Efficient Reasoning upon Fusion of Many Data Sources

Chase Yakaboski and Eugene Santos, Jr.
Thayer School of Engineering at Dartmouth College
14 Engineering Dr.
Hanover, NH 03755
{chase.th, esj}@dartmouth.edu

Abstract
Bayesian Knowledge Bases (BKB), a graphical model for representing structured probabilistic information, allow for efficient fusion of knowledge from multiple sources. Past research has focused on knowledge fusion situations that only involve a limited number of sources. In this work, we extend the BKB fusion research by exploring the effect of fusing information from many sources. This extension quickly yields a reasoning bottleneck that we overcome by leveraging a representation modification algorithm. With this algorithm, we show how reasoning complexity upon fusion of many sources can be significantly reduced while maintaining the underlying probability semantics of all sources. This develops a means for BKBs to be used in various data fusion problems, allowing previously intractable problems to be studied. We further illustrate our solution empirically using two simulated problems as well as practically through survival time analysis of breast cancer data taken from The Cancer Genome Atlas (TCGA) Program.

Introduction
Knowledge fusion plays a central role in all decision-making scenarios. We would never expect to decide based solely on a limited context when other information is potentially available, so why should our algorithms? By leveraging heterogeneous sensor data and feature extraction techniques, knowledge fusion contextualizes predictions to produce more robust and accurate intelligent systems. Knowledge and information fusion is so critical that the US Department of Defense developed the Joint Directors of Laboratories model (Hall and Llinas 1997) to formalize the information fusion problem with the aim of improving communication between military researchers and system developers.

Many researchers have extended and refined the Joint Directors of Laboratories model (Dasarathy 1997; Meng et al. 2020), and most agree that the information fusion process can be separated into three hierarchical categories: signal level fusion, feature level fusion, and decision level fusion. Signal level fusion generally deals with raw data fusion, such as concatenating raw datasets. Feature level fusion attempts to extract the important parts of the fused signal data and can be accomplished through various feature selection techniques such as principal component analysis. Decision level fusion is the highest level of abstraction where high level deductions about the data, extracted features, and encoded symbols/rules are made.

These data fusion principles have recently emerged in modeling complex computational biology problems, the primary motivation for our approach. For example, researchers have developed means for clustering heterogeneous cancer data to identify potential driver gene mutations (Mo et al. 2013) and cancer subtypes (Pfeifer and Schimek 2021). However, these approaches have a critical limitation. While their clustering approach may yield high accuracy, such methods can only identify patterns. They cannot hypothesize any causal link between features or the underlying mechanism that governs the biological system under study.

In this paper, we show that Bayesian Knowledge Base (BKB) fusion with the addition of our novel representation modification algorithm can be used to address all levels of the information fusion hierarchy without this limitation. At the signal level, we can form BKB fragments (small directed graphical models) that represent each data instance. These fragments are invariant to the sensor type allowing heterogeneous data to be ingested. Further, traditional BKB fusion enables us to fuse these fragments together into a large knowledge base where all fragments are annotated and tracked. Lastly, BKB reasoning permits us to make inferences across this entire fused knowledge base yielding informed decisions based on all information sources. However, a reasoning bottleneck is encountered upon fusion.
This artifact of the BKB fusion process is an added exponential complexity that hampers reasoning tasks and limits the problems that can be addressed.

While studying this problem, we noticed many fragments often shared similar structural configurations. With this observation, we were able to develop a collapsing algorithm that would compress the complexity introduced by fusion and make these problems tractable again. Further, our solution still annotates the data to ease aggregation and de-aggregation of information while maintaining the appropriate probabilistic semantics of the underlying knowledge.

Perhaps most related to our work is research by Elias Bareinboim and Judea Pearl (2013; 2016) who leverage causal models to help alleviate fusion pain points related to transportability of observations taken from clinical studies and applied to different target populations. Using multiple selection diagrams to encode commonalities between clinical studies they could determine which casual relationships could be transported to separate target populations. However, we believe our Bayesian knowledge fusion allows for a greater representational power by means of a source reliability metric with an intuitive graphical structure that now benefits from increased reasoning efficiency even under pressure from many knowledge sources.

In the next section we will describe the structure of BKBs as well as the reasoning and knowledge fusion process. Then we will present our fusion collapse algorithm, which makes reasoning in BKBs remain efficient even in the presence of many fused sources. Finally, we will discuss our algorithm’s properties and empirically demonstrate how our algorithm can allow BKBs, and by extension BNs, to analyze previously intractable problems.

Background

The Bayesian Knowledge Base

As depicted in Fig. 1, a Bayesian Knowledge Base (BKB) is a directed graph that subsumes a Bayesian Network (BN) and provides a model for encoding structured probabilistic knowledge into an efficient representation. Unlike other knowledge bases such as probabilistic databases that store data as tuples with an associated probability value, BKBs assemble a collection of conditional probability rules (CPRs) of the form: if Acid Level = Low ∧ Over Crowd = Yes, then pH Level < 6.5 with probability 0.6. This allows BKBs to establish a dependency structure on the random variable instantiation level thereby handling incomplete information, i.e., missing conditional probability values, as well as cyclic knowledge without having to make any closed world assumptions. Since conditional independence relationships can be specified between random variable states, e.g., Acid Level = Low, as opposed to between random variables, e.g., Acid Level, BKBs offer a finer representational power and supply a mechanism for fusing knowledge from multiple sources. Next, we describe BKBs formally:

Definition 1. A correlation-graph is a directed graph $G = (I \cup S, E)$ in which $I \cap S = \emptyset$, $E \subset \{(I \times S) \cup (I \times S)\}$, and $\forall q \in S \exists a \in I$, such that $(q, a) \in E$. In this graph we refer to all nodes $q \in S$ as Support nodes, or $S$-nodes, and all nodes $a \in I$ as Instantiation nodes, or $I$-nodes. See Fig. 1.

As $I$-nodes are random variable instantiations, then Reproduce Ability = Low and Over Crowd = No would be examples of $I$-nodes in our correlation graph. Similarly, the probability 0.75 above would be encoded as an $S$-node, $q_s$, in the correlation graph with edges, $E = \{(\text{Reproduce Ability} = \text{Low}, q_s), (q_s, \text{Over Crowd} = \text{No})\}$. For the remainder of this paper, we will treat an edge $(a, b) \in E$ as denoting a directed edge of the form $a \rightarrow b$.

We say that two sets of $I$-nodes, $I_1$ and $I_2$, are mutually exclusive, or mutex, if there exists an $I$-node $(R = v_1) \in I_1$ and $(R = v_2) \in I_2$ such that $v_1 \neq v_2$. This property of mutual exclusion guarantees that only one random variable state exists at one time. We can say that two $S$-nodes are mutex if the two $I$-node tail sets of each $S$-node are mutex.

Definition 2. A Bayesian Knowledge Base (BKB) is a tuple $K = (G, w)$ where $G$ is a correlation graph, and $w:S \rightarrow [0,1]$ is a weight function, such that the following properties are met:

1. $\forall q \in S$, the set all incident $I$-nodes to $q$, denoted $\text{Tail}_C(q)$, contains at most one instantiation of each random variable.

2. For distinct $S$-nodes $q_1, q_2 \in S$ that support the same $I$-node, denoted $\text{Head}_C(q_1)$, the sets $\text{Tail}_C(q_1)$ and $\text{Tail}_C(q_2)$ must be mutex.

3. For any $Q \subseteq S$ such that (i) $\text{Head}_C(q_1)$ and $\text{Head}_C(q_2)$ are mutex, and (ii) $\text{Tail}_C(q_1)$ and $\text{Tail}_C(q_2)$ are not mutex, then $\forall q_1, q_2 \in Q, \sum_{q \in Q} w(q) \leq 1$. 

![Figure 1: An example BKB of a simple aquatic ecosystem. I-nodes are represented as squares and S-nodes as solid circles. An example BKB state is $\theta = \{\text{Acid Level} = \text{Low}, \text{Over Crowd} = \text{Yes}, \text{Reproduce Ability} = \text{Low}, \text{pH Level} < 6.5\}$.](image)
More details about BKBs are available in the work of Eugene and Eunice Santos (1999).

Fusion

The goal of Bayesian knowledge fusion is to reason over knowledge taken from disparate knowledge sources that may contain heterogenous and/or incomplete information on their own but form a valuable and more complete knowledge representation when taken together. Fusion takes the CPRs of each knowledge source in the form of Bayesian Knowledge Fragments (BKF$s), which are also valid BKBs, and fuses them together to form a final BKB containing the knowledge from all sources and associated reliabilities. The work of Santos, Wilkinson, and Santos (2011) provides a detailed description of the BKF fusion algorithm.

The BKB fusion algorithm has been used to solve three primary issues that are typically encountered in knowledge fusion problems. First, when experts disagree about a weight relationship in a BKB, fusion can be used so that both experts’ opinions are encoded in the knowledge graph. Second, BKB fusion allows the inclusion of cyclic knowledge. Take the scenario in which two experts disagree about the causality between two random variables. A traditional BN could not handle this contradiction, but since BKBs support cyclic knowledge, both opinions can be fused. Lastly, BKB fusion allows a means for subverting the mutual exclusivity requirement between two or more incompatible S-nodes. For instance, take two BKBs that each have an S-node that contains the same I-node set pointing to two different I-node heads. This condition violates mutual exclusivity but if these S-nodes were separated into two different BKBs and fused, an additional I-node would be inserted into each S-nodes’ tail making them mutex.

Work on BKB fusion has mainly focused on the fusion of a relatively small number of BKBs (<10) that have been well engineered (Santos et al. 2009). While the added complexity in these cases is often negligible, the fusion of orders of magnitude more BKBs results in complexity issues that cannot be ignored. To better understand this fusion complexity and our method for correcting it, we need to discuss reasoning in BKBs.

Reasoning

Reasoning in BKBs takes the form of a graph search to find inferences consistent with a set of target and evidence instantiations, i.e., the joint probabilities over possible worlds. An inference in a BKB is a subgraph that includes at most one instantiation of a random variable. Therefore, the joint probability of an inference, $\tau$, is simply the product of all S-node weights that are involved, $P(\tau) = \prod_{q \in \tau} w(q)$.

We consider two types of reasoning in a BKB: belief revision and belief updating. Belief revision focuses on finding the most probable world given some evidence whereas belief updating, focuses on updating the joint probability of a target I-node given some evidence. In other words, belief updating calculates the updated probability of an I-node given some known evidence about the universe. Since belief updating is generally the more difficult reasoning task, it will be the center of our focus for the remainder of the paper. Formally, belief updating is concerned with calculating the probability of a state by summing up all worlds that are consistent with $\theta$ (see Fig. 1 for an example), $P(\theta) = \sum_{\tau \in \Theta_\theta} P(\tau)$, where $\Theta_\theta$ is the set of all worlds supporting $\theta$.

Collapsing a Fused Bayesian Knowledge Base

The main goal of this work is to circumvent the added complexity of the BKF fusion algorithm in many practical scenarios in which the fusion of many BKBs is required. The intuition for our solution arises when we consider cases when multiple BKF$s$ contain S-nodes with the same structure (head and tail configurations) and identical weights. In such cases, our fusion algorithm places $M$ number of source I-nodes and $M$ number of source S-nodes onto each of these equivalent S-nodes, thereby adding $M$ number of inference paths that need to be explored during reasoning for every BKF S-node.

If we assume each fragment to be fused $K_1, K_2, \ldots, K_M$ has $N$ number of S-nodes with a shared S-node ratio of $\beta$ (the ratio of identical S-nodes shared between respective fragments), that means there are $O((\beta N)^h)$ identical inferences in the fused BKB. If you were to run a belief updating calculation of a state supported by one of these inferences, a computational bottleneck would be incurred as all these exponential number of inferences in the fused BKB would have to be explored during reasoning.

To address this bottleneck, we developed our fusion collapse algorithm to collapse source nodes that share identical S-nodes. By collapsing these source nodes, we can avoid the exponential complexity as all the $M$ inferences developed by introducing the $M$ source I-nodes and S-nodes have been reduced to a single collected source I-node and S-node.

Clearly, when $\beta \to 0$ there is not much that can be done in terms of collapsing source nodes. However, in practice many knowledge bases contain crossover knowledge making $\beta$ in these cases much higher than 0. Further, when we build BKF$s$ directly from raw data, as will be seen in the breast cancer analysis, many S-nodes are shared between BKF$s$. This makes $\beta$ very high, adding significant exponential complexity, and making the problem intractable without our fusion collapsing algorithm.

Our fusion algorithm can be described formally as follows. Let there exist $M$ input fragments $K_1, K_2, \ldots, K_M$ that have been fused into $K' = (G', w')$ with $G' = (I' \cup S', E')$. Also, for any I-node $\alpha$ in the fused BKB, let the random variable that is instantiated by $\alpha$ be denoted as $R_\alpha$. 
BKB-FUSION-COLLAPSE ($K'$)

1. $K'' \equiv (G'', w'') = K'$
2. $G'' \equiv (I'', S'', E'') \equiv (I', S', E')$
3. Let $\Gamma$ be the set of all source I-nodes.
4. Let $\Lambda$ be the set of all non-source I-nodes.
5. for all I-nodes $\alpha \in \Lambda$
6. Let $\Phi = \{q, p \mid q, p \in S', \ Head(q) = Head(p) = \alpha, Tail(q) = Tail(p), w(q) = w(p)\}$
7. Let $\Omega = \{s \mid s \in \Gamma \land s \in Tail(q) \land q \in \Phi\}$
8. $ q \leftarrow \text{Pop} q$ from $\Phi$ and remove its source I-node.
9. Define new source I-node $\omega = (S_{Ra} = \Omega)$
10. Add $\omega$ to $I''$ and add new S-node $q_\omega$ to $S''$
11. Add edges $q_\omega \rightarrow \omega$ and $\omega \rightarrow q$ to $E''$
12. Let $w(q_\omega) \leftarrow \sum_{s \in \Omega} w(q_s)$
13. $I'' \leftarrow I'' \setminus \Omega$
14. $S'' \leftarrow S'' \setminus \{q_s \mid s \in \Omega\}$
15. $E'' \leftarrow E'' \setminus \{(q_s, s) \mid s \in \Omega, q \in \Phi\}$
16. return $K''$

The keen reader will notice that this algorithm only solves our complexity bottleneck when S-nodes $q_1$ and $q_2$ have identical heads and tails (minus the source I-nodes) as well as identical weights. This seems rather restrictive since it is likely that different sources, while having the same S-node structure will likely have different weights. We hope to address this observation in future work to develop an approximate collapsing algorithm so that we can still employ the exponential saving while maintaining most of the integrity of the underlying probability distribution.

**Algorithm Properties**

Since the fused BKB is valid (Santos, Wilkinson, and Santos 2011), the collapsed BKB will also be a valid BKB if all the properties of Definition 2 are satisfied.

**Theorem 1.** The output $K'' = (G'', w'')$ of the BKB-FUSION-COLLAPSE algorithm is a valid BKB.

We now sketch a proof of this theorem:

**Proof** To prove that $K''$ is a valid BKB we just need to show that the three properties in Definition 2 hold.

1. There is at most one instantiation of each variable in the tail of any S-node because we did not add or remove any random variable. Instead, we simply removed all source I-nodes to each S-node and added a single source I-node that represents the entire removed set in its place.

2. If the original fused BKB does not violate the mutual exclusivity requirement, which it will not if all $K_i$ fragments are themselves valid BKBs (Santos, Wilkinson, and Santos 2011), then upon collapse they will also not violate mutual exclusivity. Since we only potentially alter the state of each S-node’s source I-node, there will always be a source I-node pointing to every S-node thereby maintaining the mutual exclusivity that was developed in the fused BKB.

3. Lastly, we need to show that for any $Q \subseteq S$ such that (i) $Head_0(q_1)$ and $Head(q_2)$ are mutex and (ii) $Tail(q_1)$ and $Tail(q_2)$ are not mutex, for all distinct $q_1$ and $q_2$ $\in Q$ their sum is less than or equal to 1. If the set $Q$ is pointing to a source random variable, and all source S-nodes have already been normalized by fusion, then collapsing these sources into distinct groups and summing respective weights will still result in a normalized set. If $Q$ points to a non-source node, then all S-nodes in the set must be from the same set of fragments, because if they were from different set of fragments then they would have a different source collection I-node in their tail, making the tails mutex. □

It is also important that groundness of all I-nodes in the collapsed BKB be maintained because that guarantees that the probabilistic semantics of all the fused fragments are maintained. It was shown by Santos, Wilkinson, and Santos (2011) that groundness of all I-nodes in a resultant fused BKB is maintained and we will extend this proof to show that it applies to the final collapsed BKB as well.

**Definition 3.** A node $\alpha \in I \cup S$ in a BKB is grounded if there exists an inference $\tau$ over the BKB such that $\alpha \in \tau$.

**Theorem 2.** If all S-nodes in the fused BKB are grounded, then all S-nodes in the collapsed BKB are also grounded.

We now sketch the proof of this theorem:

**Proof** All $K_i$ fragments are grounded in the fused BKB $K'$ (Santos, Wilkinson, and Santos 2011). So, let $q_i \in S'$ be an S-node that is originally in the fused BKB $K'$ and from some fragment $K_i$, i.e., it is not a support S-node of any source I-node. Since $q_i$ is grounded in $K'$, then there exists at least one inference $\tau$ such that $q \in \tau$. Furthermore, in the fused BKB there exists a set of inferences $T$ such that $\forall \tau \in T, q \in \tau$, and the only I-node differentiating $\tau$’s is the source I-node that is attached to $q$. The mechanics of our collapsing algorithm simply reduces this set of inferences $T$ to a single inference $\tau''$ where $q \in \tau''$ and the new source I-node attached to $q$ is the source collection I-node. Lastly, since we know that $\tau$ is an inference in $K'$, that means that $\tau''$ is still an inference. Then every S-node is grounded either in an inference $\tau$ from the fused BKB or a new inference $\tau''$ resulting from the collapsing algorithm. □

From this result we can show easily whether I-nodes or sets of I-nodes are grounded in $K''$.

**Corollary 1.** If all S-nodes in the fused BKB are grounded, then all I-nodes in the collapsed BKB are grounded.

Extending Corollary 1 to sets of I-nodes we want to ensure that a set of I-nodes, or state $\theta$, is represented in the same way in the collapsed BKB $K''$. It is important to prove Theorem 2 and describe Corollaries 1 and 2 as they guarantee the original underlying probabilistic semantics are maintained in collapsed BKB.

**Definition 4.** A state $\theta$ is well-represented in $K$ if there exists an inference $\tau = (I' \cup S', E)$ over $K$ s.t. $\theta \subseteq I'$.

**Corollary 2.** All well-represented states in the fused BKB $K'$, are well represented in the collapsed BKB $K''$. 
Empirical Analysis

In order to illustrate the importance of our BKB fusion collapse algorithm we will apply it to a number of belief updating tasks for two simulated problems as well as discuss a practical use case analyzing multi-omics data in breast cancer patients. All experiments were run on a Dell Power Edge M630 compute node with 502GB of RAM and 32 Intel Xeon E5-2640 v3 2.60GHz CPUs.

The first simulated problem we addressed deals with a simple aquatic ecosystem as seen in Fig. 1. This BKB has a total of four random variables, nine I-nodes, and 13 S-nodes. Each experiment featured in Fig. 2 fused the aquatic ecosystem BKB \( N \) number of times, from 1 to 50, and ran 20 belief updating calculations recording runtime and the number of partial inferences explored. The number of partial inferences explored relates to the number of paths that were searched by our reasoning algorithm on its way to finding the final inference set such that the target I-node and all evidence I-nodes are supported by each inference in the set. The number of partial inferences explored is a better measure of complexity than runtime as it remains invariant to machine architecture. However, runtime results are still reported to illustrate the practical impact of our algorithm.

Four experiments were run on our aquatic ecosystem problem to analyze the rate of increasing complexity between different updating tasks. Three of these experiments analyzed 20 randomly generated belief updating tasks with a single target I-node and \( m \) number of evidence I-nodes ranging from one to three. In each of these random experiments both the target I-node and evidence I-nodes were chosen randomly from a uniform distribution. We chose to iterate over the number of evidence I-nodes because in practice greater complexity is often associated with more specified evidence. To demonstrate worst-case complexity, a final experiment was conducted analyzing only the hardest single evidence I-node updates in the BKB. The hardest inferences were determined by ranking updates on the initial unfused BKB by total number of partial inferences explored.

It is clear from Fig. 2 that our collapsing algorithm has a dramatic effect on the complexity of all experiment types: random and worst-case. We can see only a small difference in the complexity curve between the hardest updates and the random updates suggesting that in practice even average complexity approaches the worst-case. Whereas our traditional fusion algorithm’s complexity begins to degrade at around 20 fused sources, our collapsing algorithm remains flat, consistent, and robust to any number of sources.

We extended our aquatic ecosystem problem to a harder goldfish ecosystem that consisted of 57 random variables, 165 I-nodes, and 350 S-nodes. We attempted to generate similar complexity diagrams as Fig. 2, but upon fusing just two sources the problem became immediately intractable blasting through all 502GB of RAM before any update calculation could complete. Our collapsing algorithm still succeeded in handling this problem with an average runtime of 2 seconds per update and remained robust even after thousands of BKBs were fused. While the results from these simulated experiments may seem obvious as we are merely collapsing copies of the same BKB, the approach does have significant utility when forming BKFs from data instances as will be shown in the next experiment.

With our fusion collapse algorithm, we can now address problems previously intractable by both BKB and BN modeling. We illustrate this through a survival time analysis of breast cancer patients based on their multi-omics data. In this analysis we leveraged breast cancer data taken from The Cancer Genome Atlas (TCGA) which collects participating patient’s treatment, survival, and genetic data. Due to the incompleteness in the TCGA dataset, not all patients had all data recorded. Upon preprocessing the data our breast cancer cohort consisted of 643 patients that had heterogenous survival time, treatment, and genetic information available.

Our approach was to form individual patient BKFs as shown in Fig. 3(a), with a predetermined structure and S-node probabilities of 1, since each respective I-node was observed in the patient. We then fused all these patient fragments together to construct a joint distribution of 10,589
gene mutations, 30 drugs and a modifiable discretization of survival time ranges between 0 and 6,593 days. With our initial fusion approach, the problem was immediately intractable as the fused BKB contained ~6.8M I-nodes, ~13.6M S-nodes, and an exponential number of inferences. However, with our fusion collapse algorithm we reduced the BKB to only ~53k I-nodes and ~64k S-nodes, allowing for tractable BKB reasoning given that the inference paths were already shallow due to our imposed BKF structure. This improvement highlights the fact that when forming BKFs from data the fused BKB will often contain repeated knowledge. Our collapse algorithm compresses these repetitions and permits BKBs built from large datasets to be tractable.

Fig. 3(b) studies the average effect of a drug over the top 1,000 most frequently mutated genes that were observed in the dataset. Survival time was discretized to five distinct ranges such that each range consisted of an equal number of patients. Our results indicate a general increase in average predicted survival time in patients that were prescribed Paclitaxel (ChEMBL428647). While an interesting finding that is not a result of any obvious confounders or bias in the dataset, no experiments depicted in Fig. 3(b) would have been computable under traditional BKB fusion without leveraging our collapse algorithm. Thus, our fusion collapse algorithm provides us the opportunity to study problems not originally in scope of BKBs or by extension BNs.

Conclusion and Future Work

We have described an algorithm for circumventing the complexity that is incurred during Bayesian knowledge fusion of many sources while maintaining the underlying probabilistic semantics. We have also provided empirical evidence of the complexity savings through two simulated problems and a clear practical use case in which our algorithm can allow a previously intractable problem in omics analysis to be studied in a belief network context.

In future work we hope to extend our fusion collapse algorithm to handle S-nodes that have the same graphical structure but different weights. Such an extension would allow a trade-off between reasoning complexity and representational power. Lastly, we hope to explore whether our algorithm can help learn BKBs from data. Since our fusion collapse algorithm can help us address large dataset problems like TCGA, we wonder if these methods could be used to learn interesting BKB structures and potentially hypothesize new causal relationships.

Acknowledgements
This work was supported in part by AFOSR Grant No. FA9550-20-1-0032, DURIP Grant No. N00014-15-1-2514 and NIH Award No. 1OT2TR003436-01.

References