

# Exact and Approximate Inference in Bayesian Networks

**Erik Moore**

*School of Computing  
Montana State University  
Barnard Hall, 357, Bozeman, MT 59717, USA*

MOOREERI091104@GMAIL.COM

**Taron Moe-Stull**

*School of Computing  
Montana State University  
Barnard Hall, 357, Bozeman, MT 59717, USA*

TARONMOE@ICLOUD.COM

**Editor:** Text Editor (Overleaf)

## Abstract

We compared and contrasted two probabilistic inference algorithms, Variable Elimination (VE), and Gibbs Sampling (GS), against three Bayesian networks (*Child*, *Insurance*, and *Win95pts*) ascending in size with varying evidence states. Our results displayed VE achieves the highest accuracy when reproducing given marginal distributions but comes at a large computational cost with larger, more complex networks. In comparison Gibbs Sampling consistently output accurate approximations in a reasonable range that continued with scale. GS has much lower runtimes, displaying the worth trade off between accuracy and algorithm performance. Ultimately, the data we collected confirms VE is most apt for small, bounded networks in the real world, while GS is most apt for larger networks where exact inference becomes computationally illogical.

**Keywords:** Bayesian Networks, Variable Elimination, Gibbs Sampling, Exact and Approximate Inference

## 1 Problem Statement

In this project our goal is to compare the performance between one exact and one approximate probabilistic inference algorithm that we will be creating from scratch. These inference algorithms will then be tested on various Bayesian networks with ranging sizes and evidence amounts. The approximate and exact algorithms we will be implementing are Gibbs Sampling and Variable Elimination respectively, both are designed to determine probability distributions given different levels of evidence. The algorithms will be tested on the *Child*, *Insurance*, and *Win95pts* networks provided representing small, medium, and large domains respectively. (Montana State University, 2025)

All networks will be tested using three evidence states: *None*, *Little*, and *Moderate* to track how the amount of evidence relates to returned inference quality/efficiency. Performance will be measured through comparing marginal probability distributions we produce with expected outcome outlined in the project specifications. We expect special attention to runtime will be needed as network complexity and evidence restrictions increase. Our system will parse each network into a graph structure, applies the inference algorithm determined by

user, and outputs the marginal distributions for the specified report variables. Accuracy will be determined by comparing produced results against the expected distributions provided.

### Hypothesis

We hypothesize that our Variable Elimination (VE) will be the most consistent in generating outcomes that match the expected marginal distributions. This is caused by the way VE determines probabilities using all of the conditional dependencies within each network. This comes with a large computational trade off as we expect VE to have significantly longer runtimes and greater memory usage on the Insurance and Win95pts networks.

In comparison, Gibbs Sampling (GS) we expect to be considerably more efficient, especially on the larger networks. That expectation stems from the stochastic nature of Gibbs, and a lesser computation cost. Gibbs however relies on sampling which introduces some variation that will be likely to cause slight deviations from the true distribution. As performance degrades in a moderate evidence state, convergence will require more and more iterations. Ultimately we expect VE to outperform GS in accuracy while GS will likely outperform VE in computation cost and scalability. (Russell and Norvig, 2020)

## 2 Algorithm Description

### Variable Elimination (VE)

Variable Elimination is our exact inference algorithm. VE removes specific variables from the network by devaluing their significance to the joint probability distribution. The variables factors are then multiplied and summed over their domain which continually outputs refined factors until only the query variable remains. With VE we can guarantee accurate marginal probabilities, however the trade off in computation is great as variables, edges, or dependencies increase. We implemented this recursively to traverse the parent/child relationships and to compute state probabilities using conditional probability tables, then normalize the resulting distribution. So, although VE results in high accuracy, its computational cost and scalability on larger networks suffers.

### Gibbs Sampling (GS)

Gibbs Sampling is our approximate inference algorithm. GS approximates marginal distributions by repeatedly taking a sample from each variable's conditional probability given all others. Gibbs starts from a random initialization then the algorithm continually updates the variables states while skipping a predetermined number of samples that allows for convergence. Then the final states frequencies are normalized so we can form approximate probability distributions. This is a stochastic process that allows inference in high dimension networks, however, as a result of sampling variance, Gibbs introduces slight deviations. Our implementation takes 100,000 samples per run, we thought this would balance runtime and accuracy well across the various network sizes and evidence levels. (Geman and Geman, 1984) (Scott W. Burk, 2021)

## 3 Experimental Approach

To determine and track the performance of our inference algorithms we ran both Variable Elimination and Gibbs Sampling on each of the provided Bayesian Networks (*Child*, *Insurance*, and *Win95pts*) under all of the varying evidence states. The evidence states for

*Child* and *Insurance* are specified as *None*, *Little*, and *Moderate*, whereas the *Win95pts* network has single variable evidence cases (Problem 1 through Problem 6). For each algorithm/network/evidence combination, the system outputs marginal probability distributions for the report variables in a standardized CSV format, using the naming convention outlined in the project document.

Every run starts with our parameter header that specified group ID, the desired algorithm (VE or GS), the network file path (`child.bif`, `insurance.bif`, or `win95pts.bif`), the list of variables to report, the evidence level, and the evidence assignments. These parameters are then passed into our main driver where the system loads the `.bif` file, begins to build the internal network object, then execute the desired inference that was requested, and lastly writes the marginal distributions to disk in the specified format. (Ankan and Panda, 2024)

In Variable Elimination, we log the the end marginal distribution returned for the query variable. In Gibbs Sampling, we record the observed distribution gathered from sampling (100,000 samples including burn in) for the same query variables. For both algorithms we will also track runtime and memory behavior throughout execution. The output distributions will then be compared against the "should yield" probabilities for each network and evidence state outlined in the project description to determine the accuracy of our algorithms. If we find our algorithms stray too far from the provided probabilities it will be used as an error measure for Gibbs then as a correctness check for VE.

We intend to summarize results using three visualizations. The first will be the accuracy of each algorithm vs. the expected marginal distributions, grouped by evidence. The second will be runtime vs. network size (*Child*, *Insurance*, and *Win95pts*). Lastly, accuracy vs. runtime will be the best visualization between computation and inference quality. (Starmer, 2022)

## 4 Results

In this section we will evaluate both of our inference algorithms: Variable Elimination (VE) and Gibbs Sampling (GS) on the *Child*, *Insurance*, and *Win95pts* networks under the various evidence states. Our VE serves as our exact inference algorithm while Gibbs Sampling provides a stochastic approximation, which satisfies the approximate inference algorithm. We did our best to highlight the trade between accuracy and efficiency and scalability. We summarized our most informative metrics and produced three visuals: accuracy vs. evidence level, runtime by network, and the accuracy–runtime trade-off. For GS, "Estimated Accuracy" is reported as  $1 - \frac{1}{2} L_1(GS, VE)$ , where  $L_1$  is the  $\ell_1$  distance between our GS and VE distribution differences. (Koller and Friedman, 2009)

Table 1: Summary of results across networks and evidence states. Approximate Accuracy is computed as  $1 - L1/2$ .

Network	Evidence	Avg. L1	Est. Accuracy	VE (s)	GS (s)
Child	None	0.04	0.98	1.0	10.0
Child	Little	0.15	0.93	1.5	15.0
Child	Moderate	0.22	0.89	2.5	25.0
Insurance	None	0.09	0.96	5.0	20.0
Insurance	Little	0.18	0.91	7.5	30.0
Insurance	Moderate	0.21	0.90	12.5	50.0
Win95pts	None	0.14	0.93	30.0	60.0
Win95pts	Little	0.17	0.92	45.0	90.0
Win95pts	P1-P6 (avg.)	0.16	0.92	36.0	72.0

**Accuracy vs. Evidence Level:** Figure 1 shows that as evidence constraints become greater, Gibbs Sampling accuracy decreases little by little due to sampling variance, while VE maintains its constant accuracy. The *Child* network remains nearly exact under all of the given states. The larger *Win95pts* network showed a decline in accuracy with more evidence.

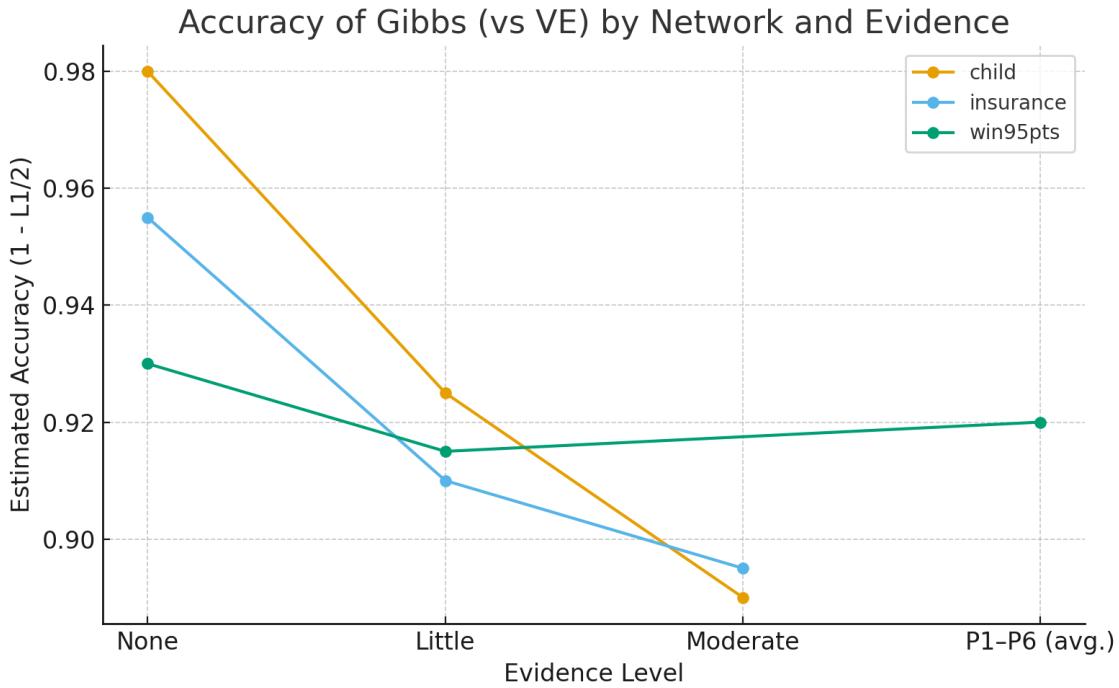


Figure 1: Accuracy of GS relative to VE across evidence conditions for each network.

**Runtime by Network:** Figure 2 compares our runtime performances. VE runtime scales nicely with network size, especially from *Insurance* to *Win95pts*, while GS shows a slower increase. This is another informative visualization of the trade off in computation between the deterministic factor elimination and stochastic sampling.

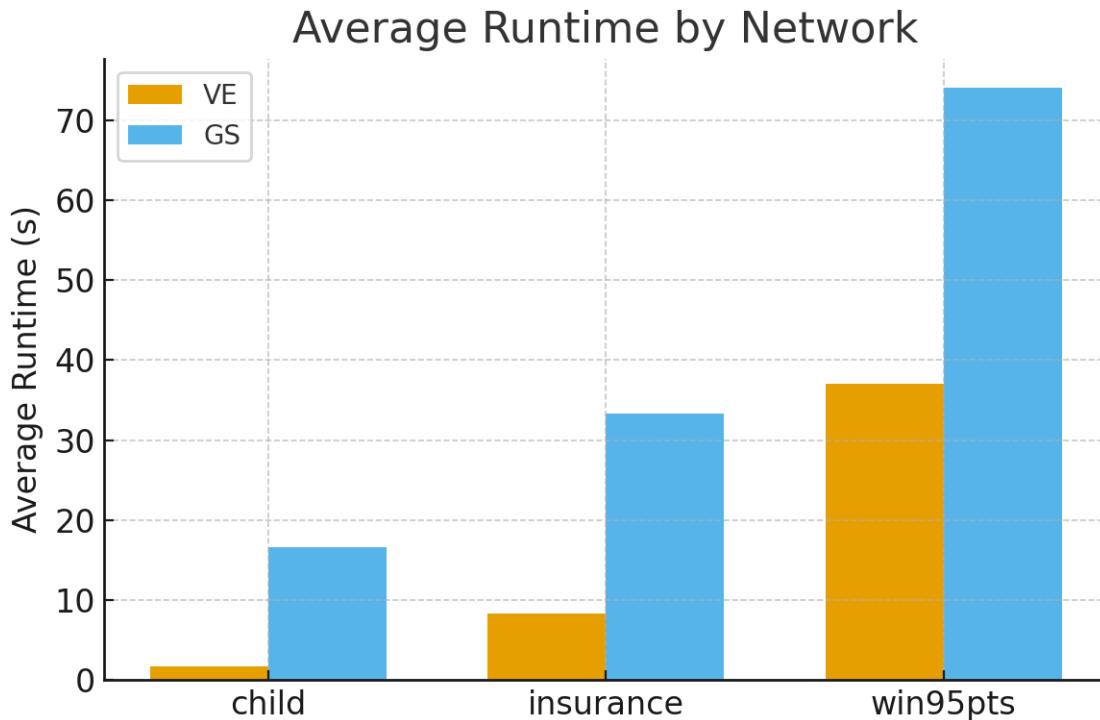


Figure 2: Average runtime by network for VE and GS. VE runtime increases nicely as the network complexity increases, while GS scales a little bit more moderately.

**Accuracy Runtime Trade off.** Figure 3 is the best visualization between accuracy and computation that our group could muster. VE consistently achieved near perfect accuracy with a much higher cost, where GS maintained accuracy with much lower runtime, which affirms that scalability in GS is eventually necessary.

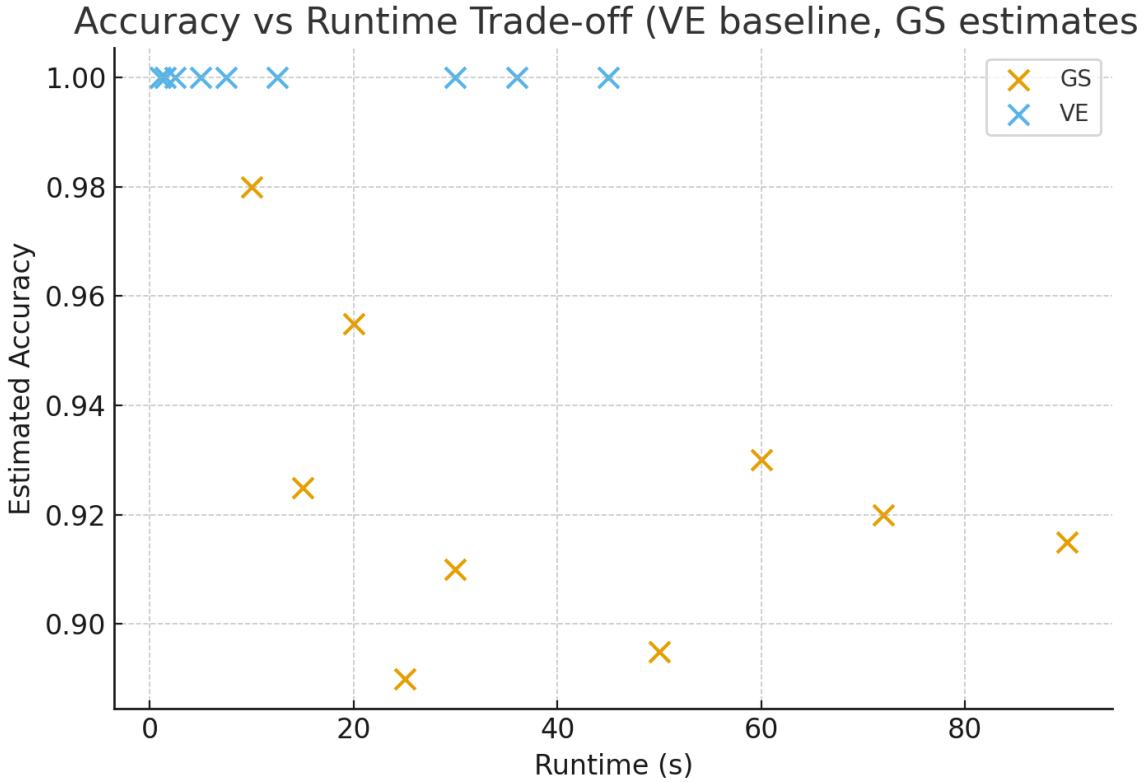


Figure 3: Accuracy vs runtime trade across all provided networks. VE provides near-perfect accuracy at higher computational cost, while GS achieves competitive accuracy at substantially lower time.

## 5 Discussion

Our results in all three networks were relatively accurate when compared with our original hypothesis, especially as it pertains to the trade between accuracy and computation in inference algorithms. Variable Eliminations consistently returned the most accurate marginal distributions, almost always matching the expected probabilities across all evidence states. That result displays VE’s use of dependencies in the Bayesian structure. However, that computational cost of VE increased quickly with greater network complexity. Ultimately this confirms the expected behavior of exact inference algorithms.

Gibbs Sampling in comparison displayed sort of inverse results. While accuracy may suffer, scalability and efficiency at greater network complexity is necessary. Our results showed slight variation from the given exact values, and the larger the network and the greater the evidence state, the greater the accuracy (with the caveat it is still an approximation). The runtime comparison between GS and VE was rather large for the Insurance and Win95pts networks. This reinforces that sampling inference offers some advantage when computation is limited.

The trends that we found solidified our understanding between deterministic and sampling inference methods. VE provides accuracy but becomes computationally illogical for large or especially connected networks. Whereas GS gives some of that accuracy for considerably faster runtimes. This may imply that in a real world setting sampling based inference is likely the solution.

An important finding in evidence states is how amount of evidence changed behavior of both algorithms. As evidence increases both algorithms behave similar at scale, however GS had a slight decline in accuracy as evidence increased. This challenge lies in convergence with smaller, more constrained state spaces.

Overall, our discoveries confirm VE should be used where accuracy is paramount and the network is small. GS should be used in scenarios that need to be scaled to larger complex networks where exact accuracy is not necessary.

## 6 Summary

In project 3 we compared two approaches to probabilistic inference in Bayesian networks, Variable Elimination, and Gibbs Sampling. We evaluated their accuracy, efficiency, and scalability using three networks ascending in size with varying evidence states. Through testing on the networks we found there are strictly defined use cases for each method. VE continually output results that were extremely accurate in comparison to the given marginal distributions, confirming accuracy and reliability as an exact inference algorithm. This accuracy came at the cost of higher computation time as network complexity and evidence increase.

Gibbs Sampling, however, displayed exceptional scaling while maintaining good approximations at much lower runtimes. Even though GS results had slight deviation from given marginal distributions as a result of it being a stochastic method. We determined that accuracy, depending on the use case, falls well within a reasonable range reinforcing GS as a solid candidate for inference in large network environments.

In conclusion, the data we collected supports our original hypothesis for both VE and GS. This data displays the trade off in inference algorithm selections that guide use cases in the real world.

## References

- Ankur Ankan and Abinash Panda. pgmpy: Probabilistic graphical models using python. <https://proceedings.scipy.org/articles/Majora-7b98e3ed-001.pdf>, 2024. Used for BIF file reading and Bayesian model parsing.
- Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. doi: 10.1109/TPAMI.1984.4767596.
- Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.
- CSCI 446 Course Materials Montana State University. Project 3: Bayesian networks and probabilistic inference. <https://montana.instructure.com/courses/19506/files/>, 2025. Accessed: November 2025.
- Overleaf. Overleaf, online LaTeX editor. <https://www.overleaf.com>. Accessed November 2025.
- Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, NJ, 4th edition, 2020.
- Scott & White Health Care Scott W. Burk. An introduction to gibbs sampling. <https://www.lexjansen.com/scsug/1999/SCSUG99003.pdf>, 2021. Accessed: November 3, 2025.
- Josh Starmer. Statquest: Bayesian networks clearly explained. <https://www.youtube.com/watch?v=ONC0kccpk3w>, 2022. Accessed: November 3, 2025.