relationships is identified as represented in Table 6. Arc strength in the label of a score function is measured by the score gain or loss which is caused by the removal of the arc.

Table 6. Arc Strengths for the linked variables.

From	**To**	**Arc Strength**
Age	Education	-58.619681
Age	Occupation	-249.260712
Occupation	Salary	-409.172957
Race	Network	-103.311865
Salary	**Gender**	***-8.582701***
Service	Convenience	-111.205557
Coverage	Cost	-124.674523
Coverage	FamilyFriends	-38.084501
Salary	**Race**	***-17.274102***
Convenience	Marketing	-77.055932
Convenience	Coverage	-44.293219

From Table 6, it is noticed that the nodes from salary to gender and salary to race show the lowest negative values of arc strengths which are -8.582701 and -17.274102 respectively. Thus, the arcs in these two pairs of linked nodes indicate that there exists a strong relationship between them.

Ultimately, a final network diagram which shows the strength of the relationships can be constructed as shown in Figure 7.

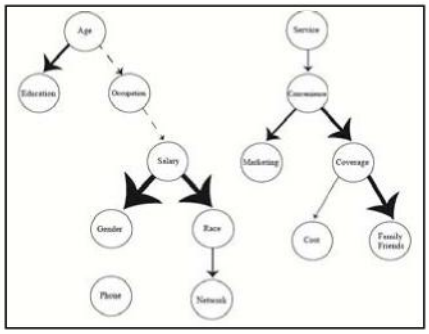


Figure 7. The final results of the score learned network using the Hill-Climbing and Tabu Search algorithms.

Based on Figure 7, the thicker edge shows the stronger relationship among the linked nodes from its arc strength. Thus, it can be concluded that an individual's salary is strongly linked to the person's gender and race. This is always true that an individual's salary is linked to the location as types of occupation will vary depending if the location is town or sub-urban. Similarly in Shaver and Tudbull (2002), it is said that the job opportunities in a particular area is shaped by the resources and industries that surrounds the area while types and levels of economic activity do much to determine the incomes and living standards of the families that live there. Also, the most significant differences among the areas proposed by Shaver and Tudbull (2002) are rural and urban.

From the Department of Statistics, Malaysia (2011), female citizens in Penang island are more in town area, that is the northeast district as compared to Penang island male citizen who are distributed more in southwest sub-urban area. Also, it is observed that Penang Island Chinese are more in town area as compared to the Malays and Indians who are more in sub-urban area. So, it sounds logical that gender and races of citizen in Penang island have different job opportunities as their distributions in Penang island is significantly different between town and sub-urban areas. As a result, the salary is different according to their types of occupation from different areas and it is concluded that the salary is related to gender of Penang island mobile phone users.

Nevertheless, the racial background of Penang island mobile phone users is related to the mobile network service provider. This finding is substantiated since the distribution of race is related to the districts in Penang Island. Therefore, users will tend to choose the mobile network which is convenient to them within the district. Besides, the trend and culture of a reference group for a particular race in the context of choosing mobile networks plays a vital role too. Kumar (2008) said that cultural factor exert the broadest

and deepest influences on consumer buying behavior and cultures are not homogeneous entities with universal values. Therefore, different races have different cultures of buying behaviors. Hence, a mobile phone user in Penang Island will usually tend to choose the mobile network according to what his or her reference group of the same race uses as well as family members. This is supported by Kaapanda (2012) that customers prefer subscribing to the same service provider as their family members. It is always cheaper to use the same mobile network to connect among the family members since the call rates are less and there are free data and Short Message Service (SMS). So, this may be the reason for customers to subscribe to the same service providers as family members. Thus, the mentioned trend and culture of reference group for a particular race will determine the type of mobile network chosen. Also, it is found that a particular race of a person will automatically be attracted to their own language or celebrity used by the commercial advertisement. The credibility of celebrities is apparently different among different ethnic groups. Thus, Kaapanda (2012) stated that a mobile network service provider is very important in choosing the appropriate celebrity so that the credibility of advertisement can be enhanced. Furthermore, one will also be fascinated if the commercial advertisement depicts the culture or belief with respect to the race of an individual.

As seen in Figure 7, the absence of arc between the marketing variables and mobile network suggests that none of these marketing variables are related to the selection of mobile network service provider. So, it is indicated that all mobile network companies perform comparatively in terms of the marketing variables. Hence, marketing variables in this study have no significant impact on the choice of the mobile networks.

4. Conclusion

From the results and comparison, the network built by Hill Climbing and Tabu Search algorithms is chosen as the final network. Furthermore, arc strength is obtained and it is found that a Penang island mobile phone user's salary is strongly linked to the gender and race. Meanwhile, it is also observed that the race of a mobile phone user in Penang island is related to his or her choice on mobile network. Therefore, the mobile network's preferences of Penang island mobile phone user is directly affected by race and indirectly affected by salary.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. **19** (6), 716 - 723.

Cooper, G.F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*. **9** (4), 309 - 347.

Department of Statistics, Malaysia (2011). *Population Distribution and Basic Demographic Characteristics 2010*, ISBN: 9789839044546. Malaysia.

Friedman, N., Linial, M. & Nachman, I. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*. **7**, 601 - 620.

Ge, Y., Li, C. & Yin, Q. (2010). Study on factors of floating women's income in Jiangsu province based on Bayesian networks. *Advances in Neural Network Research & Applications*, *Lectures Notes in Electrical Engineering 2010*. **67** (9), 819 - 827.

Henriksen, H.J. & Barlebo, H. C. (2007). Reflections on the use of Bayesian belief networks for adaptive management. *Journal of Environment Management*. **88** (4), 1025 - 1036.

Holmes, D.E. & Jain, L.C. (Eds.) (2008). *Innovations in Bayesian Networks: Theory and Applications*, volume 156 of *Studies in Computational Intelligence*. New York: Springer.

Kaapanda, L.N. (2012). An evaluation of factors determining the selection of mobile telecommunications service providers in the northern region of Namibia : In : *3rd International Conference of Business and Economic Research (3rd ICBER 2012) Proceeding*, 12-13 March 2012, Bandung. Bandung : 3rd International Conference of Business and Economic Research (3rd ICBER 2012) Proceeding.

Kojima, K., Perrier, E., Imoto, S. & Miyano, S. (2010). Optimal search on clustered structure constraint for learning Bayesian network structure. *Journal of Machine Learning Research*. **11**, 285 - 310.

Kumar, C.R. (2008). *Research Methodology*. New Delhi: APH Publishing Corporation.

Margaritis, D. (2003). *Learning Bayesian network model structure from data*. Ph.D thesis, Carnegie-Mellon University.

Neil, M., Fenton, N. & Tailor, M. (2005). Using Bayesian networks to model expected and unexpected operational losses. *Risk Analysis*. **25** (4), 963 - 972.

Sany, S.M.M., Ahmed, A.M. & Norzaini (2011). The relationship between service quality and satisfaction on customer loyalty in Malaysian mobile communication industry. *School of Doctoral Studies (European Union) Journal*. **3** (0), 32 - 38.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*. **6** (2), 461 - 464.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*. **35** (3), 1 - 22.

Shaver, S. & Tudbull, J. (2002). *Literature Review on Factors Contributing to Community Capabilities*, ISBN: 0733419984. Australia.

Tang, Y., Cooper, K., Cangussu, J., Tian, K. & Wu, Y. (2010). Towards effective improvement of the Bayesian belief network structure learning : In : *International Conference on Intelligence and Securing Informatics (ISI)*, 23-26 May 2010, Canada. Canada: IEEE Press.

Tsamardinos, I., Aliferis, C. F. & Statnikov, A. (2003). Algorithms for large scale markov blanket discovery : In : *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, 12-14 May 2003, Florida. California : AAAI Press.

Yaramakala, S. & Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection : In : *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, 27-30 November 2005, Texas. Washington : IEEE Computer Society.