Database Security Protection via Inference Detection[1]

Yu Chen and Wesley W. Chu

Computer Science Department
University of California, Los Angeles, CA 90095
Email: {chenyu,wwc}@cs.ucla.edu

Abstract. Access control mechanisms are commonly used to provide control over who may access sensitive information. However, malicious users can exploit the correlation among the data and infer sensitive information from a series of seemingly innocuous data access. In this paper, we proposed a detection system that utilizes both the user's current query and past query log to determine if the current query answer can infer sensitive information. A semantic inference model (SIM) is constructed based on the data dependency, database schema and semantic relationship among data. After the SIM is instantiated via specific instances, it can then be mapped to a Bayesian network and used for evaluating the inference probability. The decision of answering the current query is based on if any of the sensitive attributes can be inferred with a probability higher than their pre-specified thresholds. This detection system is being extended to the cases of multiple collaborative users based on the query history of all the users and their collaborative levels for specific sensitive information.

1. Introduction

Access control mechanisms are commonly used to protect users from the divulgence of sensitive information in data sources. However, such techniques are insufficient because malicious users may access a series of innocuous data, and from the received answers, the malicious users may employ inference techniques to derive sensitive information.

Database inferences have previously been studied. Delugach and Hinke [DH96, HDW96] and Garvey *et al*. [GLQ92] developed approaches that use schema level knowledge for inference detection at database design time. However, Yip *et al*. has pointed out the inadequacy of schema level inference detection, and he identifies six types of inference rules from data level [YL98]. An inference controller prototype was developed to handle inferences during query processing. Rule-based inference strategies were applied in this prototype to protect the security [TFC93]. Furthermore, to provide scalable inference in large systems, feasible inference channels that are based on query and database schema are generated to guide the data inference [CCH94].

In this paper, we propose to develop an inference detection system that resides at the central directory site. The system keeps track of users' query history and when a new query is posed, all the channels where sensitive information can be inferred will be identified. If the probability to infer sensitive information exceeds a pre-specified threshold, the current query request will then be denied. Further, we analyze user social relations to detect collaborative inference attacks. Therefore, our proposed system can prevent malicious users from obtaining sensitive data.

2. The Inference Framework

As shown in Figure 1, the proposed inference detection system consists of three modules: knowledge acquisition, semantic inference model (SIM), and security violation detection including user social relation analysis.

The *Knowledge Acquisition* module extracts data dependency knowledge, data scheme knowledge and domain semantic knowledge. Based on the database schema and the data sources, we can extract data dependency between attributes within the same entity and among entities. Domain semantic knowledge can be derived by semantic links with specific constraints and rules. A semantic inference model can be constructed based on the acquired knowledge.

[1] This research is supported by NSF grant number IIS-03113283

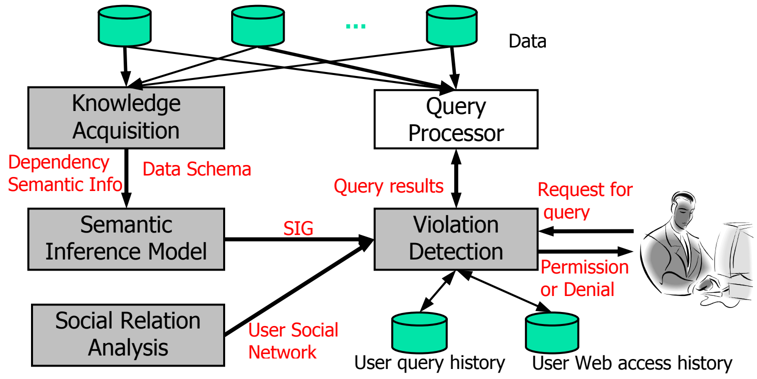


Fig. 1. The framework for an Inference Detection System

The *Semantic Inference Model (SIM)* is a data model that combines data schema, dependency and semantic knowledge. The model links related attributes and entities as well as semantic knowledge needed for data inference. Therefore SIM represents all the possible relationships among the attributes of the data sources. A *Semantic Inference Graph* (*SIG*) can be constructed by instantiating the entities and attributes in the SIM. For a given query, the SIG provides inference channels for inferring sensitive information.

Based on the inference channels derived from the SIG, the *Violation Detection* module combines the new query request with the request log, and it checks to see if the current request exceeds the pre-specified threshold of information leakage. If there is collaboration according to social relation analysis, the *Violation Detection* module will decide whether to answer the current query based on the acquired knowledge among the malicious group members and their social relation to the current user.

3. Knowledge Acquisition for Data Inference

Since users may pose queries and acquire knowledge from different sources, we need to construct a semantic inference model for the detection system to track the users' inference intention. The semantic inference model requires the system to acquire knowledge from data dependency, database schema and domain-specific semantic knowledge.

- *Data Dependency*:

Data dependency represents causal relationships and correlations between attribute values. Let E_i be entity *i*, e_i be the instance of E_i, A and B be attributes of E_i. In the relational model, functional dependency, $A \rightarrow B$, is a type of data dependency where the value of attribute A decides the value of attribute B. The concept of data dependency includes non-deterministic relationships and therefore is more general than functional dependency. We use conditional probability $p_{i|j}=Pr(B=b_i|A=a_j)$ as a parameter to represent the data dependency from B to A.

Data dependency can be divided into two types: *dependency-within-entity* and *dependency-between-related-entities*. Let A and B be two attributes in an entity E. If B depends on A, then for all the instances of E, the value of attribute B depends on the value of attribute A. In this case, we say A and B are dependent within entity. Let A be an attribute in entity E_1, B be an attribute in E_2, and E_1 and E_2 are related by R, which is a relation that can derived from database schema.

- *Database Schema*:

In relational databases, database designers use data definition language to define data schema. The owners of the entities specify the primary key and foreign key pairs. Such pairing represents a relationship between two entities. If entity E_1 has primary key *pk*, entity E_2 has foreign key *fk*, and $e_1.pk=e_2.fk$, then dependency-between-related-entities from attribute A (in e_1) to attribute B (in e_2) can be derived.

- *Domain-Specific Semantic Knowledge*:

For a given database, there are certain semantic relationships among attributes and/or entities which can be represented by the constraints for the attribute values. Since users often pose query with semantic constraints, domain-specific semantic knowledge is needed to transform these constraints into non-semantic terms for query processing [CYC96]. Therefore, such semantic knowledge needs to be acquired and should play a part in the data inference.

Semantic knowledge among attributes is not defined in the database and may vary with context. We can acquire the corresponding set of semantic knowledge based on the set of semantic queries posed by the users. For example, in the following query, WHERE clause #3 and #4 are the semantic conditions that specify the semantic relation "can land" between entity Runways and entity Aircrafts. Based on this query, we can extract semantic knowledge and integrate it into the Semantic Inference Model shown in Figure 3.

■ **Query: Find airports that "*can land*" a C-5 cargo plane.**

```
SELECT AP.APORT_NM
FROM AIRCRAFTS AC, AIRPORTS AP, RUNWAYS R
WHERE AC.AC_TYPE_NM='C-5' and   #1
      AP.APORT_NM = R.APORT_NM and    #2
      AC.WT_MIN_AVG_LAND_DIST_FT <= R.RUNWAY_LENGHT_FT and    #3
      AC.WT_MIN_RUNWAY_WIDTH_FT <= R.RUNWAY_WIDTH_FT;    #4
```

4. Semantic Inference Model

The Semantic Inference Model (SIM) represents dependent and semantic relationships among attributes of all the entities in the information system. As shown in Figure 2, the related attributes (nodes) are connected by links that represent their relationships. There are three types of relation links: dependency link, schema link and semantic link, as follows.

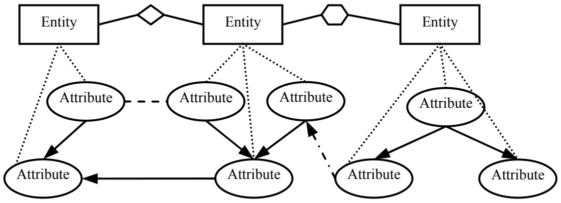


Fig. 2. A Semantic Inference Model. Entities are interconnected by schema relations (diamond) and semantic relations(hexagon).The related attributes (nodes) are connected by their data dependency, schema and semantic links.

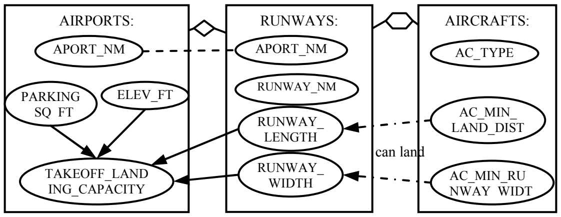


Fig. 3. A Semantic Inference Model example for Airports, Runways and Aircraft.

Dependency link connects dependent attributes within the same entity or related entities.

Schema link connects an attribute of the primary key to the corresponding attribute of the foreign key in the related entities. For example, in Figure 3, APORT_NM is the primary key in AIRPORTS and foreign key of RUNWAYS. Therefore, we connect these two attributes via schema link.

Semantic link connects attributes with a specific semantic relation. The specific semantic relation (e.g., "can land") can be obtained from the domain knowledge or by mining the data sources. The set of candidate semantic relations can be derived from the set of semantic queries.

4.1 Semantic Inference Model Reduction

The large number of related attributes in the SIM can generate a vast number of links. Many of these links are either redundant or superfluous. Therefore, it is desirable for us to simplify the model by reducing the number of redundant links. A SIM consists of linking related attributes (structure) and their corresponding conditional probabilities (parameters). To reduce the model complexity, we generate a set of candidate structures with their corresponding parameters, and select the one that best matches the data sources [FGK96, GTK01, GFK01]. Using a simplified model significantly reduces the complexity in deriving the set of inference channels.

4.2 Semantic Inference Graph

To perform inference at the instance level, we instantiate the SIM with specific entity instances and generate a semantic inference graph (SIG). Each node in the SIG represents an attribute for a specific instance. Related attributes are then connected via instance-level dependency links, instance-level schema links and instance-level semantic links. As a result, the SIG represents all the instance-level inference channels in the SIM.

- *Instance-level dependency link:*

When a SIM is instantiated, the dependency-within-entity is transformed to dependency-within-instance in the SIG. Similarly, the dependency-between-related-entities in the SIM is transformed to dependency between two attributes in the related instances. This type of dependency is preserved only if two instances are related by the instantiated schema link. Consider two dependent attributes A and B. Let A be the parent node and B be the child node. The degree of dependency from B to A can be represented by the conditional probabilities $p_{i|j} =Pr(B=b_i|A=a_j)$. The conditional probabilities of the child node given all of its parents are summarized into a conditional probability table (CPT) that is attached to the child node. For instance, Figure 4b shows the CPT for the node "T" of the SIG in Figure 4a. The conditional probabilities in the CPT can be derived from the database content [FGK99, GFK01].

- *Instance-level schema link:*

The schema links between entities in SIM represent “key, foreign-key” pairs. At instance level, if the value of the primary key of an instance e_1 is equal to the value of the corresponding foreign key in the other instance e_2, which can be represented as R(e_1, e_2), then connecting these two attributes will represent the schema link at the instance level. Otherwise, these two attributes are not connected.

- *Instance-level semantic link evaluation:*

Let T be the target node of the semantic link, P_S be the source node, and P_1, …, P_n be the other parents of T, as shown in Figure 4a. The semantic inference from a source node to a target node can be evaluated as follows:

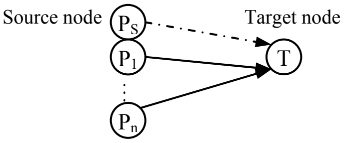
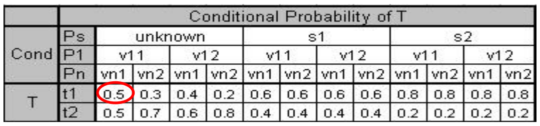


Fig. 4a. Target node T with semantic link from source node P_S and dependency links from parents P_1, …, P_n.



		Conditional Probability of T											
Cond	Ps	unknown				s1				s2			
	P1	v11		v12		v11		v12		v11		v12	
	Pn	vn1	vn2	vn1	vn2	vn1	vn2	vn1	vn2	vn1	vn2	vn1	vn2
T	t1	0.5	0.3	0.4	0.2	0.6	0.6	0.6	0.6	0.8	0.8	0.8	0.8
	t2	0.5	0.7	0.6	0.8	0.4	0.4	0.4	0.4	0.2	0.2	0.2	0.2

Fig. 4b. The CPT of target node T summarizes the conditional probabilities of T given values of P_S and P_1,…,P_n. For example, $Pr(T=t_1|P_S=unknown, P_1=v_{11}, P_n= v_{n1})=0.5$.

If the semantic relation between the source and the target node is unknown or if the value of the source node is unknown, then the source and target node are independent. Thus, there is no semantic link between them. To represent the case of the unknown semantic relationship, we need to introduce the attribute value “unknown” to the source node and set the value of the source node to “unknown.” In this case, the source and target node are independent, i.e., $Pr(T=t_i|P_1=v_1, \ldots P_n=v_n, P_S=unknown) = Pr(T=t_i|P_1=v_1, \ldots P_n=v_n)$. When the semantic relationship is known, the conditional probability table of the target node is updated with the known semantic relationship. If the value of the source node and the semantic relation are known, then $Pr(T=t_i| P_1= v_1, \ldots P_n= v_n, P_S=s_j)$ can be derived from the specific semantic relationship, e.g., in Figure 4b, the semantic relationship decides that $Pr(T=t_1| P_1, \ldots P_n, P_S=s_1)=0.6$ and $Pr(T=t_1| P_1, \ldots P_n, P_S=s_2)=0.8$.

4.3 Evaluating Inference in Semantic Inference Graph

For a given SIG, there are attribute dependencies within an entity, between related entities, and semantic relationships among attributes. As a result, there are many feasible inference channels that can be formed via linking the set of dependent attributes. Therefore, we propose to map the SIG to a Bayesian network to reduce the computational complexity in evaluating users’ inference probability for the sensitive attributes.

For any given node in a Bayesian network, if the value of its parent node(s) is known, then the node is independent of all its non-descending nodes in the network [HMW95, Pea88]. This independence condition greatly reduces the complexity in computing the joint probability of nodes in the network. More specifically, let x_i be the value of the node X_i, pa_i be the values of the parent nodes of X_i, then $P(x_i|pa_i)$ denotes the conditional probability of x_i given pa_i where i=1,2,…,n. Thus, the joint probability of the variables x_i is reduced to the product of $P(x_i | pa_i)$:

$$P(x_1, \ldots, x_n) = \prod_i P(x_i|pa_i) \qquad \textbf{(1)}$$

The probability for users to infer the sensitive node $S=s$ given known evidences $D_i=d_i$, $i=1, 2, \ldots, n$ is:

$$P(s \mid d_1, d_2 \ldots, d_n) = \frac{P(s, d_1, d_2 \ldots, d_n)}{P(d_1, d_2 \ldots, d_n)} \qquad \textbf{(2)}$$

which can be further computed using Equation (1). Thus, the probability of inferring a sensitive node can be computed from the conditional probabilities in the Bayesian network. Many algorithms have been developed to efficiently perform such calculations [Dec96, JLO90, LS88, ZP94].

Probabilistic Relational Model (PRM) is an extension of Bayesian network that integrates schema knowledge from relational data sources [FGK99, GTK01, GFK01]. Specifically, PRM utilizes relational structure to develop *dependency-between-related-entities*. Therefore, in PRM an attribute can have two distinct types of parent-child dependencies: *dependency-within-entity* and *dependency-between-related-entities*, which matches the two types of dependency links in the SIM. Since the semantic links in SIM are similar to dependency links, we can convert each SIM to a PRM-based model. The corresponding Bayesian network can be generated after instantiating the model to instance level. Thus, for a given network, the probability of inferring a specific sensitive attribute can be evaluated via efficient Bayesian inference algorithms. In our test bed, we use SamIam [Sam], a comprehensive Bayesian network tool developed by the Automated Reasoning Group at UCLA, to carry out the inference calculation.

5. Inference Violation Detection

Semantic inference graphs provide an integrated view of the relationships among data attributes, which can be used to detect inference violation for sensitive nodes. In such a graph, the values of the attributes are set according to the answers of the previous posted queries. Based on the list of queries and the user who posted these queries, the value of the inference will be modified accordingly. If the current query answer can infer the sensitive information greater than the pre-specified threshold, then the request for accessing the query answer will be denied.

Consider our previous example in Figure 3, let the TAKEOFF_LANDING_CAPACITY of any airport be the sensitive attribute, and it should not be inferred with probability greater than 70%. If the user has known that:

1. Aircraft C-5 can land in airport LAX runway r1.
2. C-5 has "aircraft_min_land_dist = long" and "aircraft_min_runway_width = wide."

Then this user is able to infer the sensitive attribute LAX's "TAKEOFF_LANDING_CAPACITY=large" via Equation (2) and (1) with probability 58.30%, as shown in Figure 5a.

Now if the same user poses another query about the "Parking_sq_ft of LAX", and if this query is answered (as shown in Figure 5b, "LAX_Parking_Sq_Ft=large"), then the probability of inferring "LAX_TAKEOFF_LANDING_ CAPACITY = large" will increase to 71.50%, which is higher than the pre-specified threshold. Thus, this query request should be denied.

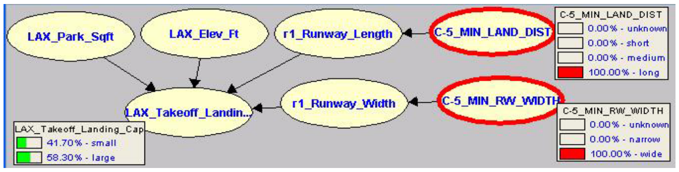


Fig. 5a. Example of inference violation detection for single user. This is a portion of the Bayesian network for the example. The probability distribution of each node is shown in a rectangular box. The values of the bold nodes are given by previous query answers; the probability values of sensitive nodes are inferred.

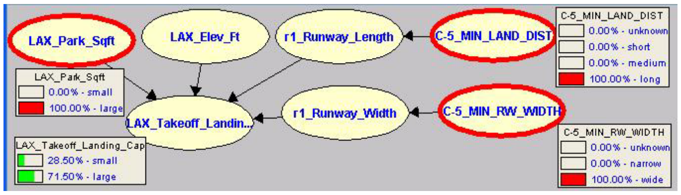


Fig. 5b. Given the additional knowledge "LAX_Parking_Sq_Ft=large", the probability for inferring the sensitive information "LAX_TAKEOFF _LANDING_CAPACITY =large" is increased to 71.50%.

We are currently extending our inference violation detection system from a single user to multiple user cases, where users may collaborate with each other to jointly infer sensitive data. We propose to employ a user social network to model the relationships among cell members for deriving their collaborative inference. A social network is a graph structure that represents the relationship among the user population. The edges of the network represent the influence level of one user to another. Because social relationships vary for each task, the structure and parameters for the social network is also task sensitive. Such a network can be constructed from the answers of questionnaires such as those used in security clearances, personal profiles and interviews. For a given specific task, the amount of information that flows from one user to another depends on how close their relationships are. The collaborative inference probability can be derived based on whether the users' posed query sets are independent or overlap on the inference paths, the users' collaborative relationship (direct or indirect) and their collaboration to each other. Thus, the collaborative inference for a specific task can be derived by tracking and combining each user's query history together with their collaborative levels from the user social network.

6. Related Work

The inference problem has been studied in Privacy-Preserving Data Mining (PPDM). The goal of PPDM is to analyze data through collaborative data mining, and at the same time preserve data confidentiality. One branch of PPDM, Secure Multi-party Computation (SMC), computes certain functions on multiple inputs in a distributed network where each participant holds one of the inputs. SMC wants to ensure that no more information is revealed to a participant in the computation than what can be inferred from the participant's input and final output [CKV02, Pin02, VC03]. This inference took place in the process of multi-party computation, which is a specific scenario of generic data inference.

7. Conclusion

We proposed a technique to prevent users to infer sensitive information from a series of seemingly innocuous queries. Based on the data dependency, the database schema and the semantic knowledge, we constructed a semantic inference model (SIM) that links all the related attributes and thus, represent all possible inference channels from any attributes to the set of pre-assigned sensitive attributes. The SIM is then instantiated by specific instances and reduced to a semantic inference graph (SIG) for inference violation detection to control query access. To reduce computation complexity for inference, the SIG can be mapped into a Bayesian network, where the nodes represent the attributes and links represent the relationships among attributes. Available Bayesian network tool can then be used for evaluating the inference probability along the inference channels. When a user poses a query, the detection system will examine his/her past query log and calculate the probability of inferring sensitive information from answering this posed query. The query request will be denied if it can infer sensitive information with probability exceeding the pre-specified threshold. We are currently extending the detection system to multiple collaborative users based on query history of all the users as well as their social relations.

References

[CCH94] Wesley W. Chu, Qiming Chen and Andy Y. Hwang. “Query Answering via Cooperative Data Inference.” *Journal of Intelligent Information Systems* (JIIS), Volume 3(1): 57-87, 1994.

[CYC96] Wesley W. Chu, Hua Yang, Kuorong Chiang, Michael Minock, Gladys Chow, and Chris Larson. “CoBase: A Scalable and Extensible Cooperative Information System.” *Journal of Intelligence Information Systems* (JIIS). Vol 6, 1996, Kluwer Academic Publishers, Boston, Mass.

[CKV02] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. “Tools for privacy preserving data mining.” *SIGKDD Explorations,* 4(2), December 2002.

[Dec96] Rina Dechter. “Bucket elimination: A unifying framework for probabilistic inference.” In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 211-219, 1996.

[DH96] Harry S. Delugach and Thomas H. Hinke. “Wizard: A Database Inference Analysis and Detection System.” In *IEEE Trans. Knowledge and Data Engeneering*, vol. 8, no. 1, 1996. pp. 56-66.

[FGK99] N. Friedman, L. Getoor, D. Koller and A. Pfeffer. “Learning Probabilistic Relational Models.” *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, August 1999, pages 1300--1307.

[GLQ92] T.D. Garvey, T.F. Lunt, X. Quain, and M. Stickel, “Toward a Tool to Detect and Eliminate Inference Problems in the Design of Multilevel Databases.” *6th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, 1992.

[GTK01] L. Getoor, B. Taskar, and D. Koller. “Selectivity Estimation using Probabilistic Relational Models.” *Proceedings of the ACM SIGMOD (Special Interest Group on Management of Data) Conference*, 2001.

[GFK01] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. “Learning Probabilistic Relational Models.” Invited contribution to the book *Relational Data Mining*, S. Dzeroski and N. Lavrac, Eds., Springer-Verlag, 2001.

[HMW95] Guest Editors: David Heckerman, Abe Mamdani, and Michael P. Wellman. “Real-world applications of Bayesian networks.” *Communications of the ACM*, 38(3):24-68, March 1995.

[HDW96] Thomas H. Hinke, Harry S. Delugach, and Randall Wolf. “Wolf: A Framework for Inference-Directed Data Mining.” *10th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, 1996.

[JLO90] F. V. Jensen, S.L. Lauritzen, and K.G. Olesen. “Bayesian updating in recursive graphical models by local computation.” *Computational Statistics Quarterly*, 4:269-282, 1990.

[LS88] S.L. Lauritzen and D.J. Spiegelhalter. “Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems (with Discussion).” *Journal of the Royal Statistical Society*, Series B, 50(2): 157-224, 1988.

[Pea88] Judea Pearl. “Probabilistic Reasoning in Intelligence Systems.” Morgan Kaufmann, San Mateo, CA, 1988.

[Pin02] B. Pinkas. “Cryptographic techniques for privacy-preserving data mining.” *SIGKDD Explorations,* 4(2), December 2002.

[Sam] SamIam by Automated Reasoning Group, UCLA. http://reasoning.cs.ucla.edu/samiam/

[TFC93] Bhavani M. Thuraisingham, William Ford, M. Collins, and J. O'Keeffe. “Design and Implementation of a Database Inference Controller.” *Data Knowl. Eng.* 11(3), page 271, 1993

[VC03] J. Vaidya and C. Clifton. “Privacy-preserving k-means clustering over vertically partitioned data” In *Proceedings of the 9th ACM SIGKDD*, 2003.

[YL98] Raymond W. Yip, and Karl N. Levitt. “Data Level Inference Detection in Database Systems.” PCSFW*: Proceedings of the 11th Computer Security Foundations Workshop*, 1998.

[ZP94] Nevin Lianwen Zhang and David Poole. “A simple approach to bayesian network computations.” In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 171-178, 1994.