Determining Probabilities of Handwriting Formations using PGMs

Pratik Pravin Kubal
Department of Computer Science
University at Buffalo
pkubal@buffalo.edu

Abstract

Handwriting Formation is a vast field and every researcher uses different features to carry out various tasks. In this work, we determine model Probabilistic Graphical Models over different styles of modelling features in handwriting formations by two previous works. While performing the former task we survey methods used in Probabilistic Graphical Model settings and various challenges related to that, bias. In the latter part of the paper, we discuss the possibility of using Distant Supervision for one of the data absent work.

1 Introduction

Probabilistic Graphical models are a form of Generative models which involve probability distributions and factors. In the work, we survey different concepts involved in Probabilistic Graphical Models in building upon two relevant works in the domain of Handwriting Formations. In the following sections we first familiarize with the methods used in the paper and then use these methods in experimentation over 4 tasks of constructing candidate networks by structure learning through correlation between variables in 'th' data; finding the best network through data generation by ancestral sampling; using 'and' data set for structure learning,parameter estimation; and converting the Bayesian networks to Markov Models through process of moralization and running Inferences over one best network from 'th' work and its Markov Model, and Bayesian Network and Markov model of 'and' data set.

2 Relevant Work

Muehlberzer et. al paper[3] is our main reference. This paper surveys 200 individuals and while using Palmer copybook as a template to categorize their handwriting according to the most common pair of letter *'th'*. We extend this work by using the Conditional Probability distributions(CPD) tables and running three tasks over this(Structure Learning, Inference). Furthermore, we are going to make the same assumption as the paper assumes any variable not defined in the CPD table is independent. Due to the exponential search space to find the structure based on CPDs and having no data available, we limit ourselves to 10 graphs and limit up to 2 parents for any given node. The Features for this study are given in table 3.
Another research was done by Srihari et. al in terms of Forensic/Questioned Document Examination. In this study, 'and' as a word was used (Like the previous work used 'th'). There were 9 features, which can be seen in table 4. Unlike previous work, we have a data available for Structure learning, Inference and Parameter estimation for Bayesian Network and Markov Network.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

3 Methods

3.1 Limitations

The first limitation is that any probabilistic graphical model can be connected by connecting all nodes to all other nodes, this might model the ground truth perfectly, but the model is too complex to construct and run inferences on it. Another reason why this hypothetical scenario might not be possible because there are some random variables which are independent of each other. In our case, one such example could be x1(Height Relationship of t to h) is independent of x3(Shape of Arc of h) based on the assumption given in the paper(CPD doesn't exist). Furthermore, even after accounting for Independence the number of the possible graphs is exponential. Therefore we take only 10 graphs for this project. Furthermore, we stick to classical or conventional Bayesian networks as Directed Acyclic graphs and do not take cycles. Instead, we decompose the cycle into two networks.
The second limitation in this work is complex functions required to derive the conditional probability distributions(CPDs) of the child given parents more than one. Due to time constraints, we restrict ourselves to V-Structures, that is, any given node can have a maximum of two parents.
The third limitation is Bias in evaluating the models. The K2Score will be biased towards the structure for which the data was sampled or generated. Also, we don't have any data relating to the features, except 'and' data set(Which could be used, discussed later in Section 6. Ideally to generate data free of Bias we have to take into consideration n number of models(Exponential case) but due to time constraints to implement the method to generate a fair approximation on the number of nodes, we just consider 3 networks. However, even after using this method selection of these three networks is biased, See 6.

3.2 Independence between Random Variables

We use the correlation between the variables with a threshold as a measure for approximating the Independence relationships.

$\sum_{I=0,J=0}^{i,j} \left| P(X_I|Y_J) - P(X_I)P(Y_J) \right|$

The Independence can then be found using appropriate threshold following the first and second limitation discussed, See Section 3.1.

3.3 V-Structure

V Structure is an interesting structure, observing the central node makes the nodes on two arms of V independent. However, the implementation to find the V-Structure is complex. The method which we have used to find the V-Structure consists of two parts.
First, it naively classifies the relationship between the three variables as Dependent if there is a CPD already defined and just a lookup would find the value, Bayesian if there is an inverse relationship defined and needs Bayes theorem to find the value, independent if both are independent and a lookup on marginal probabilities would get this value. After three helper functions for dictionary switching, we finally get a function to get any given relationship between two variables.
Second, using the above methods and the formula given below, it finds the CPDs,

$P(Y|X1,X2) = \frac{P(X1|Y)P(X2|Y)P(Y)}{P(X1,X2)}$

However, we have found that the above function doesn't give valid CPDs. There can be two reasons such as inaccurate data or the assumption that the nodes not defined in the CPD are not independent. In such cases, we Normalize the CPD to get a valid CPD.
Another problem which we faced was a lookup of 0 values. These created a problem during getting the V-Structure of some nodes such as (x6 -> x2 <- x3). To solve this we simply added a very small number whenever 0 was returned from the look-ups. We created a pair of Lambda Functions for this purpose. Another ambiguity arises when finding the denominator for the V-Structure. Sometimes, in some scenarios, the CPD is defined, but while thresholded result assumes these as independent, we found that both outcomes are consistent with Best network in Section 5 when assuming them as independent and looking up just the marginal table and using the CPDs if defined. For accuracy purposes and generalization of V-Structure function, we divide the denominator lookup into two cases.

3.4 Evaluation Methods

We use K2 score as evaluation metric to find best Bayesian model for 'th' data set and structure learning for 'and' data set. K2 score is a scoring function which rates each network structure based on a data set given. Due to possible Bias in Conventional methods as discussed in 3.1 we use sample data from three networks as the data set and shuffle this data set as a parameter for K2Score. We evaluate the structures on this shuffled data set. We use model 1,5,and 8 according to three categories which we have experimented on in 4.1

3.5 Ancestral Sampling

Ancestral sampling is forward sampling where after a topological sort, the marginal nodes are sampled, based on the state or outcome of these marginals, child nodes are sampled. When we sample all the nodes in the network, we get one data sample. For this project we use 10000 samples for each data set network(The three networks discussed in 4.1).
We use forward_sample method from PGMPY library to implement ancestral sampling.

3.6 Inferences in Probabilistic Graphical Models

There are two types of inferences possible for PGMs: Exact Inferences and Approximate Inferences. Exact inferences are tied to the complexities of the network, therefore run into time complexity issues on dense networks. In this case, since we have a fairly small network compared to most practical applications, we use variable elimination inference method from exact inferences.
Variable elimination in short sums out or pushes in the desired random variables onto the Bayesian chain product obtained from the network. We run a MAP (maximum a posteriori) query to find the high probability assignments on different models('th' data set and 'and' data set for Bayesian and Markov network) based on Variable elimination. Furthermore, we also find the joint distribution to find the low probability 'th' for same settings given above.

3.7 Structure Learning

There are two techniques for structure learning, Score-based structure learning and Constraint-based Structure learning. The score-based structure learning algorithms exhaustively or for a given set of edges and nodes evaluate the score based on the data if they are independent. While Constraint-based methods find various tests such as correlation, cross entropy or chi-square independence tests.
Hybrid methods such as Max-Min Hill-Climbing(MMHC) combine both constraint-based and score based methods. It first learns skeleton based on constraint-based methods and orients the edges using score-based methods. This avoids the exponential space of possible graphs by moving in the right direction.
For 'and' data set structure learning we used Hill Climbing and MMHC algorithm for finding the best structure scored by K2Score. The documentation of pgmpy doesn't give any references to which type of Hill-climbing it uses(There are many variants), but we assume that the algorithm is greedy and starts from a base state having disconnected nodes and tries to maximize a scoring mechanism by connecting the edges. The scoring mechanism which we are using is K2Score. Progressively, the hill-Climbing algorithm converges to a single structure. The Hyperparameter for this algorithm is max_indegree which we vary and check the score.

3.8 Parameter Estimation

To estimate the conditional probability distributions, the method depends on the network which we want to create. For Bayesian Networks both closed form and iterative methods exist, however for Markov networks there are only Iterative solutions. One can create a Bayesian network with closed-form solution and then convert it to Markov, however as discussed by Daphne Koller in her book [1], not every Bayesian network can be converted into a faithful Markov Model. The next section discusses this in detail.
For this project we are using Maximum Likelihood Estimator for parameter estimation for the CPDs for 'and' data set.

Table 1: Models in 'th' study

Name	Threshold	K2Score
Network 1	~0.13	-190217.5323
Network 2	~0.13	-190285.8216
Network 3	~0.13	-191041.1161
Network 4	~0.16	-190110.1044
Network 5	~0.16	-190120.9370
Network 6	~0.16	-190205.5980
Network 7	~0.14	-189715.7333
Network 8	~0.13	-189588.7209
Network 9	~0.13	-190404.7179
Network 10	~0.13	-189563.1522

3.9 Moralization

The formal definition of the moral graph seems sufficient to understand the process of moralization. The Moral graph will have an undirected edge between two node X and Y if

1. There is a directed edge between them (In any direction).
2. X and Y are both parents of the same node.

One word of caution is that moralizing leads to loss of some independence properties such as V Structures. In our case,considering V-Structure x4 -> x6 <- x1 in Bayesian Network,see network 10 in Figure 1, in Markov network, see Figure 5 it can be seen that there is a new undirected edge from x1–x4, which given x6, they will still be dependent.

4 Experimentation

4.1 Models in 'th' Study

We discuss the models which we have used in this project. See table 1 for the thresholds and K2Scores. The first type of models are constructed by using the information about the correlations in the paper to model the network. After computing the correlations we can eliminate some of the nodes. Network 1, 2, and 3 corresponds to these type of networks. From this subset, we use network 1 as a network to generate data.
Another subset of networks we compare are graphs with minimum edges between nodes, such type of networks have many nodes independent and found by keeping the threshold high. Examples of such networks are network 4,5,6,7. These models were easy to construct because we model simple dependencies between the nodes. Network 5 had an edge which was derived from Bayes theorem, for variability in directions of edges. Just one network(Network 6) amongst this subset had a V-Structure test if V-Structure helps in Sparse Bayesian networks. Network 7 was testing how just 1 independent node affects the score of a Bayesian network. We use network 5 to be a representative of this subset.
The third type of networks makes a lot of dependence assumptions, more than the above two types of networks. These nodes have no independent random variables and the non-V-Structure edges are low. Networks 8,9, and 10 correspond to this subset. Network 8 is used in this case to generate data for ancestral sampling. The networks can be found in Figure 1.

4.2 Models in 'and' Study

The Basic Bayesian network is obtained by the Hill Climbing Algorithm with default parameters. Then we vary the max_indegree to 1. Since 2 is the max in the degree we get two such Bayesian networks. These networks can be found in 3 and 2.
For the implementation of MMHC algorithm, we first created a skeleton of edges with most of the edges disconnected keeping the significance level fixed at 0.0001 and gave this as a start parameter

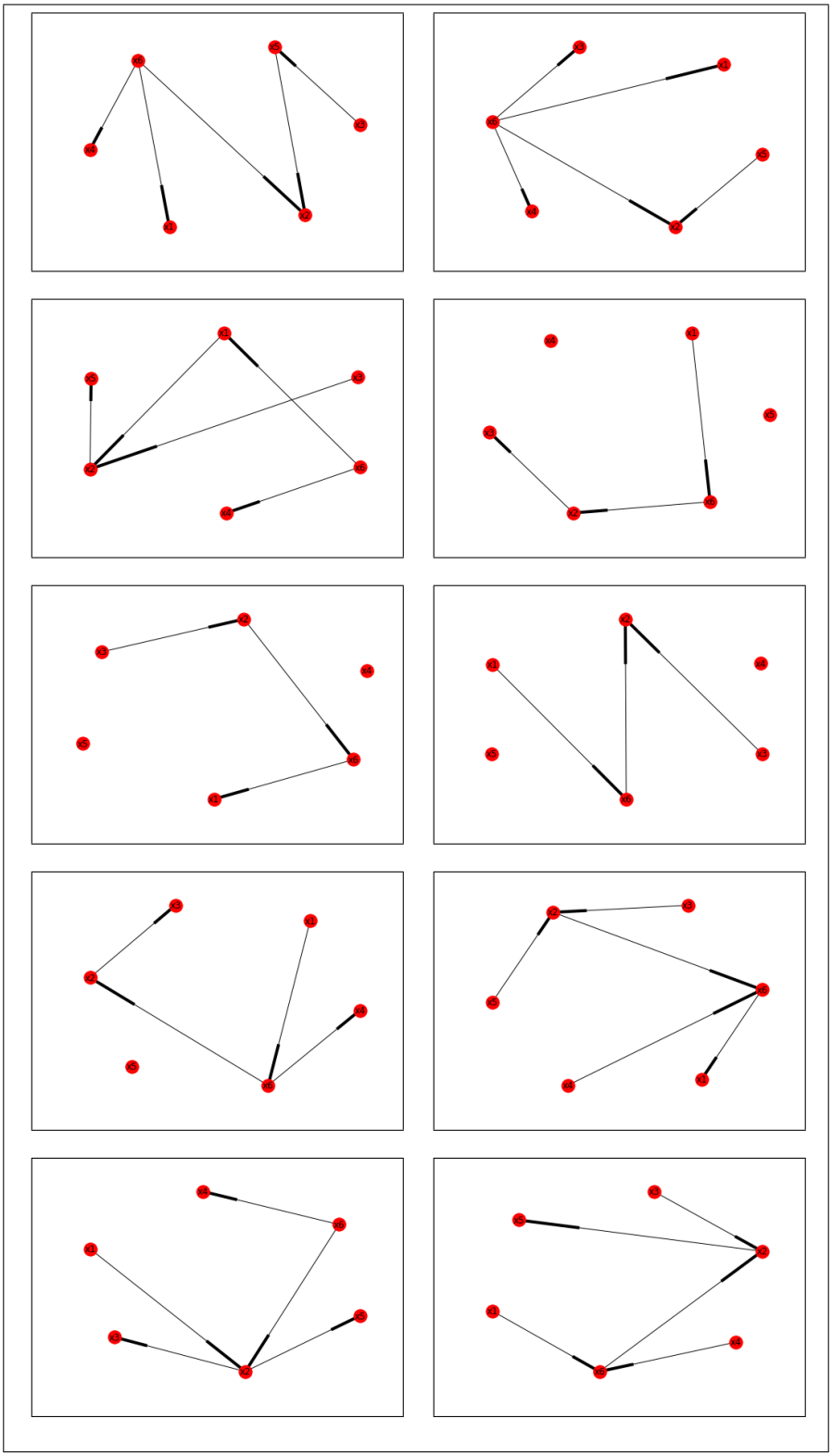

Figure 1: Left to right, top to bottom networks 1 through 10

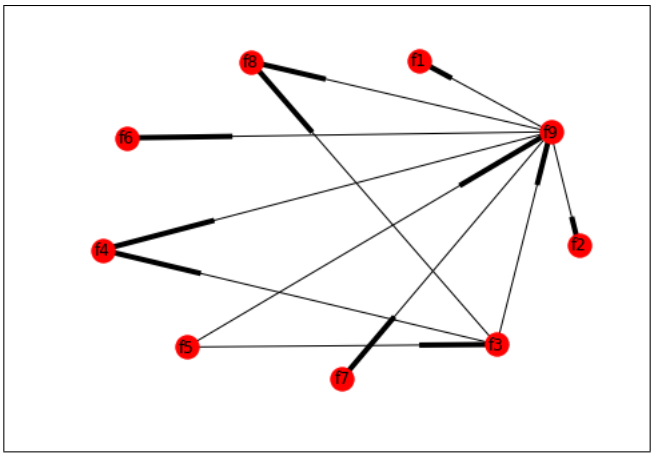

Figure 2: Bayesian Network for 'and' data set

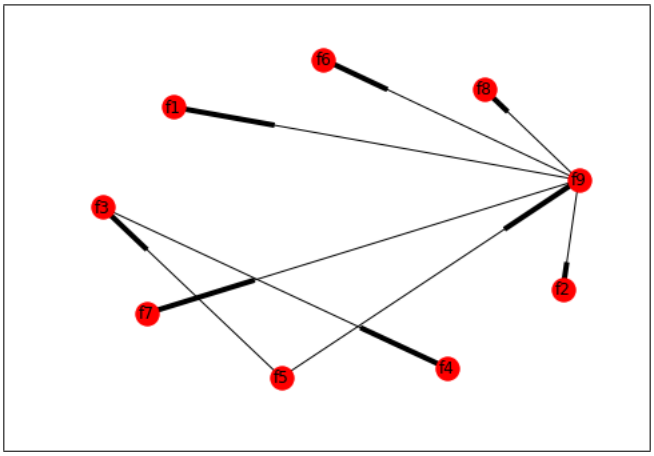

Figure 3: Bayesian Network for 'and' data set (indegree 1)

to the hill climbing algorithm. The hill climbing algorithm then orients the edges and builds a structure over the skeleton edges. Referring to the 4, we see that in skeleton the direction is x4 -> x3, but during Hill Climbing the algorithm orients it as x3->x4.

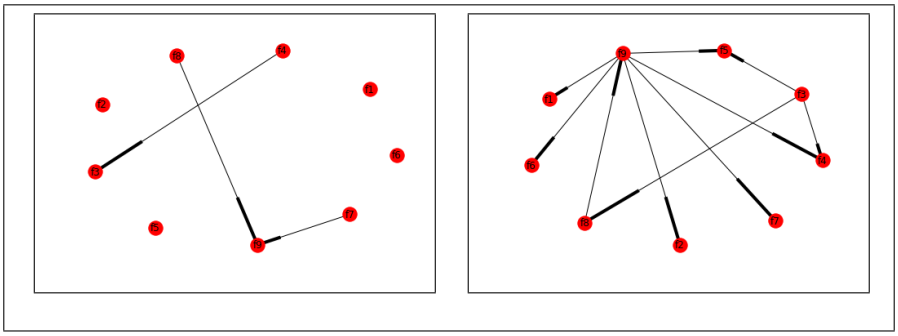

Figure 4: Left: Skeleton, Right:Final Network for MMHC Bayesian Structure Learning in 'and' Study

Table 2: Models in 'and' Study

Type of Network	K2Score
Hill Climbing Indegree 2	-9462.7049
Hill Climbing Indegree 1	-9472.9720
MMHC	-9464.6952

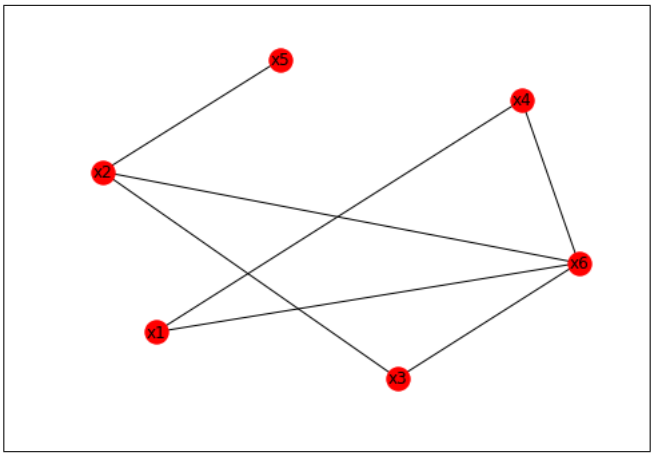

Figure 5: Markov Network for 'th' data set

4.3 Bayesian to Markov Conversion

We use the default package of PGMPY library to convert the Bayesian network to Markov network by Moralization discussed in Section 3. We use network 10 as the best network for 'th' study and 'and' data to covert them into Markov Network, See Figures 5 and 6.
We also run inference on these networks to find the differences if any. The results are discussed in 5.

5 Results

Using Table 1 We find that network 10 is the best network in the 'th' data set with the highest score. Among the networks which came close was Network 7 and 8. However, K2Score of network 8 could

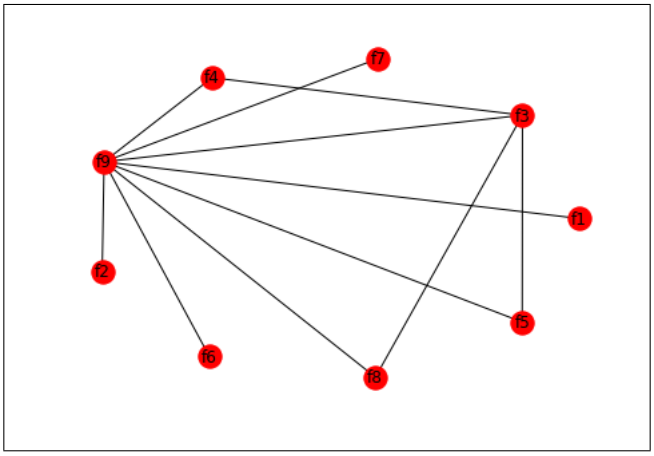

Figure 6: Markov Network for 'and' data set

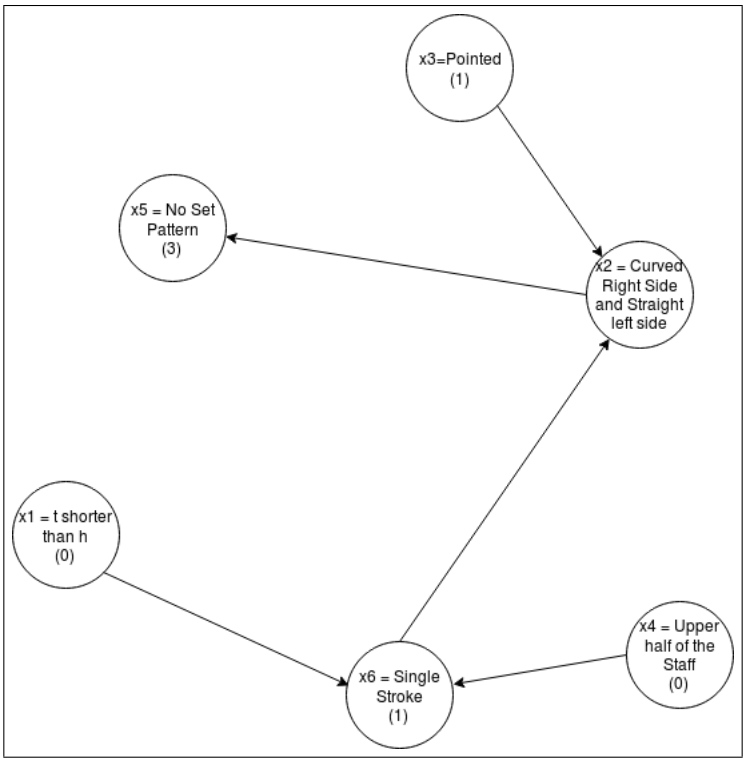

Figure 7: Visual High Probablity 'th'

be biased given that we have used it to generate data by ancestral sampling.
For 'and' Study the results are given in 2. We see that Hill Climbing with indegree 2 outperforms Hill Climbing with indegree 1. MMHC is closer to this network.
In the best Bayesian Network for 'th' we see that 'Shape of Arch of h' and 'Shape of t' affects 'Shape of the loop of h'. Moreover, the 'Shape of t' is affected by both 'Height of Cross on t Staff' and 'Height relationship of t and h' which seems an interesting area to look for discussions.
For the Bayesian Network for 'and' data set is dense and there are a lot of ways for inference to travel from one node to other. Such as one possibility for inference endpoint f4 would be from f3 or f5->f3->f9->f4. This network is more densely connected than 'th' data set which captures the causal flow of inference such as for V-Structure differently. Example, for f5 -> f9 <- f3, if x9 is observed then f5 and f3 won't be independent, we will still need to observe f5 or f3 to make either of them independent since they are part of a dense network.
So far the inferences of the Most probable and least probable assignment are same before and after moralization. But will change for evidential reasoning inferences due to Moralization(Discussed in 3). Based on the different inferences necessary whether to use a Bayesian or Markov network can be evaluated on a case-by-case basis.
We follow up on the results and inferences of this project in Section 6.
The Summarized high probability 'th' can be seen in figure 7

6 Discussion

Regarding inherent bias in the evaluation method for 'th' networks, we find that if we consider model 1,4,6 we see that Model 1 is better than all other models on an average. But if we consider models 4,5,6(Minimal connection models) we find that network 5 is better. It seems that the 'Independence assumptions from the correlation threshold affects the evaluation model in the form of Bias'. Furthermore, we find that heavily connected networks such as network 1, network 2, and network 3 have more effect on the bias(Possibly due to more connections and fewer independence assumptions) than networks 4,5,6.
The inferences generated by Bayesian and Markov networks for both most probable and least

Table 3: Inference on 'th' data set for Bayesian and Markov networks

NW	Inference	Height Relationship of t and h (x1)	Shape of loop of h (x2)	Shape of Arch of h (x3)	Height of Cross on t staff (x4)	Baseline of h (x5)	Shape of t (x6)
BN	Most Probable	t shorter than h (0)	Curved right side and straight left side (1)	Pointed (1)	Upper half of staff (0)	No Set Pattern (3)	Single Stroke (1)
	Least Probable	t even with h (1)	Curved left side and straight right side (2)	No set pattern (2)	Above staff (2)	baseline even (2)	Tented (0)
MN	Most Probable	t shorter than h (0)	Curved right side and straight left side (1)	Pointed (1)	Upper half of staff (0)	No Set Pattern (3)	Single Stroke (1)
	Least Probable	t even with h (1)	Curved left side and straight right side (2)	No set pattern (2)	Above staff (2)	baseline even (2)	Tented (0)

Table 4: Inference on 'and' data set for Bayesian and Markov networks

NW	Inference	Initial stroke of formation of a (x1)	Formation of staff of a (x2)	Number of arches of n (x3)	Shape of arches of n (x4)	Location of mid-point of n (x5)	Formation of staff of d (x6)	Formation of initial stroke of d (x7)	Formation of terminal stroke of d (x8)	Symbol in place of the word and (x9)*
BN	Most Probable	Center of staff (2)	Retraced (1)	Two (1)	Pointed (0)	At Baseline (2)	Looped (2)	Overhand (0)	Curved down (2)	Symbol (1)
	Least Probable	Right of staff (0)	No Staff (3)	No fixed pattern (2)	No fixed pattern (4)	No fixed pattern (3)	Tented (0)	Straight across (2)	No fixed pattern (4)	Formation (0)
MN	Most Probable	Center of staff (2)	Retraced (1)	Two (1)	Pointed (0)	At Baseline (2)	Looped (2)	Overhand (0)	Curved down (2)	Symbol (1)
	Least Probable	Right of staff (0)	No Staff (3)	No fixed pattern (2)	No fixed pattern (4)	No fixed pattern (3)	Tented (0)	Straight across (2)	No fixed pattern (4)	Formation (0)

probable pattern of 'th' and 'and' are the same. These are summarized in table 3 and 4. An interesting observation which we find is that the tented shape of t is least probable for 'th' data set, and similarly, the tented formation of staff of d is least probable for 'and' data set. Does this mean there is a possibility of noisy distant supervision for data set for 'th' task?
Another interesting inference obtained was least probable and most probable assignment is completely inverse for 'th' for Shape of Loop of h. Furthermore, we see that Shape and arch (x2 and x3) of n are highly correlated.
It can also be seen that very few individuals keep the baseline of h even, which is the convention usually in cursive books which 'th' study is using.
For the 'and' Bayesian network, an intuitive structure emerges without explicitly stating it from the data, out of 9 nodes, node f9(Symbol in place of the word and) affects 6 of them. This is obvious since if there is an '' symbol, it would affect most of the features. But, in the data set, this is modelled differently, they have labelled the 'Word Formation' and 'Symbol Formation' on the same states across other variables. Moreover, while investigating we found that MAP query is different from manual approaches such as finding mode, if we find the mode of the feature 9, we would get None as the most probable assignment, but if we use MAP query we get Symbol as the Most probable assignment. This is because of priors or the parents of feature 9. Another disparity we are not able to explain is that is in spite of 'Symbol' being part of the most probable assignment, the No-fixed pattern is found in four features out of 9 as least probable assignment. It shows that the 'symbol and' and 'word formation and' are mapped to the same variable states.

7 Conclusion

In this work, we have completed all the tasks and explored PGMs in a concise way. We have discussed important inferences obtained from the two past works. We have also extended the pgmpy library in the form of adding a V-Structure function to calculate the CPDs of a V Structure from a CPD list dictionary and other parameters. In the latter part of the paper we have also converted Bayesian networks to Markov Models using Moralization. Based on the work, it seems that it is easy predicting PGMs from data but it is difficult to get PGMs from CPDs alone. We have also discussed a possibility to handle cases where we only have the CPDs, Distant Supervision. We, however, do not follow this path and future work could dive into the possibility of using 'and' data set as distant supervision for 'th' paper.

References

[1] Probabilistic Graphical Models Principles and Techniques (2009) Daphne Koller and Nir Friedman, The MIT Press. Retrieved March 08, 2019, from https://mitpress.mit.edu/contributors/daphne-koller

[2] The max-min hill-climbing Bayesian network structure learning algorithm (2006), Tsamardinos, I., Brown, L.E. Aliferis, C.F. Mach Learn (2006) 65: 31. https://doi.org/10.1007/s10994-006-6889-7

[3] A Statistical Examination of Selected Handwriting Characteristics, Muehlberzer et al. (Link in UBLearns, other citation link needed). Retrieved March 08,2019