Leveraging Data Relationships to Resolve Conflicts from Disparate Data Sources

Romila Pradhan, Walid G. Aref, and Sunil Prabhakar

Purdue University, West Lafayette IN 47907, USA,
{rpradhan, aref, sunil}@cs.purdue.edu,

Abstract. Recently, a number of data fusion systems have been proposed that offer conflict resolution as a mechanism to integrate conflicting data from multiple information providers. State-of-the-art data fusion systems largely consider claims for a data item to be unrelated to each other. In many domains, however, the observed claims are often related to each other through various entity-relationships. We propose a formalism to express entity-relationships among claims of data items and design a framework to integrate the data relationships with existing data fusion models to improve the effectiveness of fusion. We conducted an experimental evaluation on real-world data, and show that the performance of fusion was significantly improved with the integration of data relationships by (a) generating meaningful correctness probabilities for claims of data items, and (b) ensuring that the multiple correct claims output by the fusion models were consistent with each other. Our approach outperforms state-of-the-art algorithms that consider the presence of relationships over claims of data items.

1 Introduction

With the advent of the collaborative web, while innumerable data providers furnish increasing amounts of information on diverse data items, often there is little to no restraint on the quality of data from different providers. Data sources often provide conflicting information either unknowingly (e.g., failing to furnish updated data, making errors during data collection, copying from other sources) or deliberately (e.g., to mislead facts). A number of data fusion techniques have been proposed [1] to resolve data discrepancies from disparate sources and present high-quality integrated data to users. Recently, [2,3] studied the problem of dependence among sources in the context of data fusion whereas [4,5] studied the interdependence among data items in the fusion of spatial and temporal data. However, the space of existing associations between claims of data items has largely been unexplored. Failing to acknowledge these relationships has been observed to account for as much as 35% of false negatives in data fusion tasks [6]. The rich space of relationships among claims of data items makes it challenging to distinguish correct from incorrect information as illustrated next.

ID	**Data Item**	**S$_1$**	**S$_2$**	**S$_3$**	**S$_4$**	**S$_5$**
O$_1$	**Silent Night**		*Christmas*	*Pop**	*Pop/Rock**	
O$_2$	**Feel It Still**	*Pop**	{*Alt Pop Rock*, Rap*}	*Rock**	*Pop/Rock**	*Pop**
O$_3$	**Perfect**		*Pop**	*Classical*	*Pop/Rock**	*Classical*
O$_4$	**Unforgettable**	*Rap**	{*Pop, Alt R&B**}	*Classical*	*Hip Hop**	

Table 1: Table shows five websites providing information about music genres of four songs. Correct claims are marked with a (*).

Example 1. Consider an example of information provided by five websites on music genres of certain songs (Table 1). Sources provide conflicting information for the same data item, e.g., S_2 provides Christmas as the genre for song Silent Night whereas S_3 claims it to be Pop and S_4 provides Pop/Rock as the genre.

Claims for data items exhibit various entity-relationships: (a) Sometimes, claims are *hierarchically* related, e.g., Pop/Rock is a sub-genre of genres Pop and Rock whereas Alt R&B has stylistic origins in Hip Hop; (b) a claim may be referred to by different names, e.g., in the context of music, Hip Hop and Rap are widely considered to agree with each other; (c) claims may be mutually exclusive to other claims. For example, the song Unforgettable may not be simultaneously of the Classical and the Hip Hop genres. Note that entity-relationships among claims can be obtained from domain-specific databases (e.g., structured vocabulary input [7], map databases) and general purpose knowledge bases [8,9]. (The relationships among claims for this example have been obtained from DBpedia [9] and AllMusic [1], the popular online music guide.)

Single-truth data fusion models [10,2] mostly regard claims to be mutually exclusive while some consider implications (or similarities) among the various observations. The approaches adopt ad hoc measures, such as string edit distance, difference between numerical values, and Jaccard similarity, to identify whether or not one claim implies another. These measures, however, may not be directly applicable to data that exhibit relationship semantics different from notions of implications addressed in prior work, e.g., when claims are real-world entities related to each other beyond string edit-distance. On the other hand, multi-truth fusion models [11,3] completely disregard the existence of relationships among claims of data items. Implications between observations may offer completely new scenarios in the multi-truth setting, e.g., integrity constraints may mandate that multiple true claims be associated to each other.

Furthermore, the correctness probabilities produced by different data fusion models often do not reflect the true likelihood of a claim being true: without any integrity constraints, a data fusion model may generate correctness probabilities such that for the song Perfect, sub-genre Pop/Rock rather than genre Pop has a higher probability of being correct. However, since the latter is a broader genre, one would expect it more likely to be true. Existing data fusion models do not account for these kinds of constraints on the correctness probabilities of claims.

[1] www.allmusic.com

Given the knowledge of how different music genres are related to each other, a data fusion system that considers Pop and Pop/Rock to be distinct genres (for the song Perfect) would benefit from the knowledge by re-evaluating the correctness probabilities of these claims and by reconsidering claims provided by sources S_2 and S_4 to improve the output of fusion on other data items. There are, however, certain challenges in integrating the domain knowledge information on entity-relationships among claims with the data fusion process. First, there can be permutations of agreement or disagreement among sources at different *granularities* of information. For example, sources may: (a) agree on a broader concept but disagree on specifics, (b) agree on a specific concept and disagree on broader ones, or (c) may not reach a consensus at any granularity. A naïve solution will gather evidence for and resolve the 'general' claims; however, the downside to the approach is that while we gain confidence about broader claims, no additional evidence is obtained on the correctness of specific claims. Second, existing data fusion models vary widely in their underlying conflict resolution mechanisms (e.g., Bayesian-based, optimization-based, probabilistic-graphical-model-based). We need a way to represent the data relationships that facilitates seamlessly integrating it with the various fusion models. To address the aforementioned issues, we require principled strategies to represent the domain knowledge information on relationships among claims and leverage it effectively to jointly assess data sources and infer correctness probabilities of claims.

In this paper, we address the problem of integrating entity-relationships among claims with data fusion process to improve the effectiveness of existing data fusion models. Our main contributions can be summarized as follows:

- We propose to represent the knowledge of data relationships among claims in the form of an arbitrary directed graph. We outline pre-processing steps for effective representation and efficient traversal of the graph. (Section 4)
- We propose an approach to integrate the directed graph of data relationships with existing data fusion models and propose an algorithm to leverage the graph to generate consistent correct claims for each data item. (Section 5)
- Our experimental evaluation on real-world data shows the applicability of our approach to a wide range of data fusion models and demonstrates that incorporating the domain knowledge of entity-relationships among claims can significantly improve fusion results. (Section 6)

2 Related Work

Data fusion. The problem of conflict resolution as a way to integrate conflicting data from a multitude of data sources has been extensively studied, and a number of data fusion systems have been proposed in the past [1].

The present work provides a general framework to effectively integrate relationships among claims of data items with existing data fusion models.

Leveraging data correlations. The problem of dependencies and correlations among data sources has been studied in the context of data fusion [2,3] whereas correlations between data items have been explored during the fusion of spatial and temporal data [4,5].

While the hierarchical structure of relationships among object labels has been studied extensively in the past, especially in the area of image annotation [12] and classification [13], it has not been exploited fully in data fusion. Few single-truth data fusion systems (that assume each data item to have a single correct claim) have found the approach of considering implications or similarities between claims to improve the effectiveness of fusion [10,2]; the adopted techniques, however, are limited to ad hoc similarity measures between claims. Multi-truth models [3,11], on the other hand, do not consider any associations among claims of data items.

The closest to our work is [14] that proposed using information on partial ordering among claims to discover truth from synthetically generated data and showed that considering partial ordering reduces the error-rate of source quality estimation. Their approach, however, does not capture relations other than partial ordering, e.g., it does not address representation of relations among claims that are equivalent to each other or are mutually exclusive. Moreover, in a bid to limit overestimation, the approach does not take the partial order into account for evaluating source metrics, and considers it partially in determining correct claims of data items resulting in a low overall recall for fusion.

Extracting entity-relationships. The present work does not focus on modeling the entity-relationships for integration with data fusion. With the undeniable success of large-scale knowledge bases [8,9] and the ongoing research on learning entity-relationships [15,16], our framework relies on knowledge bases and domain-specific databases to extract the relationships among claims of data items. The extracted relations are then fed into the data fusion framework to improve the effectiveness of fusion.

3 Problem Formulation

We consider database instance $\mathcal{D}$, data fusion model $\mathcal{F}$ and binary relation $\mathcal{R}$ denoting the entity-relationships among claims of data items in $\mathcal{D}$, and formulate the problem of leveraging relation $\mathcal{R}$ to improve the effectiveness of fusion.

Data Model. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a set of sources that provide claims about data items in set $\mathcal{O} = \{O_1, \ldots, O_m\}$. For a particular data item, say O_i, $S^i = \{S_1^i, S_2^i, \ldots\}$ denotes the ordered list of sources that provide claims $\psi^i = \{\psi_1^i, \psi_2^i, \ldots\}$ about O_i, where source S_j^i provides claim ψ_j^i. The set of unique claims of O_i is denoted by $V_i = \mathit{distinct}(\psi^i) = \{v_i^1, \ldots, v_i^{|V_i|}\}$. The set of sources that provide claim $v \in V_i$ is represented by $S^i(v) \subseteq S^i$, and the set of claims that S_j provides for O_i is denoted by $V_i(S_j) \in V_i$. We represent the observations and distinct claims of data items in $\mathcal{O}$ by $\Psi = \{\psi^1, \ldots, \psi^{|\mathcal{O}|}\}$ and $V = \{V_1, \ldots, V_{|\mathcal{O}|}\}$, respectively.

Example 2. Consider data item O_2 in the example presented in Table 1. $\psi^2 = \{\textsf{Pop, Alt Pop Rock, Rap, Rock, Pop/Rock, Pop}\}$ is the ordered list of claims made by sources in $S^2 = \{S_1, S_2, S_2, S_3, S_4, S_5\}$, where source S_3 provides claim Rock. Also, $V_2 = \{\textsf{Pop, Alt Pop Rock, Rap, Rock, Pop/Rock}\}$. For data item O_2, the set of sources for claim Pop is denoted by $S^2(\textsf{Pop}) = \{S_1, S_5\}$.

Definition 1. *A database $\mathcal{D}$ is a tuple $\langle \mathcal{O}, \mathcal{S}, \Psi, V \rangle$, where $\mathcal{O}$ is the set of data items, $\mathcal{S}$ is the set of sources, $V = \{V_1, \ldots, V_{|\mathcal{O}|}\}$ is the set of claims, and $\Psi = \{\psi^1, \ldots, \psi^{|\mathcal{O}|}\}$ is the set of observations for all data items.*

Definition 2. *A data fusion system $\mathcal{F}$ is a function that takes database $\mathcal{D}$ as input and outputs a set of probability assignments P denoting correctness probabilities of claims and source quality measures $Q^{\mathcal{F}}$:*

$$\mathcal{F} : \mathcal{D} \rightarrow \langle P, Q^{\mathcal{F}} \rangle$$

where $\forall O_i \in \mathcal{O}$, $P(v_i^k) = p_i^k \in [0,1]$ is the correctness of claim v_i^k, i.e., the probability that claim $v_i^k \in V_i$ is correct and $\forall S_j \in \mathcal{S}$, $Q_j^{\mathcal{F}}$ is a vector indicating the quality of source S_j.

Definition 3. *A binary relation $\mathcal{R} \subseteq V \times V$ denotes the entity-relationships among claims $V = \{V_1, \ldots, V_{|O|}\}$ of data items in $\mathcal{O}$.*

Problem Statement. It is required to develop a *relation-aware* data fusion framework, denoted by $\mathcal{F}_{\mathcal{G}}$, that integrates data fusion model $\mathcal{F}$ with relation $\mathcal{R}$ to infer the correctness probabilities of claims in database $\mathcal{D}$.

4 Exploring Entity-Relationships

In this section, we review the various entity-relationships existing between claims of data items and propose a formalism to express the prior domain knowledge of entity-relationships among claims.

4.1 Observations

As an extension to existing relationships among real-world entities, we observe subsumption, overlaps, equivalence and disjointedness among claims of data items (also detailed in Example 1). In the following, we provide an intuition of what these relationships mean in the context of correctness of claims:

- **Subsumption/Overlaps.** A claim may be part of one or more claims, e.g., Pop and Rock, as music genres, are generalization of the Pop/Rock genre. Any source that provides Pop/Rock definitely agrees with the Pop and Rock genres. We say that genre Pop/Rock *implies* or *supports* genres Pop and Rock.
- **Equivalence.** Real-world entities may be referred to differently by different sources and contexts, e.g., Hip Hop music is referred to as Rap in some cultures and contexts. Therefore, any source that provides Hip Hop as a genre agrees with Rap and vice versa. The relation between such claims elicits a bidirectional implication, i.e., both the claims imply each other.
- **Mutual exclusion.** In certain settings, the correctness of a claim may require all other claims to be declared false. For example, a song-listing integration system may mandate that a song be either of genre Alt R&B or Classical but not both. Therefore, if Alt R&B is considered the correct genre for data item O_4, Classical cannot be correct and vice versa.

From these observations, we recognize two themes, namely *implication* and *mutual exclusion*, in the relationship among claims of data items. Implication summarizes subsumption, overlaps and equivalence relationships, and indicates claims that can be correct or incorrect at the same time. Mutual exclusion dictates the set of claims that cannot be simultaneously correct.

4.2 Relationship Model

Based on these two themes, we define relation $\mathcal{R} \subseteq V \times V$ to describe implication (relationship) between two claims: that is $(u, v) \in \mathcal{R}$ if and only if u implies or supports v. We observe that $\mathcal{R}$ is reflexive, transitive and neither symmetric nor antisymmetric (because given $(u, v) \in \mathcal{R}, (v, u)$ may or may not exist in $\mathcal{R}$). Relation $\mathcal{R}$ can be represented in the form of a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = V$, i.e., vertices in $\mathcal{G}$ represent the set of distinct claims in V and edges in $\mathcal{E}$ represent the relation between claims at the corresponding vertices. $\forall(u, v) \in \mathcal{R}, \exists(u, v) \in \mathcal{E}$ denoting the fact that claim represented by vertex u supports that represented by v. In the rest of the paper, where applicable, we will use claim $v \in V$ and the vertex represented by claim $v \in \mathcal{V}$ interchangeably. Subgraph $G_i = (V_i, E_i) \subseteq \mathcal{G}$ represents the relations over claims of data item O_i.

Following standard graph notation, if $e = (u, v) \in \mathcal{E}$, then v is a parent of u and u is a child of v. If there is a path from u to v (denoted by $u \rightsquigarrow v$), then v is an *ancestor* of u and u is a descendant of v. An arbitrary directed graph thus defined captures the observed relations among claims in the following way:

Implication relation is captured by reachability among vertices. If $u \rightsquigarrow v$ in $\mathcal{G}$, then u implies or supports v. Under this definition of implication,

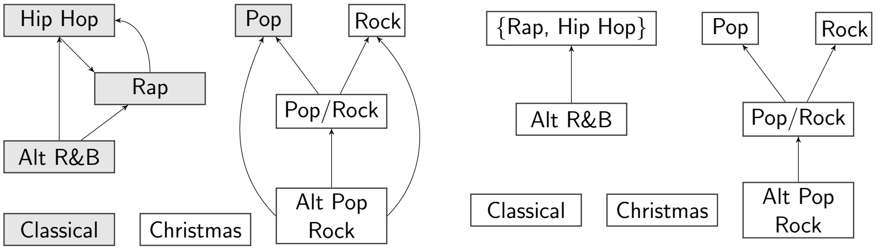
1. v represents *coarser* information than u and encapsulates subsumption.
2. Overlapping claims have a common descendant. Formally, u overlaps with v if there exists w such that $w \rightsquigarrow u$ and $w \rightsquigarrow v$.
3. If $u \rightsquigarrow v$ and $v \rightsquigarrow u$, then u and v represent equivalent claims such that $\mathcal{G}$ contains a cycle which is incident with both u and v. Equivalent claims are represented by *equivalence classes* of vertices in $\mathcal{G}$.

Mutual exclusion is expressed by identifying claims that do not have a common descendant, i.e., u and v are mutually exclusive if $\nexists w$ such that $w \rightsquigarrow u$ and $w \rightsquigarrow v$.

A directed graph (defined as above) over the claims of a data item presents general to specific information as we move from its root (top) to leaves (bottom). When claims are not related, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be seen as a graph with claims as vertices with no edges in between, i.e., $\mathcal{E} = \emptyset$.

Example 3. Figure 1a shows the directed graph of relations over claims of data items in Table 1. Rock and Pop are overlapping claims that have a common descendant: Pop/Rock. Hip Hop and Rap are considered equivalent claims as they are on a cycle incident with both the claims. Moreover, claims Rap and Christmas are mutually exclusive because they do not have a common descendant.

Removing redundancies. The aforementioned directed graph representation can have a large number of redundant edges and vertices as illustrated next. Consider subgraph $G_2 = (V_2, E_2) \subseteq \mathcal{G}$ consisting of claims of data item O_2. Since edge (Alt Pop Rock, Pop/Rock) $\in E_2$ and edge (Pop/Rock, Rock) $\in E_2$, by transitivity, Alt Pop Rock $\rightsquigarrow$ Rock causing edge (Alt Pop Rock, Rock)$\in E_2$ to be redundant. Furthermore, in the subgraph $G_4 = (V_4, E_4) \subseteq \mathcal{G}$ of claims of



(a) Directed graph $\mathcal{G}$. **(b) Modified directed graph $\mathcal{G}^{\mathcal{T}}$.**

Fig. 1: Figure on left shows the directed graph $\mathcal{G}$ of entity-relationships among claims of data items as shown in Table 1. The shaded subgraph denotes relations between claims specific to data item O_4. Figure on right shows modified graph $\mathcal{G}^{\mathcal{T}}$ obtained as transitive reduction of the equivalent acyclic graph of $\mathcal{G}$.

data item O_4, claims Hip Hop and Rap are in the same equivalence class and therefore, can be represented by a single vertex.

We process graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in the following two steps to achieve a concise representation that facilitates effective summarization and efficient navigation:

1. **Redundant Vertices.** We remove redundant vertices in $\mathcal{V}$ by forming the equivalent acyclic graph [17] of $\mathcal{G}$, denoted by $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$. Vertices in $\mathcal{G}^*$ represent equivalence classes in $\mathcal{G}$ and edges in $\mathcal{G}^*$ represent edges between the equivalence classes. $\mathcal{G}^*$ can be obtained by identifying strongly connected components [18] of $\mathcal{G}$. Consider vertices $u, v \in \mathcal{V}$. Let u^* and v^* respectively represent the equivalence classes for claims u and v in $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$. For $u^* \neq v^*$, if $\exists (u, v) \in \mathcal{E}$, then edge $(u^*, v^*) \in \mathcal{E}^*$. Note that $\mathcal{G}^*$ may still have redundant edges because of the transitivity property.
2. **Redundant Edges.** We identify the unique transitive reduction [17] of $\mathcal{G}^*$, denoted by $\mathcal{G}^{\mathcal{T}} = (\mathcal{V}^*, \mathcal{E}^{\mathcal{T}}) \subseteq \mathcal{G}^*$. $\mathcal{G}^{\mathcal{T}}$ has no redundant edge, i.e., for $u, v \in \mathcal{V}^*$, if v is not a parent of u and $u \rightsquigarrow v$, then edge $(u, v) \notin \mathcal{E}^{\mathcal{T}}$. Transitive reduction $\mathcal{G}^{\mathcal{T}}$ has the fewest possible edges and has the same reachability relation as $\mathcal{G}^*$.

Subgraph $G_i^T = (V_i^*, E_i^*) \subseteq \mathcal{G}^{\mathcal{T}}$ represents the transitive reduction of G_i. Figure 1b shows the graph obtained after processing directed graph in Figure 1a.

Complexity Analysis. Equivalent acyclic graph $\mathcal{G}^*$ is obtained in $O(|\mathcal{V}| + |\mathcal{E}|)$ time [18] whereas transitive reduction $\mathcal{G}^{\mathcal{T}}$ can be derived in $O(|\mathcal{V}|^{\beta})$ steps [17] where $\beta \geq 2$. Note that processing directed graph G_i depends only on the number of distinct claims for data item O_i (which is usually not very large) and not on the number of sources that provide information on the item.

In the rest of this paper, we use $\mathcal{G}$ to represent the modified directed graph representation $\mathcal{G}^{\mathcal{T}}$ and G_i to denote G_i^T.

Supporting and Supported Claims. To integrate directed graph $\mathcal{G}$ with existing data fusion models, we need to identify the following two sets of claims for each claim $v \in V_i$: (a) set of claims in V_i that support v, denoted by $\delta(v, G_i)$;

and (b) set of claims in V_i that v supports, denoted by $\alpha(v, G_i)$. After identifying the vertex or equivalence class in G_i that v belongs to, we add claims in the equivalence class and claims that are its descendants in G_i to $\delta(v, G_i)$, and add claims in the equivalence class and claims that are its ancestors in G_i to $\alpha(v, G_i)$. The notion of supporting and supported claims will be used in Section 5.1 to estimate source qualities and correctness of claims.

5 Integration with Data Fusion

Given the entity-relationships among claims as described in $\mathcal{G}$ (obtained in Section 4), in this section we outline the steps for leveraging $\mathcal{G}$ to resolve conflicts during integration of data from multiple sources. We first describe how existing data fusion models can be modified in the presence of $\mathcal{G}$ and then discuss how to utilize $\mathcal{G}$ to determine correct claims for data items.

5.1 Revised Data Fusion Methodology

To determine which of the provided claims are correct and which incorrect, state-of-the-art data fusion models [10,2,3] consider sources to play a pivotal role and usually function in two steps: first, obtain source quality estimates; second, compute the correctness of claims based on the computed source qualities. Given a data fusion model $\mathcal{F}$, characterized by computations of source quality measures $Q^{\mathcal{F}}$ and correctness of claims P, we describe how to modify these two computations for $\mathcal{F}$ given directed graph $\mathcal{G}$ over claims of data items.

- **Estimating Source Quality.** Existing fusion models evaluate sources either in terms of a single measure (e.g., accuracy [2], trustworthiness [10]) or multiple measures (e.g., precision, recall, accuracy, false positive rate [3,11]). The quality of source S_j, denoted by $Q_j^{\mathcal{F}}$, is measured based on $V_i(S_j)$, the set of claims that S_j provides for data item $O_i \in \mathcal{O}$.
  In the presence of entity-relationships among claims, a source, in addition to claims directly provided by it, also implicitly supports claims that are supported by the provided claims. Therefore, $Q_j^{\mathcal{F}}$ depends on claims in $V_i(S_j)$ and claims supported by those in $V_i(S_j)$. Given directed graph $G_i \subseteq \mathcal{G}$ for data item O_i, claim $v \in V_i(S_j)$ supports claims in $\alpha(v, G_i)$ (Section 4). Consequently, we replace $V_i(S_j)$ by $\overrightarrow{V_i}(S_j) = \{\alpha(v, G_i) \mid v \in V_i(S_j)\}$ in the computation of $Q_j^{\mathcal{F}}$. Clearly, $V_i(S_j) \subseteq \overrightarrow{V_i}(S_j)$.

  Example 4. Consider source S_2 in Table 1. Using the modified directed graph in Figure 1b, we observe that S_2 supports claims as shown in Table 2.
  Note that for each data item, we only consider the modified directed subgraph over claims of that particular data item, e.g., since claim Hip Hop $\notin V_2$, we do not consider that Rap supports Hip Hop in the context of data item O_2.
  Comparing Table 2 with Table 1, we observe that out of the 11 claims S_2 supports, 8 are correct resulting in a precision (fraction of claims provided that are correct) of $8/11 = 0.73$. Its recall (fraction of correct claims provided) is $8/11 = 0.73$ as it provides 8 out of the 11 listed correct claims. Note that

ID	$V_i(S_2)$	$\overrightarrow{V_i}(S_2)$	**Correct**
$\mathbf{O}_1$	Christmas	Christmas	Pop, Pop/Rock
$\mathbf{O}_2$	Alt Pop Rock, Rap	Alt Pop Rock, Pop/Rock, Pop, Rock, Rap	Alt Pop Rock, Pop/Rock, Pop, Rock
$\mathbf{O}_3$	Pop	Pop	Pop/Rock, Pop
$\mathbf{O}_4$	Pop, Alt R&B	Pop, Alt R&B, Hip Hop, Rap	Alt R&B, Hip Hop, Rap

Table 2: Claims provided or supported by source S_2.

in the absence of knowledge of relations among the claims of data items, the precision and recall of S_2 would be $3/6 = 0.5$ and $3/11 = 0.27$, respectively.

Procedure EstimateSourceQuality outlines pseudocode for estimating source quality measures given a fusion model and claim relationships. Note that when training data is available, $P(v)$ is defined for items in the training data and $Q^{\mathcal{F}}$ is computed over those items. Otherwise, $Q^{\mathcal{F}}$ is initialized to random values, and source quality and claim correctness are estimated iteratively.

```
Procedure EstimateSourceQuality
Input: Database $\mathcal{D}$, directed graph $\mathcal{G}$, fusion model $\mathcal{F}$, claim correctness $P$
Output: $Q^{\mathcal{F}}$, quality measures of sources
for $s \in \mathcal{S}$ do
    for $O_i \in \mathcal{O}$ do
        $\overrightarrow{V_i}(s) = \{\alpha(v, G_i) \mid v \in V_i(s)\}$
        for $v \in \overrightarrow{V_i}(s))$ do
            Compute $Q^{\mathcal{F}}(s)$ according to $\mathcal{F}$ based on $P(v)$
```

- **Estimating Correctness of Claims.** The second step in data fusion models estimates the correctness of claims by utilizing the estimated source quality measures. The correctness of claim $v \in V_i$, denoted by $P(v)$, is computed in terms of the quality measures of sources in $S^i(v)$, the set of sources that provide v. Claims provided by good sources are considered more likely to be correct than those provided by poor sources.

  Intuitively, the correctness of claim v should depend not only on sources that provide v but also on sources that implicitly support it – the latter can be identified by identifying claims that support v. Given directed graph $G_i \subseteq \mathcal{G}$ for data item O_i, claim v is supported by claims in $\delta(v, G_i)$. In estimating the correctness of v by a particular data fusion model, we replace $S^i(v)$ by $S^i(\overrightarrow{v}) = \{S^i(u) \mid u \in \delta(v, G_i)\}$. Again, $S^i(v) \subseteq S^i(\overrightarrow{v})$. This step ensures that general claims gather greater evidence with support from specific claims and have higher correctness probabilities than them.

  In the presence of directed graph G_i, instead of computing the correctness of each provided claim for data item O_i, we compute the correctness of each vertex in G_i. Doing so, we avoid having to separately estimate the correctness of equivalent claims. Procedure EstimateClaimCorrectness outlines the pseudocode for computing correctness probabilities given the knowledge of relations among claims.

```
Procedure EstimateClaimCorrectness
  Input: Database $\mathcal{D}$, directed graph $\mathcal{G}$, fusion model $\mathcal{F}$, source measures $Q^{\mathcal{F}}$
  Output: $P$ correctness probability of claims
  for $O_i \in \mathcal{O}$ do
      for claim $v \in V_i$ do
          $S^i(\overrightarrow{v}) = \{S^i(u) \mid u \in \delta(v, G_i)\}$
          for $s \in S^i(\overrightarrow{v})$ do
              Compute $P(v)$ according to $\mathcal{F}$ based on $Q^{\mathcal{F}}(s)$
```

Given observations ψ, data fusion model $\mathcal{F}$ and directed graph representation $\mathcal{G}$, as discussed above, we integrate $\mathcal{G}$ with the processes of estimating source quality measures and correctness of claims. We present the pseudocode for modifying $\mathcal{F}$ using $\mathcal{G}$ in Algorithm 1.

```
Algorithm 1: ModifyDataFusion
  Input: Database $\mathcal{D}$, directed graph representation $\mathcal{G}$, data fusion model $\mathcal{F}$
  Output: $P$ correctness probabilities of claims
1 $Q^{\mathcal{F}}$ = EstimateSourceQuality($\mathcal{D}, \mathcal{G}, \mathcal{F}, P$)
2 $P$ = EstimateClaimCorrectness($\mathcal{D}, \mathcal{G}, \mathcal{F}, Q^{\mathcal{F}}$)
```

Iterative fusion models [10,2] randomly initialize source quality estimates and iterate over lines 1 and 2 until $Q^{\mathcal{F}}$ converges. When ground truth data is available, fusion models [3] utilize it to compute source quality estimates.

5.2 Determining correct claims

Having obtained the correctness probabilities, single-truth fusion models will consider claim with the highest probability to be correct and multi-truth fusion models will consider claims with probability greater than a threshold (usually 0.5) to be correct. However, determining correct claims in the standard manner has certain limitations: (a) single-truth fusion models will miss multiple correct claims, and (b) multi-truth fusion models may output correct claims that are indeed constrained to be mutually exclusive.

To address the aforementioned issues, given correctness probabilities P and directed graph $\mathcal{G}$, we describe the steps to determine correct claims for data items in Algorithm 2. Lines 4-6 identify root nodes of the directed graph G_i over claims of data item O_i. Lines 8-10 consider the vertex with maximum correctness probability, `currentNode`, to be correct and add claims in `currentNode` to the list of correct claims for data item O_i. The algorithm then identifies children nodes of the selected vertex for further traversal and repeats lines 8-10 until a leaf node (i.e., vertex with no children) is reached.

6 Experimental Evaluation

This section presents an empirical evaluation of the proposed approach on a real-world dataset. Our objectives are: (1) to assess the effectiveness of using

Algorithm 2: DetermineCorrectClaims

```
Input: Directed graph representation $\mathcal{G}$, correctness probabilities $P$
Output: $V^*$, set of correct claims for data items in $\mathcal{O}$
for $O_i \in \mathcal{O}$ do
    Initialization: considerNodes $= \emptyset$; $V_i^* = \emptyset$
    Let $G_i = (V_i, E_i) \subseteq \mathcal{G}$ be the directed graph over claims in $V_i$
    for vertex $V \in V_i$ do
        if $\nexists \{(V, b) \in E_i\}$ then
            considerNodes $=$ considerNodes $\cup \{V\}$    /* identify root nodes */
    do
        currentNode $= \operatorname{argmax}_{w \in \texttt{considerNodes}} P(w)$
        for claim $v \in$ currentNode do
            $V_i^* = V_i^* \cup \{v\}$
        considerNodes $=$ children of currentNode
    while ( $\exists u \mid (u, \texttt{currentNode}) \in E_i$)
```

the knowledge of entity-relationships among claims in improving the accuracy of existing data fusion models, and (2) to compare the effectiveness of using arbitrary directed graphs against existing approaches that consider prior domain knowledge of entity-relationships among claims of data items.

Competing Methods

We evaluate the effectiveness of using the domain information on entity-relationships among claims on the following single- and multi-truth data fusion models:
Voting: Naïvely assumes correct data to be more frequent than inaccurate data and considers the most frequent claim of a data item to be correct.
TruthFinder [10]: Iteratively computes trustworthiness of sources and confidence in claims, and selects claim with the highest confidence to be correct.
ACCU [2]: Iteratively computes accuracy of sources and correctness of claims by assuming only one claim of a data item to be correct and rest incorrect.
PrecRec [3]: Computes source quality metrics assuming access to ground truth for a subset of data items and uses the estimates to determine correctness of claims. The method outputs multiple correct claims for a data item.

We further compared our approach of using arbitrary directed graphs (denoted by DG) to the partial ordering solution [14](denoted by PO). We implemented all the algorithms in Java.

Performance Metrics

To evaluate effectiveness of the approaches, we present results according to their *precision*, *recall* and F_1*-score*. We measure the precision of an approach as the fraction of claims output by the algorithm that are indeed true. Recall is measured as the fraction of all correct claims that are output by the particular algorithm. We measure the overall performance of an approach in terms of the harmonic mean of its precision and recall, that weighs the two metrics evenly

	Voting	**TruthFinder**	**ACCU**	**PrecRec**
Recall	0.210	0.243	0.251	0.919
Precision	0.758	0.874	0.904	0.835
F1	0.329	0.380	0.393	0.875
% Inconsistent	-	-	-	0.146

Table 3: Effectiveness of data fusion models on Restaurants. Multi-truth fusion model PrecRec is effective in identifying correct claims but outputs claims that may be inconsistent with each other.

$\left(\text{i.e., } F_1 = \frac{2.precision.recall}{precision+recall}\right)$.

% inconsistency: We use the entity-relationships among claims of data items to measure the fraction of pairs of claims considered correct by a data fusion model that are unrelated and *inconsistent* with each other.

Real-World Data
We conducted experiments on the Restaurants dataset in [19] that lists information on restaurants in New York's Manhattan area as provided by 12 sources. We observed that the locations of these restaurants are conflicting but related and, therefore, chose to determine their correct values for the snapshot of data collected on the last available date (3/12/2009).

We identified restaurants by their names and removed those that were chains: if a single source provides inconsistent claims for a restaurant, we consider it to be a chain that may have multiple locations and remove all instances of such restaurants. For example, if a source provides two neighborhoods or two street addresses for the same restaurant, we consider the possibility that it is part of a chain of restaurants. The resulting dataset had 11, 589 unique restaurants (we collected ground truth for 500). It should be noted that, we assume sources to be self-consistent (i.e., a source by itself does not provide inconsistent claims) and ignore errors arising during data collection by humans and sensors.

We extracted the different granularities of locations for restaurants as provided by sources into separate claims. For example, claim "357 East 50th St, Midtown East" was broken down into claims: 357 East 50th St and Midtown East. We extracted relations among the claims using Wikipedia [2] and corroborated with DBpedia and Google Maps. Using the neighborhood definitions, we extracted relations of streets and avenues with neighborhoods. We identified $\sim 1\%$ of restaurants for manual review of relations. Their claims included buildings that were represented by alternate street addresses because of the difference in data collection strategies of different sources.

As a result of inconsistencies across data sources, the resulting directed graph of relations among claims is not just a tree (as in the partial order solution [14]) but can be any arbitrary directed graph with cycles. A partial ordering solution, therefore, will not be directly applicable to resolve such conflicting data.

The case for consistency. To demonstrate the need for approaches that generate consistent correct claims, we run the described data fusion models (Voting,

[2] https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods

TruthFinder, ACCU and PrecRec) on Restaurants and report their performance as measured by precision, recall and F1-measure in Table 3. We observe that while the multi-truth model (PrecRec) is, expectedly, able to retrieve a larger fraction of correct claims, it is less accurate than the single-truth models TruthFinder and ACCU. We dig deeper into the recall of PrecRec and observe that $\sim 15\%$ of pairs of claims considered correct by PrecRec are, in fact, inconsistent with each other (similar results were obtained with synthetic data). The reason for this behavior is that the model considers most of the claims to be correct but is unable to *distinguish* correct from incorrect information. Moreover, the other methods output a single true claim, and hence are inadequate for the current problem. This experiment proves that multi-truth data fusion models are not sufficient for such interrelated data, and that there is indeed a need for approaches that present consistent and accurate data to users.

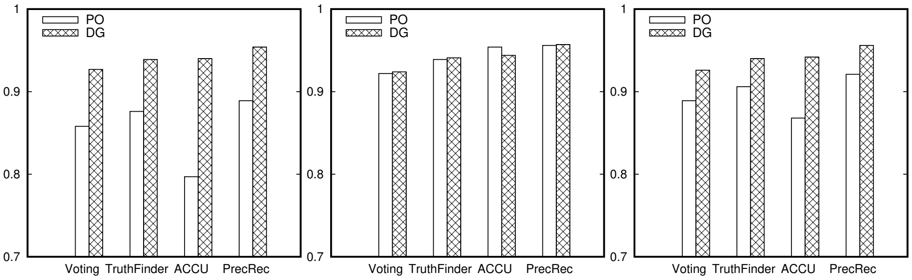
Effectiveness of using data relationships during fusion. We evaluate the advantage of using the knowledge of relations among claims of data items over the effectiveness of different data fusion models. In particular, we have three goals: (a) to evaluate whether the knowledge of relations among claims improves fusion results, (b) to compare the two approaches, PO and DG, and (c) to evaluate how the different data fusion models perform with the knowledge of relations.

	Voting		**TruthFinder**		**ACCU**		**PrecRec**	
	PO	**DG**	**PO**	**DG**	**PO**	**DG**	**PO**	**DG**
Recall	0.889	**0.950**	0.876	**0.939**	0.797	**0.940**	0.889	**0.954**
Precision	0.948	**0.951**	0.939	**0.941**	**0.954**	0.944	0.956	**0.957**
F1	0.917	**0.950**	0.906	**0.940**	0.868	**0.942**	0.921	**0.956**

Table 4: Effect of integrating the entity-relationships among claims on the effectiveness of different fusion models.

We present in Table 4, the results of using DG and PO, entity-relationships among claims, in conjunction with the data fusion models. Comparing the results with Table 3, we find that leveraging data relationships results in an overall improvement in the precision, recall and F1-measure of all data fusion models. The reason for this improvement is that using the knowledge of entity-relationships among claims: (a) single-truth fusion models are converted into multi-truth models, thus retrieving more than one correct claims for each data item and resulting in higher recall, and (b) proper traversal of the graph structures results in less false positives compared to that obtained without the information on relations.

In Figure 2, we compare how the entity-relationship models (PO and DG) fare in conjunction with different data fusion models. Since PO does not support partial orders between claims that result in graphs with cycles, to evaluate PO, we removed edges on cycles in the directed graphs. To determine correct claims in PO, we set the probability threshold, $\theta = 0.05$, i.e., claims with correctness probability higher than 0.05 are considered correct. While both approaches exhibit comparable improvement in precision, DG has consistently higher recall for corresponding data fusion models. This is because DG considers a wide range of



(a) **Recall** (b) **Precision** (c) **F1-score**

Fig. 2: Comparing relationship models PO and DG during fusion of Restaurants. For PO, we set probability threshold $\theta = 0.05$.

relations existing among claims whereas PO is limited only to hierarchies and leaves out ancestors of an overlapping claim that are not reachable from the parent of the claim in question. With an increase in the value of θ, we observe that PO is able to retrieve far fewer correct claims than DG (a difference of around 20% in recall when $\theta = 0.1$ and $\sim 70\%$ with $\theta = 0.3$).

It is worth mentioning how the data fusion models compare against each other in the presence of information about relations. Unsurprisingly, our best case is using DG with PrecRec, when we have access to ground truth for computing source quality measures and have all the information on relations among claims, thus outperforming the other data fusion models across all performance metrics. This is in line with earlier efforts in data fusion that emphasize upon the need for accurate initialization of source quality metrics toward obtaining superior fusion results. It is, however, interesting to note that with the knowledge of data relationships, even the most naïve data fusion technique (Voting) achieves significant improvement in precision and recall – it outperforms state-of-the-art multi-truth model PrecRec that has access to ground truth but no access to domain knowledge (comparing Voting + DG in Table 4 vs. PrecRec in Table 3).

Experiment Takeaways. (1) Leveraging the knowledge on relations among claims improves fusion results. (2) Arbitrary directed graph representation DG is more effective at identifying correct claims than partial ordering solution PO. (3) Unsupervised data fusion models (Voting, TruthFinder, ACCU) perform comparable to supervised models (PrecRec) with DG. This experiment gives rise to an important result: in the presence of domain knowledge, we may not need sophisticated models or ground truth to benefit from the domain knowledge.

7 Conclusions

In this paper, we proposed a formalism to express the prior knowledge of entity-relationships among claims of data items that enables representing a wide range of relationship semantics existing between claims. We designed a framework to integrate the data relationships with the process of fusing conflicting data from disparate sources. We demonstrated the applicability of our approach to a number of existing fusion models, evaluated our approach against other methods that incorporate such relation information in the data, and showed that, compared to other methods, our algorithm achieves significant improvement in fusion results. The effectiveness of our approach depends on completeness of the extracted knowledge and can be improved by accounting for ambiguity in relations. Moreover, directed graphs formalize binary entity-relationships that could be improved with more expressive knowledge representation formalisms (e.g., logic-based, conceptual graphs). We plan to explore these issues in future work.

References

1. Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *SIGKDD Explorations Newsletter*, 2016.
2. X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," *PVLDB*, 2009.
3. R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava, "Fusing data with correlations," in *SIGMOD*, 2014.
4. C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng, "Truth discovery on crowd sensing of correlated entities," in *SenSys*, 2015.
5. S. Wang, L. Su, S. Li, S. Hu, M. T. A. Amin, H. Wang, S. Yao, L. M. Kaplan, and T. F. Abdelzaher, "Scalable social sensing of interdependent phenomena," in *IPSN*, 15.
6. X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang, "From data fusion to knowledge fusion," *PVLDB*, 2014.
7. G. A. Miller, "Wordnet: A lexical database for English," *Commun. of ACM*, 1995.
8. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, 2007.
9. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *ISWC/ASWC*, 2007.
10. X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the Web," *TKDE*, 2008.
11. B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration," *PVLDB*, 2012.
12. A. Tousch, S. Herbin, and J. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognition*, 2012.
13. Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *ICML*, 2007.
14. V. Beretta, S. Harispe, S. Ranwez, and I. Mougenot, "How can ontologies give you clue for truth-discovery? an exploratory study," in *WIMS*, 2016.
15. S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *ECML PKDD*, 2010.
16. M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *ACL*, 2009.
17. A. V. Aho, M. R. Garey, and J. D. Ullman, "The transitive reduction of a directed graph," *SIAM Journal on Computing*, 1972.
18. R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing*, 1972.
19. X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," *PVLDB*, 2009.