Prediction of Intern recruitment Using Bayesian Network

by

Jiexin Chen (2130026008)

Jiayin Peng (2130031205)

Ruotong Yu (2130026185)

A group project (Bayesian Networks AI3043)

Bachelor of Science (Honours)

in

Artificial Intelligence

at

BNU-HKBU

UNITED INTERNATIONAL COLLEGE

December, 2024

DECLARATION

We hereby declare that all the work done in this Project is of our independent effort. We also certify that we have never submitted the idea and product of this Project for academic or employment credits.

The contribution of each teammate is listed below:

Jiexin Chen is responsible for inference

Jiayin Peng is responsible for UI

Ruotong Yu is responsible for data processing and Bayesian Network Construction

Date: ____2024.12. 15_____

1 Introduction

For a tech company, hiring a candidate as an intern is a serious question. HR needs to consider from multiple dimensions whether the candidate is suitable to stay in the company, so as to facilitate the continuous development of the company.

Our work is to give the probability that a HR in a company decides to give admission offer to a candidate. We learned the conditional probability distribution based on a series of datasets that include comprehensive collection of synthetic job postings. We use statistical methods to calculate the probability of a company hiring an intern under different conditions. We also constructed the logic of our Bayesian Network based on common sense and some assumptions.

Inside our Network, the final goal is to calculate the probability that a candidate will enter a tech company to be an intern. This will be affected by both candidates' yearn for the position and company's willingness to the candidate, and also, some other factors of emergency. We construct some related factors, such as work experience, interview performance, major-related work, welfare, etc., into logical networks following some reasonable logic. We define the conditional distributions and conditional dependencies according to both datasets and assumptions.

We constructed the User Interface to input the query and conditions and then output the probability. The user can also define the Baye's probability by themselves to see how the results change. This research aims to give some reference to people who are seeking for jobs.

2 Bayesian Network Construction

This section illustrates how we calculate the probability distribution of related factors from some hiring datasets by statistic method and demonstrates our Bayesian Network by a directed acyclic graph.

2.1 Data Collection

We found a job description dataset from Kaggle which contains more than 1,000 thousand job positions information and counted some factors related to the jobs to get real probability distributions.

Overall, there are 12 factors: **Educational Level, Work Experience, Age, Interview Performance, Major Related, Working Hours, Interest, Salary, Welfare, Company Offer, Offer Accepted**. All of them contribute to the final **Admission**.

From the dataset, we can directly using '**=SUBTOTAL(3, X, X)'** function to count the rate of specific values. For example, the **Educational Level** ('Qualification' in this table) node contains three main categories, which are Bachelor Degree, Master Degree and PhD, separately occupies 524461, 419037 and 105080 items, so that the rate will be 0.5,0.4 and 0.1. We count all the information by the same method for **Work Experience, Working Hours, Salary, Welfare**.

Job Id	Experience	Qualific	Salary Range	Work Type	Preference	Job Title	Role	Benefits	1048576
1.09E+15	5 to 15 Years	M.Tech	$59K-$99K	Intern	Female	Digital Marke	Social Media M	{'Flexible Spending Accounts (FSAs), Relocation Assistance, Legal Assistance, Employee Recognition Programs, Financial Counselin	
3.98E+14	2 to 12 Years	BCA	$56K-$116K	Intern	Female	Web Develop	Frontend Web	{'Health Insurance, Retirement Plans, Paid Time Off (PTO), Flexible Work Arrangements, Employee Assistance Programs (EAP)'}	
4.82E+14	0 to 12 Years	PhD	$61K-$104K	Temporary	Male	Operations M	Quality Control	{'Legal Assistance, Bonuses and Incentive Programs, Wellness Programs, Employee Discounts, Retirement Plans'}	
6.88E+14	4 to 11 Years	PhD	$65K-$91K	Full-Time	Female	Network Engi	Wireless Netwo	{'Transportation Benefits, Professional Development, Bonuses and Incentive Programs, Profit-Sharing, Employee Discounts'}	
1.17E+14	1 to 12 Years	MBA	$64K-$87K	Intern	Female	Event Manag	Conference Ma	{'Flexible Spending Accounts (FSAs), Relocation Assistance, Legal Assistance, Employee Recognition Programs, Financial Counselin	
1.17E+14	4 to 12 Years	MCA	$59K-$93K	Full-Time	Male	Software Test	Quality Assurar	{'Life and Disability Insurance, Stock Options or Equity Grants, Employee Recognition Programs, Health Insurance, Social and Recre	
1.29E+15	3 to 15 Years	PhD	$63K-$103K	Temporary	Both	Teacher	Classroom Tea	{'Flexible Spending Accounts (FSAs), Relocation Assistance, Legal Assistance, Employee Recognition Programs, Financial Counselin	
1.50E+15	2 to 8 Years	M.Com	$65K-$102K	Contract	Female	UX/UI Design	User Interface I	{'Employee Assistance Programs (EAP), Tuition Reimbursement, Profit-Sharing, Transportation Benefits, Parental Leave'}	
1.68E+15	2 to 9 Years	BBA	$65K-$102K	Temporary	Female	UX/UI Design	Interaction Des	{'Transportation Benefits, Professional Development, Bonuses and Incentive Programs, Profit-Sharing, Employee Discounts'}	
2.56E+14	1 to 10 Years	BBA	$60K-$80K	Full-Time	Both	Wedding Plar	Wedding Consi	{'Legal Assistance, Bonuses and Incentive Programs, Wellness Programs, Employee Discounts, Retirement Plans'}	
2.70E+15	3 to 10 Years	BCA	$57K-$104K	Contract	Female	QA Analyst	Performance Te	{'Flexible Spending Accounts (FSAs), Relocation Assistance, Legal Assistance, Employee Recognition Programs, Financial Counselin	
1.45E+15	4 to 12 Years	B.Tech	$64K-$98K	Contract	Male	Litigation Att	Family Law Atto	{'Employee Referral Programs, Financial Counseling, Health and Wellness Facilities, Casual Dress Code, Flexible Spending Accounts	
1.91E+15	3 to 15 Years	MCA	$65K-$122K	Part-Time	Both	Mechanical E	Mechanical Des	{'Tuition Reimbursement, Stock Options or Equity Grants, Parental Leave, Wellness Programs, Childcare Assistance'}	
2.91E+14	1 to 8 Years	B.Com	$56K-$86K	Temporary	Female	Network Adm	Network Securi	{'Legal Assistance, Bonuses and Incentive Programs, Wellness Programs, Employee Discounts, Retirement Plans'}	
1.63E+15	1 to 9 Years	MCA	$57K-$98K	Full-Time	Male	Account Man	Sales Account I	{'Casual Dress Code, Social and Recreational Activities, Employee Referral Programs, Health and Wellness Facilities, Life and Disabil	
2.69E+15	4 to 12 Years	M.Com	$65K-$100K	Part-Time	Male	Brand Manag	Product Brand	{'Health Insurance, Retirement Plans, Flexible Work Arrangements, Employee Assistance Programs (EAP), Bonuses and Incentive Pr	
2.82E+15	5 to 14 Years	M.Tech	$60K-$83K	Part-Time	Male	Social Worke	School Social W	{'Health Insurance, Retirement Plans, Flexible Work Arrangements, Employee Assistance Programs (EAP), Bonuses and Incentive Pr	
7.69E+13	0 to 11 Years	BA	$55K-$117K	Full-Time	Male	Social Media	Content Creato	{'Casual Dress Code, Social and Recreational Activities, Employee Referral Programs, Health and Wellness Facilities, Life and Disabil	
2.35E+14	3 to 12 Years	BA	$55K-$121K	Contract	Female	Email Market	Deliverability A	{'Employee Referral Programs, Financial Counseling, Health and Wellness Facilities, Casual Dress Code, Flexible Spending Accounts	
4.08E+14	5 to 9 Years	PhD	$65K-$128K	Temporary	Female	HR Generalist	HR Coordinato	{'Health Insurance, Retirement Plans, Flexible Work Arrangements, Employee Assistance Programs (EAP), Bonuses and Incentive Pr	
9.41E+14	0 to 15 Years	B.Tech	$62K-$120K	Full-Time	Both	Legal Assistar	Legal Secretary	{'Employee Assistance Programs (EAP), Tuition Reimbursement, Profit-Sharing, Transportation Benefits, Parental Leave'}	

Figure 1: Job Position Dataset

There are also some data that does not exist in the dataset, we search them on the Internet. For example, **Major Related** node values are got from '2024 College students' Employability research report' as shown in Figure 2.

66%毕业生签约工作与专业对口，对口率提高3个百分点

在签约的毕业生中，43%认为签约工作与所学专业比较对口，23%认为很对口，共有66%的应届毕业生对签约工作与专业的匹配度表示满意，较去年高3个百分点。

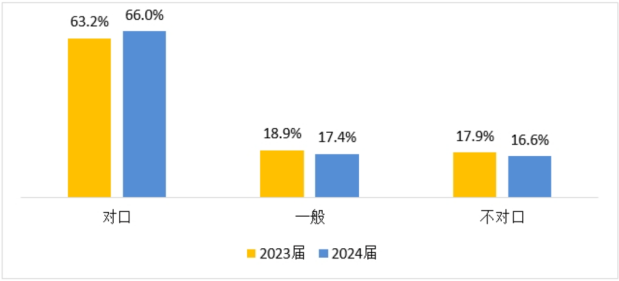


图 19　签约毕业生认为工作与专业是否对口

Figure 2: Major-Related Data

2.2 Bayesian Network Visualization

Some of the factors, include **Educational Level, Work Experience, Age, Interview Performance** will decide whether the company would like to give an offer to the candidate, i.e., **Company Offer.** Others include **Major Related, Working Hours, Interest, Salary, Welfare** will affect the rate of the candidate accept the offer, i.e., **Offer Accepted.**

Specifically, at the beginning, we place **Educational Level (3 values: PhD, M, B)** and **Work Experience** (2 values: 2 years and more, less than 2 years). These two factors will together determine **Interview Performance (3 values: Excellent, Good, Bad). Age** and **Interview Performance** jointly decide **Company Offer(2 values: Yes, No).** In another path, it begins with **Salary (3 values: 55-57 thousand, 58-61 thousand, 62-65 thousand), Major Related (2 values: Yes, No)** and **Working Duration (2 values: Full time, Part Time),** Major Related and Working Duration together determine **Interest(2 values: Yes, No)** of the job. **Welfare(2 values: good, general)** depends on Salary. Interest and Welfare together decide **Offer Accepted(2 values: Yes, No).**

Based on the conditional dependencies and conditional distributions we got in 2.1, we constructed our Bayesian network as figure 1.

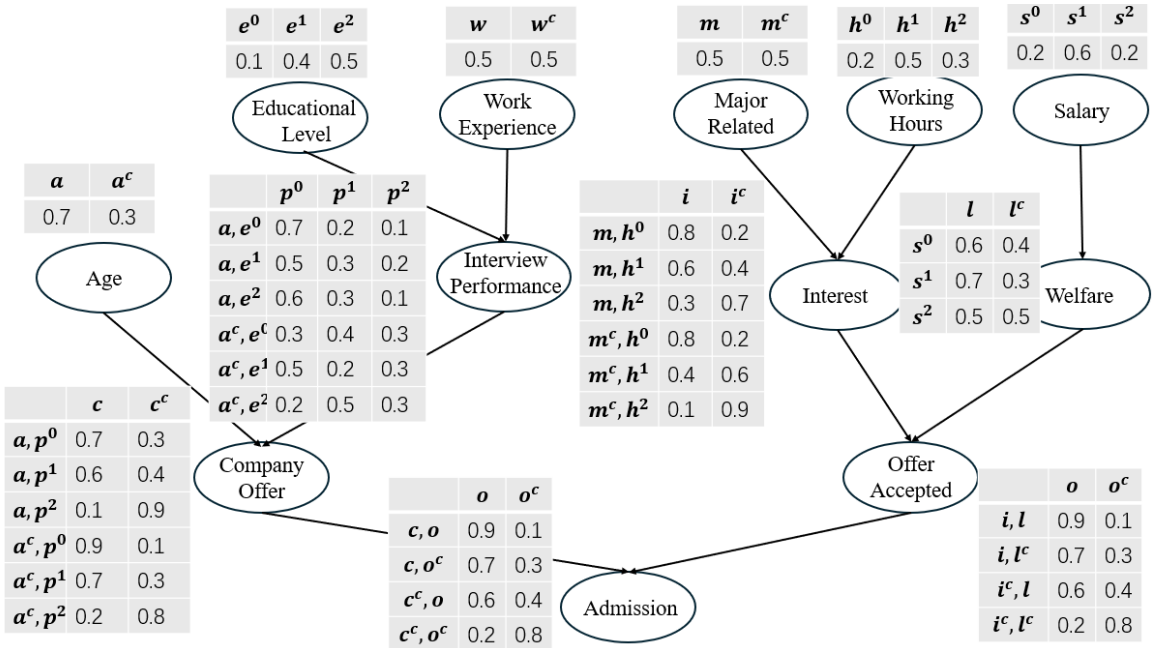


Figure 2: Intern Admission Network

3 Inference algorithms

In this project, we use two methods to inference: Variable elimination and clique tree algorithm.

3.1 Variable elimination algorithm (VE)

Variable Elimination (VE) is an efficient inference algorithm widely used for probability computations in Bayesian networks. Its key idea is to simplify the calculation of the posterior distribution of the target variable by marginalizing out irrelevant variables..

3.1.1 Key Steps in VE

Step 1: Select a Non-Evidence Variable

In VE, we choose a variable that is not part of the evidence or the query to eliminate first. For instance, if the target variable is Admission (a), and the evidence includes Educational Level (e) = Master's (e=1) and Work Experience (w) = Less than 2 years (w=0), we might select Age (age) as the first variable to eliminate.

Step 2: Combine Factors Involving the Selected Variable

To eliminate a variable (e.g., age), we find all the factors that include this variable. For age, the relevant factors are:

1. $\phi(a) = P(age)$ which represents the prior probability of age.
2. $\phi(age, p) = P(p|age, e)$, which represents how interview performance depends on age and educational level.
3. $\phi(age, c) = P(c|age, p)$, which represents the probability of receiving a company offer based on age and interview performance.

We combine these factors into a single new factor:

$$\varphi(age, p, c) = \phi(age) \cdot \phi(age, p) \cdot \phi(age, c)$$

Step 3: Marginalize the Selected Variable

Next, we sum out age from the combined factor $\varphi(age, p, c)$to create a new factor $\tau(p, c)$

$$\tau(p, c) = \sum_{age} \varphi(age, p, c)$$

Step 4: Repeat for All Non-Evidence Variables

Continue eliminating all non-evidence and non-query variables in a similar manner.

Step 5: Compute the Target Variable's Probability

After eliminating all irrelevant variables, the remaining factors are combined to compute the posterior probability of the target variable aa (admission).

3.1.2 Implementation of VE

3.1.2.1. Using the pgmpy Library

The pgmpy library provides a ready-to-use implementation of VE, significantly simplifying the inference process.

1. Model Construction: Define the network structure and CPDs..
2. Query Execution: Use query() function to specify the target variable and evidence.
3. Automated Computation: The library internally handles factor combination, variable elimination, and marginalization, returning the posterior distribution of the target variable.

3.1.2.2. Custom Implementation of VE

To gain a deeper understanding of the VE algorithm, a custom implementation was developed, following these steps:

1. Factor Initialization: Convert all CPDs in the Bayesian network into factor representations.
2. Recursive Variable Elimination: Identify and combine relevant factors, then marginalize the selected variable by summing over its values.
3. Posterior Computation: Multiply remaining factors and normalize based on evidence.
4. Result Output: Return posterior probabilities and highlight the most likely state.

3.2 Clique Tree algorithm

The Junction Tree Algorithm is an exact inference method applicable to both directed and undirected probabilistic graphical models. It leverages a specialized tree structure, known as the junction tree, to perform efficient variable elimination and probabilistic inference.

3.2.1 Key Steps in Clique Tree algorithm

Step 1: Construct the Junction Tree

To construct a junction tree, the graphical model is converted into a triangulated graph to simplify dependencies. It is then decomposed into fully connected subgraphs called cliques, organized into a tree structure. The tree satisfies the running intersection property, ensuring

variables shared across cliques remain consistent throughout the tree.

The clique tree based on our data and model is shown below.

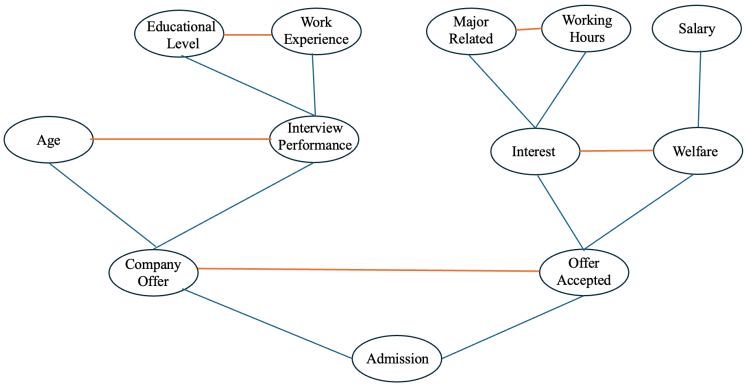


Figure. Moralization and triangulation graph.

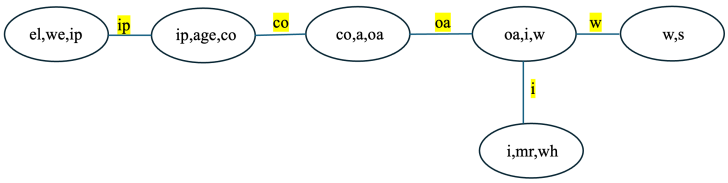
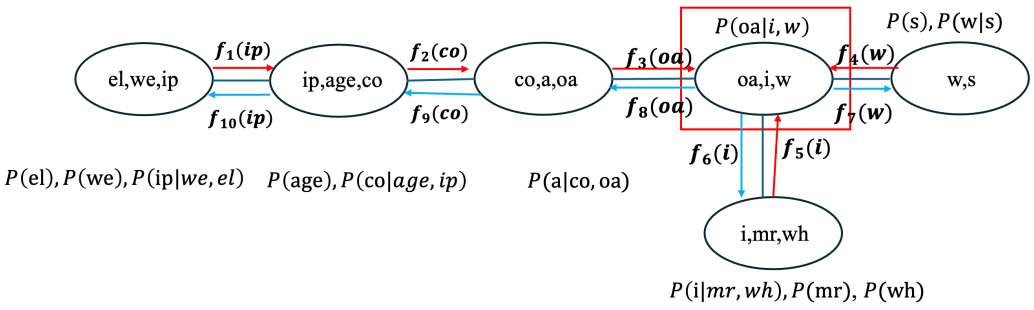


Figure. Clique tree.

Step 2. Perform Message propagation on the Junction Tree

Message passing on the junction tree assigns potential functions to cliques, derived from the original factors. Messages are exchanged between neighboring cliques to update marginal potentials, combining incoming and local information. This ensures consistent probabilities across the tree for efficient inference.

The following is the case when [oa,i,w] is selected as a pivot.



1. massage passing

$$f_1(ip) = P(el) \cdot P(we) \cdot P(ip|we,el)$$

$$f_2(co) = P(age) \cdot P(co|age, ip) \cdot f_1(ip)$$

$$f_3(oa) = P(a|co, oa)\, f_2(co)$$

$$f_4(w) = P(s) \cdot P(w|s)$$

$$f_5(i) = P(i|mr, wh) \cdot P(mr) \cdot P(wh)$$

2. massage distribution

$$f_6(i) = P(oa|i, w) \cdot f_3(oa) \cdot f_4(w)$$

$$f_7(w) = P(oa|i, w) \cdot f_3(oa) \cdot f_5(i)$$

$$f_8(oa) = P(oa|i, w) \cdot f_5(i) \cdot f_4(w)$$

$$f_9(co) = P(a|co, oa) \cdot f_8(oa)$$

$$f_{10}(ip) = P(age), P(co|age, ip) \cdot f_9(co)$$

Step 3. Construct potential functions and Compute Probabilities

For clique [el, we, ip]: $h_{[el,we,ip]} = P(el), P(we), P(ip|we, el) \cdot f_{10}(ip)$

For clique [ip, age, co]: $h_{[ip,age,co]} = P(age), P(co|age, ip) \cdot f_1(ip)$

For clique [co, a, oa]: $h_{[co,a,oa]} = P(a|co, oa) \cdot f_2(co)$

For clique [oa, i, w]: $h_{[oa,i,w]} = P(oa|i, w) \cdot f_3(oa) \cdot f_4(w) \cdot f_5(i)$

For clique [w, s]: $h_{[w,s]} = P(s) \cdot P(w|s) \cdot f_7(w)$

For clique [i,mr,wh]: $h_{[i,wr,wh]} = P(i|mr, wh) \cdot P(mr) \cdot P(wh) \cdot f_6(i)$

3.2.2 Implementation of Clique Tree algorithm

1. **Transforming the Graph**: Convert the original network into a tree-like structure that satisfies the running intersection property.

2. **Grouping Variables**: Group variables into cliques to enable efficient probabilistic computation.

3. **Visualization**: Visualize the Junction Tree to intuitively understand the connections between cliques.

4. **Belief Propagation**: Perform message passing through initialization, calibration, and probability updates.

4. **Query Probabilities**: Compute marginal or conditional probabilities of target variables based on observed evidence.

4 Comparisons

In this project, two inference methods were used: Variable Elimination and the Clique Tree Algorithm.

1. Variable Elimination

Execution Speed:

Implemented via the pgmpy library and a custom function, with the custom function being the fastest due to task-specific optimizations that reduce redundant calculations.

Applicable Scenarios:

Best for smaller or simpler Bayesian networks with fewer dependencies, where conditional probabilities can be computed quickly through variable elimination.

2. Clique Tree Algorithm

Execution Speed:

Slower than Variable Elimination as it involves transforming the network into a clique tree, adding computational and memory overhead.

Applicable Scenarios:

Suitable for large, complex networks with many interdependencies, as it efficiently manages highly coupled relationships but requires more memory and computational resources.

5 Result

5.1 Predictive inference

1. Educational Background

Result: Admission probability is 0.6149 for Bachelor's and 0.6416 for Master's..

Recommendation: Pursue a Master's degree to improve admission chances.

2. Age Over 30

result: Admission starts at 0.59 but increases to 0.6786 with work experience and 8-hour workdays.

Recommendation: While age over 30 may lower chances, relevant experience and dedication can compensate.

5.2 Diagnostic inference

1. Company Welfare

Result: Good welfare improves admission probability (0.5914 vs. 0.4086).

Recommendation: company sis suggested to enhance employee benefits

2. Working Hours

Result: Admission is highest (0.4962) for 5–8 work hours, lower for <5 (0.1831) or >8 hours (0.3208).

Recommendation: company should set reasonable expectations for daily working hours, maintain employees to work consistently within the 5–8 hour range.

3. Field of Study

Result: Science fields have slightly higher admission probability (0.5103 vs. 0.4897 for Arts).

Recommendation: Promote science fields while keeping arts students competitive.

5 conclusions

Bayesian networks provide a powerful and flexible tool for job seekers and companies in recruitment software, which not only optimizes the recruitment decision-making process and improves matching efficiency, but also provides in-depth analysis and predictions for research in the macroeconomic and sociological fields. Through Bayesian networks, recruitment software enables more intelligent decision support, while also providing a data-driven perspective on policy adjustments in the labor market and economic forecasting. With the continuous development of technology and the continuous accumulation of data, the application potential of Bayesian networks in the field of recruitment and social policy research will be further expanded, bringing more accurate and efficient solutions to society.