# Security Based Multiple Bayesian Models Combination Approach

S.RAVI KISHAN[1], S.RAJESH[2], B N SWAMY[3], J.V.D PRASAD[4]

**ABSTRACT**
Decision making in medical domain often involves incorporating new evidences into existing or working models reflecting the decision problems at hand. We propose a new framework that facilitates effective aggregation of multiple Bayesian Network models. The proposed framework aims to minimize time and effort required to customize and extend the original models through preserving the conditional independence relationships inherent in two or more types of Bayesian network models. We present an algorithm to systematically combine the qualitative and the quantitative parts of the different Bayesian models. Combination of Bayesian models involves integrating both structural and parameters of different models. We also describe how effective the presented algorithm and it can reduce total computational complexity
**Keywords:** Data privacy, Bayesian networks (BN), privacy-preserving data mining.

## 1. INTRODUCTION

A Bayesian network is a directed acyclic graph (DAG), which encodes the causal relationships between particular variables, represented in the DAG as nodes. Nodes are connected by causal links - represented by arrows - which point from parent nodes (causes) to child nodes (effects). Belief networks have been found to be useful in many applications related to reasoning and decision-making. Bayesian network (BN) is a powerful knowledge representation tool for uncertainty management and decision making. In a rapidly changing world, integrating new evidences or new fragments of knowledge in the form of multiple new models is challenging. Biomedical problems usually involve a large number of variables, complex relationships among the variables, and numerous parameters. The different evidences or models to be integrated may be from different sources, in different modeling languages, or differ in structure or in parameter, even if they may be derived from the same data sets or from experts in the same domain. Assume that a novice surgeon is planning to perform a head operation. However, he is not confident of his knowledge on nerve damnification and skin damnification. In order to make a sound decision, he needs to acquire additional knowledge related to possible nerve damnification and skin damnification in a head operation. Therefore, he seeks help from dermatology textbook and a neurology data set. Three Bayesian networks are modeled from the different sources: the dermatology textbook, the neurology data set, and the surgeon's own domain expertise respectively. There are some common variables in all the three networks, or between only two of the networks.

Combining multiple Bayesian probabilistic graphical models in a uniform manner is a tedious task; heterogamous models representing similar or overlapping pieces of information from possibly different viewpoints need to be combined both qualitatively and quantitatively. Some other efforts address topology combination in BNs, in which only two models can be combined at one time. Besides the difficulty in scaling, the resulting model can also be influenced by the order of combination, if there are more than two models to be combined. In this paper, presenting a security based Multiple Bayesian **M**odel **C**ombination (MBMC) framework to address both qualitative and quantitative combinations of an arbitrary number of graphical models simultaneously.

## 2. PROBLEM FORMULATION

Consider two parties owning private data. Those parties wish to learn the Bayesian network on combination of their databases. To achieve this one party send this data to the other in encrypted form other party receives and decrypts it. The received data is merged with the local data. The resultant data is input for the BN learning process. Learning process involves 2-steps namely structure learning and parameter learning. For computation of parameters we make use of scalar product protocol to compute parameters in a secured and privacy preserved manner.
Sending data from one party to another gives not only a chance to learn more information by the other party, it causes a breach for some security settings. Communication overhead is also caused because of sending full data.
To overcome the above mentioned limitations, we send locally learned BN model information to other in an encrypted form. Other party receives and decrypts it.

Manuscript received November 5, 2009
Manuscript revised November 20, 2009

Second party combines the received BN model information with the local BN model to produce a global BN model. Here, we are learning global BN model from local BN models and not from training data. Here we are addressing both qualitative and quantitative combinations of multiple Bayesian models. Each BN model consists of two parts: Qualitative part that represents the structure of the network and the dependencies among the variables (tree structure); the quantitative part that numerically represents the joint probability distribution over these variables. Aggregation of Bayesian networks involves Qualitative and Quantitative combination of Bayesian models. Qualitative combination involves structural combination. Quantitative combination involves combining parameters of nodes. Figure 2 shows the overall approach for combining multiple Bayesian models.

Qualitative combination of Bayesian network involves explicitly combining two BN DAGs into a single DAG, or fusing the two topologies. There are source network ($B_S$) and target network ($B_T$), $B_{ST}$ is the resultant of combination of both networks. The combination process fuses the structural information of target network into source network. The combination model $B_{TS}$ is not equivalent to $B_{ST}$.
A set of Bayesian networks that need to be integrated into one model is defined as S={B1,B2,….Bn }. *N* is the number of networks for combination. The integration process is conducted incrementally. At first, $B_1$ and $B_2$ *are* combined and the result is referred as $B_{12}$ Then, $B_3$ is combined with $B_{12}$ and it produces. $B_{123}$. Like this, the combination procedure is continued until the last Bayesian network is fused into the integrated model. The final model is referred as global Bayesian network or $B_{123....n}$

If there are N Bayesian networks, the problem is how to determine the order of Bayesian networks for combination. According to the order, the result of integration process is different and the number of edges for the model varies. For example, the result of $B_{123..N}$ and $B_{213...N}$ is not the same. For N Bayesian Networks, there are N! Cases of combination.

## 3. COMBINING MULTIPLE BAYESAIN MODELS

If there are a number of authors of Bayesian networks about the same domain, there could be a variety of models that describe similar things. Because they have different expertise about the domain, it is better to integrate them into a single model. The easiest way of combining them is to use intersection and union operations. In the case of intersection operator, the common structure of all Bayesian networks is used as a global Bayesian network. On the other hands, union operator put all of the edges and variables of the networks into a global network. Combination of multiple Bayesian models carried out 2 steps 1) Structural combination (Qualitative combination) 2) parameter combination (Quantitative combination). Qualitative combination means merging of nodes and edges in the two networks. Quantitative combination means combining conditional probability tables (CPTables.i.e.) the numerical representation of conditional probabilities.

### 3.1 QULITATIVE COMBINATION

Qualitative combination of Bayesian network involves explicitly combining two BN DAGs into a single DAG, or fusing the two topologies. There are source network ($B_S$) and target network ($B_T$), $B_{ST}$ is the resultant of combination of both networks. The combination process fuses the structural information of target network into source network. The combination model $B_{TS}$ is not equivalent to $B_{ST}$. We are classifying edges into DIR, REV, EQ. DIR means edge of target network can be directly inserted into source network. REV means edge needs to be reversing the direction. EQ means edges of two variables that have same topology value.

The algorithm has six steps.

1) Calculating the topological values of the variables in the source network.
2) Classifying the categories of the edges into DIR, REV and EQ.
3) For each edge in the REV, applying reversing operation to the target network and classifying the new edges from the operation into the three categories are done.
4) Inserting edges in the DIR into the source network from target network.
5) For each edge in the EQ, add the edge into the network and update the topological value of source network (some edges in EQ is transferred to the DIR).
6) After clearing all the edges in the three categories, the process is finished.

### 3.2 QUANTITATIVE COMBINATION

Quantitative combination refers to combination of CPT (conditional probability tables. This can be achieved by using weighted combination methods. By this method we can construct a standard CPT filled with point probability distributions in the resulting BN. This procedure relies on the observation that it is not necessary to have the actual data to learn a BN; it is sufficient to have their empirical distribution.

We can parameterize the network in top-down fashion by first computing the distribution over the roots, then joints over the second layer variables together with their parents, etc. The conditional probabilities can be computed by dividing the appropriate marginals (using Bayes Law). In many cases, that would require only local computations in sources' BNs. Since we are making only local changes to the structure, only a few parameters will need updating. If an arc is added or removed, we only need to recompute new parameters for the child node, and if an arc is switched, we only need to recomputed parameters for the two nodes involved.

**Aggregative Parameter Learning Algorithm:**

1. Learn local BN $B_{local}$ involving the variables observed at each local site.
2. Compute likelihood of variables in cross set (CS) based on local BNs.
3. Transmit the index set of low likelihood samples from each local site to the central site
4. Compute the intersection of these index sets at central site.
5. Transmit variables in cross set corresponding to the intersection set from all local
sites. At the central site, a limited number of observations $D_{obs}$ of all the cross
set variables are now available.
6. Learn a new BN $B_{new}$ using $D_{obs}$ in central site.
7. Set the cross node parameters using $B_{new}$ and local node parameters using $B_{local}$.

In the above algorithm, the selection of samples to be transmitted is based on the joint distribution of the cross set variables (not the joint distribution of all site variables). We divide the local variables into CS *(*cross set variables*)* and LS *(*local set variables*)*. The CPT of cross variables is entirely determined by the cross set.
Finally, a collective BN can be obtained by taking the union of nodes and edges of the local and the non local BNs, along with the conditional probabilities from the appropriate BNs. Probabilistic inference can now be performed based on this collective BN. Note that transmitting the local BNs to the other site would involve a significantly lower communication as compared to transmitting the local data.

## 4. EXPERIMENTAL RESULTS

In the present set up we have conducted experiments on the different types of models with different no of nodes and edges. The implementation results are as follows .the total combination time is getting reduced by the use of topological combination methods. The total time complexity is also drastically reduced because of sending common intersect variables rather than all variables presented in the given remote site. The following results will address the different experimental values

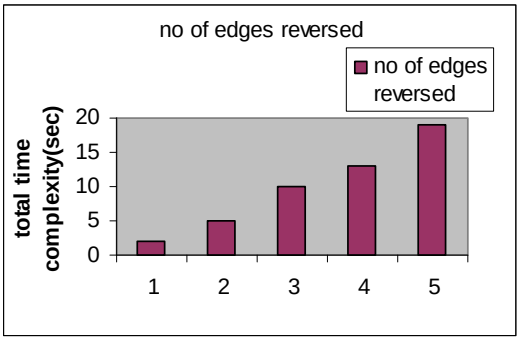


**Fig 4.1: total time complexity with respect to number of nodes**

The above figure 4.1 shows complexity of learning a combined model with total no of arc reversed to avoid cycles after combination of the model. By adapting the fusion method of graphs and classification of edges we have reduced the arc reversal and addition of new edges operation and complexity too.

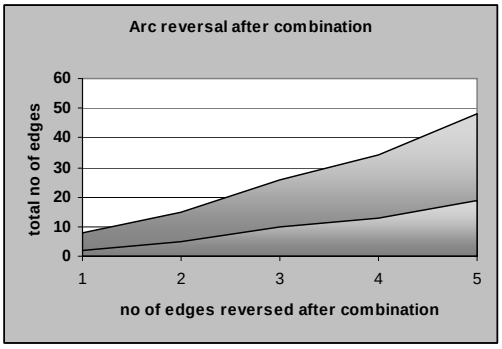


**Fig 4.2: Total no of edges reversed after combining the multiple models**

Fig 4.2 shows total no of edges reversed after combining multiple models, so that to avoid cycles to maintain the DAG property of Bayesian model. The layer of the graph also shows addition of new edges for arc reversal operation of the graph. The following fig 4.3 represents the results of edge reversals, complexities and error rate with different no of nodes and edges.

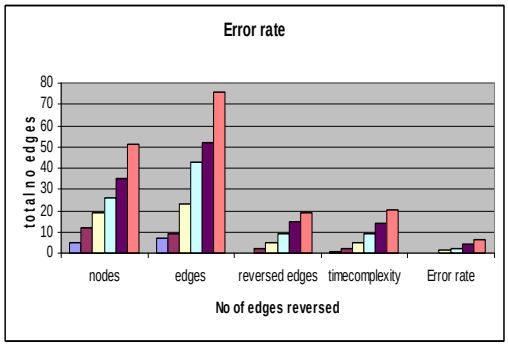


**Fig 4.3: Total no of edges reversed and Error rate of the model**

The table 4.1 depicts different combination of models with different no of nodes, edges and no of arcs reversed, computational complexity time and also error rate for the learning of a particular model.

| No of nodes | No of edges | No of arcs reversed | Combination time | Total Complexity |
|---|---|---|---|---|
| 5 | 9 | 3 | 0.852 | 2.452 |
| 12 | 20 | 6 | 1.526 | 4.325 |
| 22 | 34 | 14 | 2.963 | 5.621 |

**Table 4.1: complexities of various BN models**

## 5. CONCLUSION

We have proposed a framework to extend existing research in combining probability distributions and aggregating probabilistic graphical models. Our framework supports combinations of BNs. The resulting model would have the combined edges of both models. In this way, no cycle will be generated in the procedure of combination and DAG structure can be preserved. The target variable ordering generation can be assumed by user it self. Here we are not guaranteed to yield optimal solution as it is a NP-hard problem [10]. We have also implemented a prototype system to evaluate the feasibility and potential of the proposed approach with a set of experiments in a real medical domain. This work aims to support knowledge combination over a wide spectrum of decision problems. Our future agenda include further improvement on the target variable ordering for better combination of multiple models to yield optimal results and dynamic combination of multiple probabilistic models such as influence diagrams, decision trees.

## REFERENCES

[1] R. Chen, K. Siva Kumar, and H. Kargupta, "Learning Bayesian network structure from distributed data, "Proceedings of the 3rd SIAM International Data Mining Conference, pp. 284-288, 2003.

[2] R. Chen, K. Siva Kumar, and H. Kargupta, "Collective mining of Bayesian networks from distributed heterogeneous data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[3] J. D. Sagrado, and S. Moral, "Qualitative combination of Bayesian networks," International Journal of Intelligent Systems, vol. 18, no. 2, pp. 237-249, 2003.

[4] R. T. Clemen and R. L. Winkler," Combining probability distributions from experts in risk analysis", Risk Analysis, 19(2): 187–203, 1999.

[5] Clemen, R.T. and Winkler, R. 1999," Combining probability distributions from experts in risk analysis", Risk Analysis, 19:187-203.

[6] Pedrito M., Reid II and Urszula C., 2001. "Aggregating Learned Probabilistic Beliefs", proceedings of Uncertainty of Artificial Intelligence 2001, pages 354-361.

[7] N. Friedman, and M. Goldszmidt, "Sequential update of Bayesian network structure," Proceedings of 13th Conference on Uncertainty in Artificial Intelligence, 1997.

[8] D. M. Pennock and M. P. Wellman," Graphical representations of consensus belief", In Proceedings of UAI'99, pages 531–540, 1999.

[9] Kyung-Joong Kim and Sung-Bae Cho, " Evolutionary Aggregation and Refinement of Bayesian Networks", 2006 IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006

[10] Ng, K. C. and Abramson, B. 1992. Consensus diagnosis: A simulation study. IEEE Transactions on Systems, Man, and Cybernetics, 22:916-928.

**S.RaviKishan** received his B.Tech from Anna University, Chennai and completed post graduation from Jawaharlal Nehru Technological University, Hyderabad. He is currently pursuing Ph. D from JNTU, Anantapur and working as Assistant Professor in V R Siddhartha Engineering College, in the Department of Computer Science and Engineering, Vijayawada, Andhra Pradesh. His research interests include Data Mining and Data Warehousing. He has more than ten years of experience in teaching and in research. Education (ISTE) and also member of Computer Society of India (CSI). He has many publications in National and International conferences.

**S.Rajesh** received his B.Tech from Acharya Nagarjuna University, Andhra Pradesh and completed post graduation from Jawaharlal Nehru Technological University, Hyderabad. He is currently pursuing Ph. D from JNTU, Kakinada and working as a Lecturer in V R Siddhartha Engineering College, in the Department of Computer Science and Engineering, Vijayawada, Andhra Pradesh. His research interests include Data Mining and Data Warehousing, Bio Informatics and Fuzzy logic. He has more than five years of experience in teaching and in research.

**B.N.Swamy** received his B. Tech from Pondichery University and completed post graduation from SRM University, Chennai. He is currently working as Sr.Assistant Professor in PVP Siddhartha Institute of Technology, in the Department of Computer Science and Engineering, Vijayawada, Andhra Pradesh. His research interests include Data Mining and Data Warehousing, Computer Networks and Network security. He has more than four years of experience in teaching. He is the member of Indian Society of Technical Education (ISTE) and also member of Computer Society of India.

**J.V.D.Prasad** received his B.Tech (CSE) from JNTU, Hyderabad. He is currently pursuing M.Tech(CSE) from Acharya Nagarjuna University, Andhra Pradesh and working as a Lecturer in V R Siddhartha Engineering College, in the Department of Computer Science and Engineering, Vijayawada, Andhra Pradesh. His research interests include Data Mining and Data Warehousing.