GUIDED DATA FUSION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Romila Pradhan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Sunil Prabhakar, Chair

Department of Computer Science

Dr. Walid G. Aref

Department of Computer Science

Dr. Christopher W. Clifton

Department of Computer Science

Dr. Sonia Fahmy

Department of Computer Science

Approved by:

Dr. Voicu S. Popescu by Dr. William J. Gorman

Head of the Department Graduate Program

TABLE OF CONTENTS

Page

LIST OF TABLES x

LIST OF FIGURES xii

SYMBOLS xiv

ABSTRACT xvi

1 INTRODUCTION 1
1.1 Challenges 3
1.1.1 Involving Users During Data Fusion 3
1.1.2 Entity-relationships among Categorical Claims 4
1.2 Summary of Contributions 5
1.3 Outline 6

2 RELATED WORK 7
2.1 Conflict Resolution Systems 7
2.2 Human-in-the-loop Data Integration 8
2.3 User Interaction in Conflict Resolution 9
2.4 Leveraging Data Relationships 10

3 DATA FUSION MODELS 12
3.1 Data Model 12
3.2 Data Fusion Model 14
3.2.1 Bayesian Data Fusion Model: ACCU 15
3.2.2 Probabilistic Graphical Fusion Model: Truthfinder 16
3.2.3 Multi-truth Data Fusion Model: PrecRec 18

4 USER FEEDBACK DURING DATA FUSION 22
4.1 Introduction 22
4.1.1 Motivation and Challenges 23
4.1.2 Solution Overview 25
4.1.3 Summary of Contributions 26
4.2 Problem Formulation 26
4.3 Data Fusion Model 27
4.4 Solution 27
4.4.1 Item-level Ranking Strategies 28
4.4.2 Decision-Theoretic Framework 30
4.4.3 Further Optimizations 44

Page

4.4.4 Feedback Errors . . . 46
4.5 Experimental Evaluation . . . 47
4.5.1 Datasets . . . 47
4.5.2 Competing Methods . . . 49
4.5.3 Evaluation of Ranking Strategies . . . 51
4.5.4 Exploring Approx-MEU . . . 57
4.5.5 Effect of Batch Size . . . 59
4.5.6 Feedback Errors . . . 61
4.6 Summary . . . 64

5 LEVERAGING DATA RELATIONSHIPS TO RESOLVE CONFLICTS . . . 66
5.1 Introduction . . . 66
5.2 Problem Formulation . . . 70
5.3 Exploring Entity-Relationships . . . 70
5.3.1 Observations . . . 70
5.3.2 Relationship Model . . . 71
5.4 Integration with Data Fusion . . . 74
5.4.1 Revised Data Fusion Methodology . . . 75
5.4.2 Determining Correct Claims . . . 78
5.5 Experimental Evaluation . . . 79
5.5.1 Competing Methods . . . 80
5.5.2 Performance Metrics . . . 80
5.5.3 Real-World Data . . . 81
5.5.4 The Case for Consistency . . . 82
5.5.5 Effectiveness of Using Data Relationships during Fusion . . . 82
5.5.6 Synthetic Data . . . 85
5.5.7 Comparison with Partial Order Algorithm . . . 85
5.5.8 Discussion on the Efficiency of Directed Graphs . . . 88
5.6 Summary . . . 88

6 FUTURE WORK . . . 90
6.1 AuthIntegrate: Knowledge Management Module . . . 91
6.1.1 Information Extraction . . . 92
6.1.2 Leveraging Linguistic Cues . . . 92
6.2 AuthIntegrate: Truth Discovery Module . . . 92
6.2.1 Modeling Distrustful Scenarios . . . 93
6.2.2 Broader Characterization of Sources . . . 93
6.2.3 Leveraging Knowledge Bases and Knowledge Representations . . . 94
6.3 AuthIntegrate: Misinformation Manager Module . . . 94
6.3.1 Human-in-the-Loop Conflict Resolution . . . 95
6.3.2 Limiting the Spread of False Information. . . . 96

7 CONCLUSION . . . 97

Page

REFERENCES 99

VITA 106

LIST OF TABLES

Table Page

3.1 An example data table showing four sources providing information about directors of six movies. Correct claims are marked with a (*). 13

3.2 Output of Voting for the example in Table 3.1. 14

3.3 Output of data fusion model ACCU for the example in Table 3.1: Correctness probabilities of claims. 17

3.4 Output of data fusion model ACCU for the example in Table 3.1: Source accuracies. 17

3.5 Output of data fusion model Truthfinder for the example in Table 3.1: Confidence of claims. 19

3.6 Output of data fusion model Truthfinder for the example in Table 3.1: Trustworthiness of sources. 19

3.7 Output of data fusion model PrecRec for the example in Table 3.1: Correctness probabilities of claims. 21

3.8 Output of data fusion model PrecRec for the example in Table 3.1: Quality measures of sources. 21

4.1 Motivating example showing four sources providing information about directors of six movies. Correct claims are marked with a (*). 23

4.2 Output of data fusion for the example in Table 4.1. Value in parenthesis shows the probability that a claim is considered correct. 27

4.3 Probabilities when `Howard` is correct. 34

4.4 Probabilities when `Spencer` is correct. 34

4.5 Expected utility of data items in Table 4.1. 34

4.6 Probabilities when `Docter` is correct. 44

4.7 Probabilities when `leFauve` is correct. 44

4.8 Expected utility of data items in Table 4.1. 44

4.9 Statistics of real-world datasets. 48

4.10 Time taken to determine the next action. 54

Table Page

4.11 Time taken (in seconds) by QBC, US and Approx-MEU$_k$ with different values of k. 59

5.1 Table shows five websites providing information about music genres of four songs. Correct claims are marked with a (*). 67

5.2 Claims provided or supported by source $\mathcal{S}_2$. 76

5.3 Effectiveness of data fusion models on Restaurants. While effective in identifying correct claims, PrecRec outputs inconsistent correct claims. 82

5.4 Effect of integrating the entity-relationships among claims on the effectiveness of different fusion models. 83

LIST OF FIGURES

Figure Page

4.1 The proposed user feedback framework. 25

4.2 Graph of data items in Table 4.1: An edge implies there is at least one source that provides information for the connecting data items. 39

4.3 Long-tail characteristics in real data. Most sources provide information on a small fraction of and few provide data about a large number of items. 48

4.4 Effectiveness of different ranking strategies measured as the reduction in distance_to_ground_truth against number of items validated. 52

4.5 Comparing methods based on entropy utility function (`MEU, Approx-MEU`) against ground-truth-based method (`GUB`). 55

4.6 Scatter plot showing the relation between different performance metrics. . 57

4.7 Hybrid approach combining `QBC` and `Approx-MEU`. Figures depict the effect of expanding the set of candidates for validation in `Approx-MEU`. 58

4.8 Effect of batch size on effectiveness of the methods and time taken to validate 200 claims from FlightsDay. 60

4.9 Conflicting feedback (Books). Each row compares methods when $x\%$ of items have conflicting feedback. 62

4.10 Feedback confidence (Books). Subscript denotes user confidence (or, worker quality). 63

4.11 Incorrect Feedback (FlightsDay). Subscript denotes fraction of items with incorrect feedback. 64

5.1 Figure 5.1(a) shows the directed graph $\mathcal{G}$ of entity-relationships among claims of data items in Table 5.1. Figure 5.1(b) shows modified graph $\mathcal{G}^{\mathcal{T}}$ obtained as transitive reduction of the equivalent acyclic graph of $\mathcal{G}$. 73

5.2 Comparing relationship models PO and DG during fusion of Restaurants. For PO, we set probability threshold $\theta = 0.05$. 83

5.3 Comparing the recall of DG with that of PO on synthetic data with different number of claims per data item. 86

Figure Page

5.4 Comparing the recall of DG with that of PO when the claims of data items are rarely related to each other vs. when they are related quite often. . . . 87

6.1 Figure depicts envisioned architecture of the AuthIntegrate system. . . 91

SYMBOLS

$\mathcal{O}$	set of data items
$\mathcal{O}_i$	the i-th data item
m	number of unvalidated data items
$\mathcal{S}$	set of data sources
$\mathcal{S}_j$	the j-th data source
n	number of sources
$\mathcal{V}$	set of claims on all data items
$\mathcal{V}_i$	set of unique claims for data item $\mathcal{O}_i$
v_i^k	the k-th claim of data item $\mathcal{O}_i$
$\mathcal{S}^i(v)$	set of sources that provide claim $v \in \mathcal{V}_i$
$\mathcal{V}_i(\mathcal{S}_j)$	set of claims that $\mathcal{S}_j$ provides for $\mathcal{O}_i$
Ψ	set of observations made by data sources on data items
ψ^i	ordered list of observations for data item $\mathcal{O}_i$
$\psi_{j,i,k}$	observation of claim v_i^k by source $\mathcal{S}_j$
$\mathcal{S}^i$	ordered list of data sources that provide respective claims in ψ^i
$\mathcal{S}(v_i^k)$	set of sources that vote for claim v_i^k
$\mathcal{D}$	database instance
$\mathcal{F}$	data fusion system
$\mathcal{P}$	correctness probabilities of claims of data items
$\mathcal{Q}^{\mathcal{F}}$	source quality measures for data fusion model $\mathcal{F}$
p_i^k	probability that claim v_i^k of $\mathcal{O}_i$ is `true`
$s(v_i^r)$	confidence in claim v_i^r (Truthfinder fusion model)
$t(\mathcal{S}_j)$	trustworthiness in source $\mathcal{S}_j$ (Truthfinder fusion model)
$\mathcal{A}_j$	accuracy of source $\mathcal{S}_j$ (ACCU fusion model)

$r(\mathcal{S}_j)$	recall of source $\mathcal{S}_j$ (PrecRec fusion model)
$\rho(\mathcal{S}_j)$	precision of source $\mathcal{S}_j$ (PrecRec fusion model)
$q(\mathcal{S}_j)$	false positive rate of source $\mathcal{S}_j$ (PrecRec fusion model)
Θ	set of possible actions
θ_i	action denoting the validation of data item $\mathcal{O}_i$
$\mathcal{H}_i$	entropy of data item $\mathcal{O}_i$
$\mathcal{U}_i$	utility of data item $\mathcal{O}_i$
$EU(i)$	expected utility of data item $\mathcal{O}_i$
$\mathcal{R}$	binary relation denoting entity-relationships among claims in $\mathcal{V}$

ABSTRACT

Pradhan, Romila Ph.D., Purdue University, August 2018. Guided Data Fusion. Major Professor: Sunil Prabhakar.

While the volume and variety of data furnished by disparate data sources has rocketed over the years, often there is little to no restraint over the quality of data available on the Internet; data sources often provide conflicting information for the same data item (a real-world entity or event).

Recent years have witnessed a number of data fusion systems that propose solutions to consolidate multiple instances of a data item, distinguish correct from incorrect information and present a unified, consistent and meaningful record to users. Most of these fusion systems are focused on automatically identifying correct information for data items. Despite their remarkable effectiveness in resolving conflicts, these fusion systems are not error-free and incorrect interpretation on certain data items quickly propagate as false judgement on other items. This dissertation studies techniques to incorporate user feedback and capitalize on the knowledge of relationships among claims of data items to improve the effectiveness of conflict resolution. In particular, the dissertation addresses two key challenges toward guided data fusion.

The first challenge relates to integrating feedback from users to rapidly resolve conflicts. The objective is to effectively and efficiently integrate user feedback for maximum benefit to data fusion. For this purpose, we develop a novel framework built on the principles of decision theory and active learning to reason about the order in which claims should be validated by users. We propose approaches that exploit the structure of interactions between data items and sources and offer interactive validation time for users of a data fusion system.

The second challenge relates to leveraging relations between claims of data items to identify multiple related correct claims. The objective is to recognize existing entity-relationships among claims and integrate them with data fusion systems that are agnostic to data relationships. Toward this goal, we leverage knowledge representations that encapsulate a wide range of relationship semantics and introduce mechanisms to integrate the knowledge representation with data fusion models to retrieve multiple correct claims that are consistent with each other.

Our experimental evaluations using real-world and synthetic datasets demonstrate the effectiveness and efficiency of our proposed approaches to improve conflict resolution of data integrated from multiple sources.

1 INTRODUCTION

This dissertation studies techniques to improve the efficacy and accuray of conflict resoultion while consolidating data from disparate data sources through the judicious inclusion of human input and exploitation of relationships present in data (among data items, data sources and claims provided by sources on particular data items). The objective is to present highly accurate data to end-users in an effective and efficient manner.

With the advent of modern information systems and services, the amount and diversity of data available on the Internet have been growing at an unprecedented pace. Moreover, the number of sources that provide data has significantly increased, spanning well-known sources, such as top news agencies (e.g., CNN and BBC), to individual contributors of Wikipedia articles. In domains such as the Web, sensor networks and social media, it is not surprising to often encounter conflicting data, e.g., financial firms publish inconsistent stock prices of the same company [1], sensors report conflicting measurements [2], different flight-tracking websites publish inconsistent ETAs (expected time of arrival for a flight), on-line bookstores list different authors for identical books [3] and so on. In fact, misleading information has become the new norm; for example, three years before the death of Steve Jobs, the founder of Apple, a top news agency published his obituary to its corporate clients[1]. Resolving such conflicts is important since inaccurate information may result in unfavorable consequences such as financial losses due to an unfortunate drop in the stock price of a corporation following a false obituary or missed flights due to incorrect status information. A perfect example of the damage inconsistent, unverified information can inflict is the steady rise of "fake" news in the media and popular culture. Increasingly, it is becoming difficult for consumers to fathom whether or not

[1]http://fortune.com/2008/08/28/how-steve-jobs-obit-got-published/

a particular piece of information should be trusted unambiguously. The urgency of this matter has prompted concern from inter-governmental agencies [2,3] that consider the dissemination of trustworthy information to be of paramount importance.

Recent years have witnessed a number of data fusion systems (see [4,5] for surveys) that propose conflict resolution, i.e., distinguishing correct from incorrect information, as a way to integrate inconsistent data from multiple sources. Most of the existing data fusion techniques automatically identify correct claims for data items. Although quite accurate, fusion systems are far from perfect and incorrect judgement about one data item quickly trickles down in the form of faulty conclusions about other data items. Particularly for crucial data items, such as medical data, it is essential to distinguish correct claims from incorrect ones. To prevent the spread of inaccurate conclusions and to ensure that fusion systems correctly determine true claims for most data items, feedback should be integrated in the form of validation from users (domain experts). Automated fusion techniques, when augmented with trusted validation of true claims, are expected to steer the system toward a state of higher efficacy.

Furthermore, even though data furnished by different sources may seem independent of each other, there are often inherent data relationships among data sources, data items and claims, and simply relying on users will not be enough. Consider the example of data integrated from disparate sources where sources provide claims at different granularities: while some sources furnish more general claims, some provide very specific details. In such cases, domain-specific databases or general purpose knowledge bases prove the most useful in presenting how the different claims are related. This highlights the need for strategies to incorporate the relation between distinct claims of data items in the presence of conflicting data from multiple sources.

This dissertation seeks to answer the following questions: (1) How can we leverage user feedback for conflict resolution in a sound manner? (2) How can we capitalize on the knowledge of relationships among data items and sources to facilitate efficient

[2]http://www.un.org/apps/news/story.asp?NewsID=56336
[3]http://reports.weforum.org/outlook-14/top-ten-trends-category-page/10-the-rapid-spread-of-misinformation-online/

use of human input? (3) How can we integrate entity-relationships among claims of data items toward improving conflict resolution in data fusion?

In this chapter, we highlight the key challenges addressed in this dissertation in Section 1.1 and present its contributions in Section 1.2. The structure of this dissertations is outlined in Section 1.3.

1.1 Challenges

In this section, we highlight some of the challenges in data fusion where we focus our discussion on two key challenges that this dissertation addresses. The challenges are related to involving the user in the data fusion process, leveraging data relationships to expedite user interaction with fusion models and exploiting relationships among distinct claims of data items during data fusion.

1.1.1 Involving Users During Data Fusion

Existing data fusion systems can be used to identify correct claims for data items. However, the identified claims are not guaranteed to be correct. To instill greater integrity in the system, we can involve users (end-users or domain experts) who can examine the output of fusion and confirm which of the output claims are indeed correct. However, validation of claims by users is a very expensive task. First, it assumes access to highly accurate feedback — preferably from a domain expert. Second, users (domain experts, more so) have limited budget of questions they can answer and typical data fusion datasets have large number of data items and possible claims to be verified. It, therefore, becomes crucial to present the user with the most useful data items or claims to be validated.

The key challenge in involving user feedback is identifying the data item best suited for validation. Since ground truth data may not always be available, we need heuristics to quantify the benefit of validating one data item over another. This task requires designing algorithms housed in the principles of decision theory and active

learning to reason about the improvement in quality of fusion output and to minimize the amount of user interaction.

1.1.2 Entity-relationships among Categorical Claims

Recent years have witnessed tremendous research effors aimed at solving the problem of source selection and source dependence (copying or correlations) during integration. The problem of dependence among claimed values of data items, however, has been unexplored to a large extent. Existing data fusion models mostly consider claims to be independent of each other; the rare augmented models that acknowledge inter-relationships among claims resort to ad hoc similarity measures depending on the data type, e.g., string edit distance, numerical tolerance values, Jaccard distance between sets and so on. This approach of impromptu similarity measures fails to capture relationship semantics that differ from existing notions of similarity.

Consider the scenario of a data item for which two or more data sources provide correct information but at different granularities, e.g., one source provides a general claim while another reports a more specific claim. Sources providing the claims may also exhibit different levels of agreement and disagreement, e.g., sources may broadly agree but disagree about the specifics, or may agree on finer details and disagree on a general level, or may disagree throughout. Single-truth data fusion models that consider a single claim to be correct for a data item would fail to output other related claims that are also correct. On the other hand, multi-truth data fusion models may output correct claims that may not necessarily be consistent with each other.

Existing data fusion models would benefit from the integration of such semantic relationships between claims during the process of conflict resolution. It is, however, not immediately clear how to best represent the various relationship semantics and integrate the knowledge of relationships among claims in a seamless manner across existing data fusion models. There is, therefore, a need to utilize domain knowledge

information on claim relationships and devise techniques to integrate this knowledge during the data fusion process.

1.2 Summary of Contributions

We argue that in order to improve the resolution of conflicting data integrated from disparate data sources, it is important to leverage data relationships (among data items, sources and claims) in an effective and efficient manner. Getting users (domain experts and otherwise) in the loop is crucial because unsupervised, automated data fusion systems are not guaranteed to correctly identify correct claims. Furthermore, leveraging domain-specific databases and general purpose knowledge bases to extract data relationships is helpful in resolving data conflicts.

Specifically, this dissertation makes the following contributions:

- We propose a novel user feedback framework that integrates user input in the form of ground truth labels of claims to rapidly improve the performance of conflict resolution systems. The framework is built on principles of decision theory and active learning to effectively and efficiently solicit validation of correct claims from the user. The objective is to involve users in an interactive pay-as-you-go manner to validate claims that are most beneficial in resolving conflicts. To assess claims efficiently, we delve into relationships among the data items and sources and generate the data item best suited for validation. We implemented a research prototype of the proposed solution demonstrating its applicability, and conducted experimental studies using real-world data to validate its effectiveness.

- We propose incorporating entity-relationships among claims during the process of data fusion where data items may have multiple correct claims. Our framework represents the knowledge of these entity-relationships in the form of an arbitrary directed graph which can be pre-processed for effective representation of relationships and efficient navigation during fusion. We propose modifica-

tions to existing data fusion models for seamless integration of the directed graph of data relationships and propose an approach to generate consistent correct claims for each data item. We implemented our general approach on top of existing fusion models and through experiments on real data, demonstrated its effectiveness in identifying multiple related truths.

1.3 Outline

The rest of the dissertation is organized as follows: Chapter 2 reviews the related work in this area and Chapter 3 presents the data model and existing data fusion models that form the basis of this dissertation. Chapter 4 describes the framework for effectively integrating human input into data fusion systems. In Chapter 5, we introduce the approach to incorporate the knowledge of entity-relationships among claims of data items. We outline directions for future work in Chapter 6 and conclude this dissertation in Chapter 7.

2 RELATED WORK

Integrating data from disparate data sources in a bid to present users with consistent and correct data has been the focus of a large body of research for decades. In this chapter, we briefly review these research efforts. Our work is related to the following research areas: (i) conflict resolution in data fusion, (ii) leveraging user interaction, and (iii) leveraging intrinsic relations present in the data.

2.1 Conflict Resolution Systems

The problem of conflict resolution as a way to integrate conflicting data from a multitude of data sources has been extensively studied in the past and a number of techniques have been proposed (see [5] for a survey). The objective is to identify correct information amidst a multitude of conflicting data from multiple data sources.

The naïve method of majority voting, which considers the most frequently provided value to be true, is not effective when sources report outdated claims, unknowingly provide wrong information or copy from erroneous sources. The earliest approaches to counter majority voting were devised in [6, 7] that propose to identify authoritative sources on the Web. However, in the context of integration of data from a subset of sources, these techniques may not be employed directly as the relatively smaller set of sources does not reflect their true trustworthiness.

Over the last decade, data fusion techniques proposed the following general principle of conflict resolution in data fusion to truly represent the credibility of sources: the amount of trust in a source is measured directly in terms of the correctness of claims that it provides, and the correctness of a claims depends on the trustworthiness of sources that invest in it. Following this general principle, most of the data fusion techniques can be categorized as: Bayesian-based [2, 3, 8, 9], optimization-based [10],

or probabilistic graphical models-based [11]. Other variants have been proposed that leverage source dependencies [8, 12] and incorporate prior knowledge to improve the performance of fusion [9, 10, 13–15]. Most of the data fusion techniques offer automated (unsupervised) solutions to resolve conflicting data whereas a few [12, 15, 16] engage some level of supervision by utilizing ground truth data to identify trustworthy sources. None of these works address the problem of careful selection of ground truth data to improve and expedite conflict resolution.

2.2 Human-in-the-loop Data Integration

As a product to end-users, data management systems should work closely with people in primarily three stages: (a) in determining the objectives for the system; (b) in providing the necessary input to drive the system to a better state; and (c) in evaluating the system through outputs.

Recent years have witnessed an increase in research efforts that involve humans in the data management pipeline. Specifically, user feedback has been previously employed in a number of data management problems such as schema matching [17,18], dataspaces [19], entity resolution [20], classification [21] and data cleaning [22,23]. The goal of most of these works is an optimized utilization of human input by asking a minimal number of questions that would maximize an improvement in the quality of the system. Toward this goal, concepts from decision theory and active learning have proved useful.

Active learning [24] is based on the key idea that machine learning algorithms can achieve greater accuracy from ground truth data provided they are supplemented with a careful selection of ground truth labels. Active learning has been studied in prior research on estimating parameters in Bayesian networks [25] that provides an approximate algorithm to find the query that reduces the expected risk the most, and is heavily dependent on the specific querying algorithm. Dealing with observed and hidden variables in the context of data fusion, approximate solutions from [25] are

not directly applicable. To make informed decisions on determining the sequence in which human input is received, utility elicitation [26] details classical utility functions to narrow down user preferences under uncertainty.

Involving users is often associated with a fixed budget; utilizing an expert in such a scenario is costlier than employing a readily-available crowd of workers. Ongoing research in collecting input from a crowd [27–29] is a related area of work because of the possibly varied characteristics of users in the feedback framework. The problem of noisy labels has been extensively studied in [30, 31] that jointly estimate user quality and true labels of data items. Crowd workers and modeling their behavior add an orthogonal dimension to the problem of selecting the best data items for validation. In the presence of noisy feedback from a crowd of workers, any of the existing crowdsourcing approaches can be used to obtain the most accurate label for data items and plugged into our user feedback framework.

2.3 User Interaction in Conflict Resolution

Solicitation of user feedback has been studied before in the context of conflict resolution [23,32,33] where the focus is to primarily use master data along with editing rules and integrity constraints. The problem of determining the next data item to validate is related to the suggestion generation task in [23] that aims at asking the user a minimal set of attributes knowing which the true values of all attributes of an entity could be deduced. Toward this goal, it specifies the currency of data in terms of available temporal information and currency constraints derived from semantics of the data. Their approach, however, does not address the qualities of data sources in resolving data conflicts.

The problem of validating correct claims is also similar to the task of deducing certain regions in [32] that leverages user feedback and editing rules to deduce deterministic regions to further facilitate fixing errors in data. These approaches, however, are not data-agnostic and assume domain-specific constraints. By disregarding the

role of sources, these approaches are unable to draw conclusions about attributes of data items other than the true values of the one in question.

In particular, [12,15,16] propose conflict resolution mechanisms in the presence of ground truth data and [32] incorporates master data for resolving conflicts. Both of these forms of prior knowledge could be considered a form of user input. However, the main drawback of these approaches is that such prior knowledge is considered static and no effort is made to maximize the benefit of user input while also incorporating minimal amount of user interaction. We realize that the benefit from incorporating pre-meditated user input could be less than that achieved when the user is actively involved in confirming correctness of claims.

2.4 Leveraging Data Relationships

Source relationships. Significant research has been done on source selection and dependence and correlations among data sources [8, 12, 34–36]. In [8, 34, 35] the authors assume that data sources are either original contributors or copiers that obtain their information from the originators, and provide solutions to discover copying relationships for improved data fusion. In contrast, [12] consider positive and negative correlations among data sources — positive, when sources have similar data extraction patterns and negative, when they provide complementary information or information on different data types.

Relationships between data items. Data fusion systems largely consider data items to be independent to each other. However, increasingly, it is becoming evident that data items are often inter-related. The problem of correlations between data items have been explored during the fusion of spatial and temporal data [37–39]. It still remains a challenge to automatically discover the relationships among data items — a task that becomes difficult with the large scale of data items integrated from different sources.

Entity-relationships among claims. Real-world categories are often observed to exhibit relationships that can be extracted from rich authoritative information resources. The undeniable success of large-scale knowledge bases [40,41] and domain-specific databases, such as structured vocabulary input [42], medical databases (e.g., RxNorm[1]) and map databases, offers unforeseen opportunities to leverage entity-relationships. Furthermore, the ongoing research on learning entity-relationships [43–46] presents complementary solutions to and is the backbone of the problem of integrating data relationships with data fusion systems. The hierarchical structure of relationships among object labels has been studied extensively in the past, especially in the area of image annotation [47–49] and classification [50,51]. However, it has not been exploited fully in the context of data fusion. 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 [3,8]; the adopted techniques, however, are limited to ad hoc similarity measures between claims such as edit distance for similarity between strings, tolerance for proximity in numbers, Jaccard similarity index to gauge how similar two sets are, etc. Multi-truth models [12,16], on the other hand, neither consider any associations among the different claims of data items nor mandate the various truths about data items to be consistent with each other.

In [52], the authors proposed that the information on partial ordering among claims can be used to discover truth from synthetically generated data and showed that this approach reduces the error-rate of source quality estimation. There are certain limitations of this work. First, it 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. Second, in a bid to limit over-estimation, 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.

[1]https://www.nlm.nih.gov/research/umls/rxnorm/

3 DATA FUSION MODELS

In this chapter, we describe the preliminaries for the rest of the dissertation. We present the data model and several data fusion models used in Chapter 4 and Chapter 5.

3.1 Data Model

In this section, we describe the data model of a data fusion system and formulate the problem of ordering user feedback for effective conflict resolution in data fusion.

Let $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be a set of sources that provide claims about data items in set $\mathcal{O} = \{\mathcal{O}_1, \ldots, \mathcal{O}_m\}$. Sources provide specific claims for data items modeled as a set of observations $\Psi = \{\psi^1, \ldots, \psi^{|\mathcal{O}|}\}$. Observations for data item $\mathcal{O}_i$ are represented by $\psi^i = \{\psi_{j,i,k}\}$ where

$$\psi_{j,i,k} = \begin{cases} 1 & \text{if } \mathcal{S}_j \text{ votes for claim } v_i^k \text{ of } \mathcal{O}_i \\ 0 & \text{otherwise} \end{cases}$$

Each data item o_i can have a number of claims. For data item $\mathcal{O}_i$, $\mathcal{S}^i = \{\mathcal{S}_1^i, \mathcal{S}_2^i, \ldots\}$ denotes the ordered list of sources that provide (with slight abuse of notation) claims $\psi^i = \{\psi_1^i, \psi_2^i, \ldots\}$, where source $\mathcal{S}_j^i$ provides claim ψ_j^i. The set of unique claims of $\mathcal{O}_i$ is denoted by $\mathcal{V}_i = \mathit{distinct}(\psi^i) = \{v_i^1, \ldots, v_i^{|\mathcal{V}_i|}\}$. The set of claims on all data items is denoted by $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_{|\mathcal{O}|}\}$.

The set of sources that provide claim $v \in \mathcal{V}_i$ is represented by $\mathcal{S}^i(v) \subseteq \mathcal{S}$, and the set of claims that $\mathcal{S}_j$ provides for $\mathcal{O}_i$ is denoted by $\mathcal{V}_i(\mathcal{S}_j) \in \mathcal{V}_i$.

Example 3.1.1 *Consider data item* CATCH-22 *in the example presented in Table 3.1. The set of all claims about it is* $\mathcal{V}_{\text{CATCH-22}} = \{$*Joseph Heller, J. D. Salinger*$\}$ *and*

Table 3.1.: An example data table showing four sources providing information about directors of six movies. Correct claims are marked with a (*).

Source	Data Item	Claim	Correct?
Catch-22	$\mathcal{S}_2$	Joseph Heller	✓
Catch-22	$\mathcal{S}_3$	Joseph Heller	✓
Catch-22	$\mathcal{S}_4$	J. D. Salinger	
Fahrenheit 451	$\mathcal{S}_1$	Michael Wolff	
Fahrenheit 451	$\mathcal{S}_3$	Ray Bradbury	✓
Lord of the Flies	$\mathcal{S}_2$	William Golding	✓
Lord of the Flies	$\mathcal{S}_3$	Arundhati Roy	
Haroun and the Sea of Stories	$\mathcal{S}_2$	Salman Rushdie	✓
Haroun and the Sea of Stories	$\mathcal{S}_4$	Chris Haroun	

the fact that source $\mathcal{S}_2$ *provides claim* Joseph Heller *and not* J. D. Salinger *is represented by setting* $\psi_{S_2,\text{CATCH-22,Joseph Heller}} = 1$ *and* $\psi_{S_2,\text{CATCH-22,J. D. Salinger}} = 0$. *The ordered list of claims respectively made by sources in* $\mathcal{S}^{\text{CATCH-22}} = \{\mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ *is denoted by* $\psi^{\text{CATCH-22}}$=$\{$Joseph Heller, Joseph Heller, J. D. Salinger$\}$ *where source* $\mathcal{S}_4$ *provides claim* J. D. Salinger. *The set of sources for claim* Joseph Heller *is* $\mathcal{S}^{\text{CATCH-22}}$(Joseph Heller) $= \{\mathcal{S}_2, \mathcal{S}_3\}$.

Definition 3.1.1 *A database* $\mathcal{D}$ *is a tuple* $\langle \mathcal{O}, \mathcal{S}, \Psi, \mathcal{V} \rangle$ *where* $\mathcal{O}$ *is the set of data items,* $\mathcal{S}$ *is the set of sources,* $\mathcal{V} = \{V_1, \ldots, V_{|\mathcal{O}|}\}$ *is the set of claims, and* $\Psi = \{\psi^1, \ldots, \psi^{|\mathcal{O}|}\}$ *is the set of observations for all data items.*

Given all components defined above, we formally introduce a data fusion system with its input and output structures:

Definition 3.1.2 *A data fusion system* $\mathcal{F}$ *is a function that takes database* $\mathcal{D}$ *as input and outputs a set of probability assignments* $\mathcal{P}$ *denoting correctness probabilities of claims and may or may not output source quality measures* $\mathcal{Q}^{\mathcal{F}}$*:*

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

where $\forall \mathcal{O}_i \in \mathcal{O}$, $\mathcal{P}(v_i^k) = p_i^k \in [0,1]$ *is the correctness of claim* v_i^k*, i.e., the probability that claim* $v_i^k \in \mathcal{V}_i$ *is correct. When* $\mathcal{F}$ *outputs* $\mathcal{Q}^{\mathcal{F}}$, $\forall \mathcal{S}_j \in \mathcal{S}$, $\mathcal{Q}_j^{\mathcal{F}}$ *is a vector of source metrics that indicate the quality of source* $\mathcal{S}_j$.

3.2 Data Fusion Model

We start this section with describing the details of data fusion: consider a set of data sources $\mathcal{S}$ that provide conflicting claims on data items in $\mathcal{O}$; the goal of data fusion is to identify the correct claim of each data item $\mathcal{O}_i \in \mathcal{O}$.

Traditional conflict resolution systems often resort to *majority voting* to determine the correct claim for data items – a claim that is provided by the maximum number of sources is considered to be correct and rest are considered false. Under this naïve assumption, the probability of claim v_i^r of data item $\mathcal{O}_i$ being true is computed as:

$$p_i^r = \frac{\sum_{j=1}^{n} \psi_{j,i,k}}{\sum_{r=1}^{|V_i|} \sum_{j=1}^{n} \psi_{j,i,r}} \tag{3.1}$$

Table 3.2 presents the correctness probabilities of claims as obtained through voting. For each data item, the claim with the highest correctness probability is considered correct.

Table 3.2.: Output of Voting for the example in Table 3.1.

Data Item	**Correctness Probabilities**
Catch-22	Joseph Heller (0.67), J. D. Salinger (0.33)
Fahrenheit 451	Ray Bradbury (0.5), Michael Wolff (0.5)
Lord of the Flies	William Golding (0.5), Arundhati Roy (0.5)
Haroun and the Sea of Stories	Salman Rushdie (0.5), Chris Haroun (0.5)

Majority voting technique disregards the role of data sources in determining the correctness of claims. Recent years have witnessed a surge of data fusion models that

consider the characteristics of data sources as important in assessing the quality of claims they provide. The key idea behind this approach is the intuition that trusted sources often furnish trustworthy information whereas it is difficult to completely trust data provided by untrusted or less trusted data sources. On the basis of this intuition, the correctness of claims in these data fusion models depends upon the quality of sources providing those claims. In the following, we describe a few such data fusion models:

3.2.1 Bayesian Data Fusion Model: ACCU

This model, proposed in [8] (as `AccuNoDep`) and [53], considers sources to be characterized by their accuracies, i.e., how often do they publish correct information, and use this metric to compute the correctness of claims they provide. The model, in its basic form, assumes sources to be independent; because of its ease of understanding and interpretation, this fusion model forms the basis for a number of other variants of fusion [8, 35, 54] that consider source dependence in assessing the qualities of sources.

ACCU is a Bayesian data fusion model that has observations (the votes of sources on claims, Ψ), and hidden variables ($\mathcal{A}$: the accuracies of sources, and p_i^k: the correctness probabilities of claims); the objective is to infer the hidden variables given the observations. This goal is achieved in the following two iterative steps:

1. **Correctness of a claim.** The model uses Bayesian analysis to compute the correctness of a claim from the accuracies of sources that support it. The probability of claim v_i^r of data item $\mathcal{O}_i$ being `true` is computed as:

$$p_i^r = p(v_i^r{=}\texttt{true} \mid \psi_{.,i,.}) = \frac{\prod\limits_{s \in S(v_i^r)} \frac{(|V_i|-1)\mathcal{A}(s)}{1-\mathcal{A}(s)}}{\sum\limits_{v_i^o \in V_i} \prod\limits_{s \in S(v_i^o)} \frac{(|V_i|-1)\mathcal{A}(s)}{1-\mathcal{A}(s)}} \tag{3.2}$$

where $\psi_{.,i,.}$ represents the observations for data item $\mathcal{O}_i$ and $\mathcal{S}(v_i^r)$ is the set of sources that vote on claim v_i^r of $\mathcal{O}_i$. In this model, only one of the claims is considered to be `true` and the rest are considered `false`.

2. **Accuracy of a source.** Source accuracies are updated using the current probabilities of claims. The accuracy of source $\mathcal{S}_j$ is defined as the probability that its claim about a data item is `true`, and is computed as the average probability of its claims being `true`:

$$\mathcal{A}(\mathcal{S}_j) = \frac{\sum_{i=1}^{m}{}^{\psi_{j,i,k}=1} p_i^k}{N(\mathcal{S}_j)} \tag{3.3}$$

where $\mathcal{S}_j$ provides information about $N(\mathcal{S}_j)$ data items.

Sources are initially assigned default accuracies. The model alternates between the aforementioned two steps until it reaches a steady state (either sources accuracies or correctness probabilities converge) or attains the threshold for number of iterations. Note, however, that ACCU is not guaranteed to converge [34] although in practice, it does converge for datasets typical to data fusion. At the end of convergence, for each data item, the claim having the highest correctness probability is considered to be correct and the rest false.

Tables 3.3 and 3.4 show the output of fusion after the model has converged for the example in Table 3.1. The values in parenthesis in Table 3.3 show the probabilities of claims being considered correct.

3.2.2 Probabilistic Graphical Fusion Model: Truthfinder

The input of data fusion for this model [3] is represented as a probabilistic graphical model [55]. In this model, sources are characterized and assessed by their *trustworthiness* which is utilized to infer *confidence* in facts provided by them. Similar to ACCU, Truthfinder has observations (Ψ), and hidden variables (t: trustworthiness

Table 3.3.: Output of data fusion model ACCU for the example in Table 3.1: Correctness probabilities of claims.

Data Item	**Correctness Probabilities**
Catch-22	Joseph Heller (1) J. D. Salinger (0)
Fahrenheit 451	Ray Bradbury (0.98) Michael Wolff (0.02)
Lord of the Flies	William Golding (0.72) Arundhati Roy (0.28)
Haroun and the Sea of Stories	Salman Rushdie (1) Chris Haroun (0)

Table 3.4.: Output of data fusion model ACCU for the example in Table 3.1: Source accuracies.

Source	**Accuracy**
$\mathcal{S}_1$	0.02
$\mathcal{S}_2$	0.91
$\mathcal{S}_3$	0.76
$\mathcal{S}_4$	0.00

of sources, and s: confidence of claims); the following steps outline how to infer the trustworthiness of data sources and confidence of claims:

1. **Confidence of a claim.** The model computes the confidence of a claim based on the trustworthiness of sources that provide it. The confidence of claim v_i^r of data item $\mathcal{O}_i$ is calculated as:

$$s(v_i^r) = 1 - \prod_{\psi_{j,i,r}=1} (1 - t(\mathcal{S}_j)) \tag{3.4}$$

2. **Trustworthiness of a source** is defined as the expected confidence of claims it provides and is calculated as the average confidence of its claims:

$$t(\mathcal{S}_j) = \frac{\sum_{i=1}^{m}{}^{\psi_{j,i,k}=1} s(v_i^k)}{N(\mathcal{S}_j)} \tag{3.5}$$

   where $N(\mathcal{S}_j)$ is the number of claims provided by $\mathcal{S}_j$.

To avoid dealing with underflow caused by multiplication of unusually small trustworthiness values, Truthfinder uses logarithms that facilitates easy computation of trustworthiness and confidence values (details can be found in [3]). Truthfinder iteratively computes the trustworthiness of data sources (begins with default uniform values) and confidence of claims, and stops when the variables attain a steady state.

Tables 3.5 and 3.6 show the output after Truthfinder finishes computation for the example in Table 3.1. The values in parenthesis in Table 3.5 show the confidence of claims.

3.2.3 Multi-truth Data Fusion Model: PrecRec

This multi-truth data fusion model [12] characterizes sources by their recall, precision and false positive rate calculated over training data (assumes access to ground

Table 3.5.: Output of data fusion model Truthfinder for the example in Table 3.1: Confidence of claims.

Data Item	**Confidence of Claims**
Catch-22	Joseph Heller (0.65) J. D. Salinger (0.57)
Fahrenheit 451	Ray Bradbury (0.58) Michael Wolff (0.57)
Lord of the Flies	William Golding (0.58) Arundhati Roy (0.58)
Haroun and the Sea of Stories	Salman Rushdie (0.58) Chris Haroun (0.57)

Table 3.6.: Output of data fusion model Truthfinder for the example in Table 3.1: Trustworthiness of sources.

Source	**Trustworthiness**
$\mathcal{S}_1$	0.57
$\mathcal{S}_2$	0.60
$\mathcal{S}_3$	0.60
$\mathcal{S}_4$	0.57

truth for a subset of data items) and considers multiple claims for a data item to be correct. The source quality measures are defined over the entire training data as:

1. **Recall** of a data source, $r(\mathcal{S}_j)$, is defined as the fraction of all correct claims that the source provides. Formally,

$$r(\mathcal{S}_j) = \frac{\sum_{\substack{i=1 \\ \psi_{j,i,k}=1}}^{m} \left| \{ v_i^k \mid v_i^k = \texttt{true} \} \right|}{\sum_{\substack{i=1 \\ \psi_{.,i,k}=1}}^{m} \left| \{ v_i^k \mid v_i^k = \texttt{true} \} \right|} \tag{3.6}$$

   Precision of a data source, $\rho(\mathcal{S}_j)$, is calculated as the fraction of claims it provides that are correct. Formally,

$$\rho(\mathcal{S}_j) = \frac{\sum_{\substack{i=1 \\ \psi_{j,i,k}=1}}^{m} \left| \{ v_i^k \mid v_i^k = \texttt{true} \} \right|}{\sum_{\substack{i=1 \\ \psi_{j,i,k}=1}}^{m} \left| \{ v_i^k \} \right|} \tag{3.7}$$

   False positive rate of data source, $q(\mathcal{S}_j)$, is derived from its recall and precision using Bayes rule [12] as:

$$q(\mathcal{S}_j) = \frac{\alpha}{1-\alpha} \cdot \frac{1-\rho(\mathcal{S}_j)}{\rho(\mathcal{S}_j)} \cdot r(\mathcal{S}_j) \tag{3.8}$$

   where α is the *a priori* probability that of a claim being correct.

2. The **correctness probability** of a claim is then computed in terms of quality measures of both sources that provide the claim and those that do not provide the claim as:

$$\mathcal{P}(v_i^k) = \frac{1}{1 + \frac{1-\alpha}{\alpha} \cdot \frac{1}{\mu}} \tag{3.9}$$

   where

$$\mu = \prod_{\psi_{j,i,k}=1} \frac{r(\mathcal{S}_j)}{q(\mathcal{S}_j)} \prod_{\psi_{j,i,k}\neq 1} \left(\frac{1-r(\mathcal{S}_j)}{1-q(\mathcal{S}_j)} \right) \tag{3.10}$$

Table 3.7.: Output of data fusion model PrecRec for the example in Table 3.1: Correctness probabilities of claims.

Data Item	**Correctness Probability of Claims**
Catch-22	Joseph Heller (1), J. D. Salinger (0)
Fahrenheit 451	Ray Bradbury (1), Michael Wolff (0)
Lord of the Flies	William Golding (1), Arundhati Roy (0.33)
Haroun and the Sea of Stories	Salman Rushdie (1), Chris Haroun (0)

Table 3.8.: Output of data fusion model PrecRec for the example in Table 3.1: Quality measures of sources.

Source	**Precision**	**Recall**
$\mathcal{S}_1$	0	0
$\mathcal{S}_2$	1	0.75
$\mathcal{S}_3$	0.67	0.5
$\mathcal{S}_4$	0	0

Tables 3.7 and 3.8 show the output when PrecRec is run on the example in Table 3.1. The values in parenthesis in Table 3.7 show the correctness probabilities of claims. Claims having correctness probabilities higher than a certain threshold (usually, 0.5) are considered correct.

4 USER FEEDBACK DURING DATA FUSION

In this chapter, we propose a novel pay-as-you-go framework for supervised data fusion to judiciously leverage user feedback and rapidly improve the performance of fusion. We describe the various components of our framework that aims at rapidly resolving conflicts during data fusion with minimal user involvement.

This chapter is organized as follows: in Section 4.1, we present a motivating example for the problem of utilizing user feedback in data fusion, describe the solution overview and outline the summary of our contributions. We formally present our problem in Section 4.2 and describe the data fusion model in Section 4.3. Section 4.4, we discuss two broad ranking mechanisms to present questions to the user for feedback — our algorithms in Section 4.4.1 assess data items individually while the framework in Section 4.4.2 ranks data items by their ability to boost the performance of fusion. We empirically evaluate our algorithms on two real-world datasets with different characteristics in Section 4.5, and finally summarize the chapter in Section 4.6.

4.1 Introduction

Recently, a number of *data fusion* systems have been proposed to deal with conflicting data sources, and discriminate `true` and `false` claims of data items (see [5] for a survey). Most of the existing fusion techniques automatically identify correct claims for data items. Although quite accurate, fusion systems are not error-free; incorrect predictions quickly trickle down to other data items as faulty conclusions about correctness of claims. Particularly for crucial data items, it is essential to distinguish correct claims from incorrect ones. To prevent the spread of inaccurate conclusions and to ensure that the fusion system correctly determines true claims for most data items, feedback should be integrated in the form of validation from

Table 4.1.: Motivating example showing four sources providing information about directors of six movies. Correct claims are marked with a (*).

ID	Data Item	S_1	S_2	S_3	S_4
$\mathcal{O}_1$	Zootopia		Howard*	Spencer	Spencer
$\mathcal{O}_2$	Kung Fu Panda	Stevenson*		Nelson	
$\mathcal{O}_3$	Inside Out		leFauve	Docter*	
$\mathcal{O}_4$	Finding Dory				Stanton*
$\mathcal{O}_5$	Minions	Coffin*	Renaud		
$\mathcal{O}_6$	Rio	Jones		Saldanha*	

an expert. Automated fusion techniques with trusted validation of true claims are expected to steer the system toward a state of higher efficacy.

4.1.1 Motivation and Challenges

Consider an example of websites (sources) providing information on directors of certain animation movies as shown in Table 4.1 (correct claims of data items are marked with a *). Data fusion systems take the table of conflicting claims as input, and output the correctness of each claim (and, in some cases, the accuracy of each source, i.e., the probability that a claim provided by the source is correct).

Source S_1 provides `Howard` as the director for the movie `Zootopia` whereas sources S_3 and S_4 claim it to be `Spencer`. A data fusion system that predicts `Spencer` to be correct can benefit from the validation that `Howard` is instead true. With this knowledge, the fusion system can reconsider the claims provided by sources S_1, S_3 and S_4 and improve its output on other data items.

Validation of claims per se is an expensive task; to guarantee effective conflict resolution, it assumes access to highly accurate feedback (e.g., domain experts). To judiciously utilize the expert, claims should be presented for validation in an order that is most beneficial to the performance of fusion. Assuming we can validate any

data item (by asking an expert or using crowdsourcing), and know which of its claims is correct, which item should we select for validation?

The task of identifying the *best* data item for validation is challenging because we have to deal with a number of issues. First, we do not possess ground truth and, therefore, need to develop heuristics to select the best data item. Second, we need to quantify the definition of 'best' i.e., what is the basis for deciding whether or not one data item is more suitable for validation than another? Third, data fusion typically deals with a large number of claims (hundreds of thousands), thus limiting the ability to ask questions on a very small fraction of all claims. Fourth, since each data item may potentially influence any other item, the exhaustive search of estimating the impact of each item on all others by re-running fusion, is prohibitively expensive. For example, to evaluate data item $\mathcal{O}_1$ for validation, we need to assess its impact on all the $(2 + 2 + 2 + 1 + 2 + 2) = 11$ distinct claims of six data items. Similarly checking all data items to select the first item for validation would require $6 * 11 = 66$ computations. Scaling up this costly procedure to millions of claims is infeasible.

To this end, there are two major observations. First, data items have different *levels of uncertainty* because of the agreement/disagreement of sources on claims. One may expect that validating `"Minions"` would be more advantageous than validating `"Zootopia"` because S_1 and S_2 disagree on `"Minions"` while two of the three sources that vote for `"Zootopia"` agree on a common value. This is because we expect to learn more from the validation of data items with *disagreement*. Second, although a data item may have conflict over its values, validating it may not be beneficial if it does not influence enough items. For instance, validating `"Finding Dory"` would have an effect only on `"Zootopia"` whereas validating `"Zootopia"` would impact all other items.

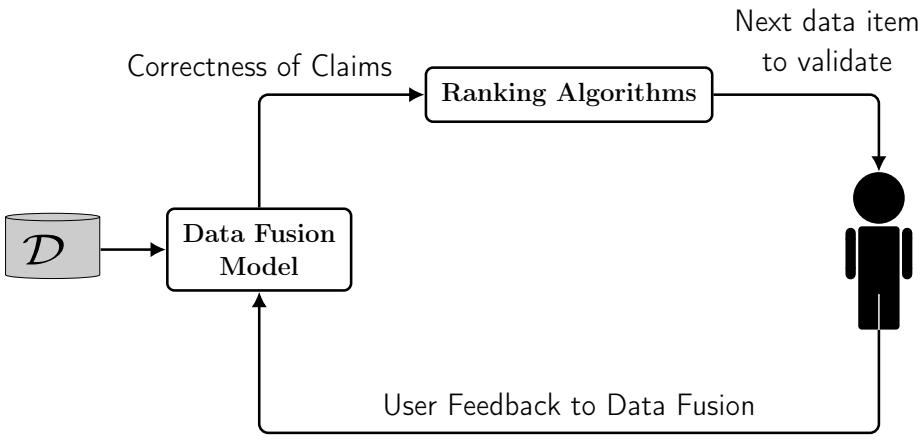


Figure 4.1.: The proposed user feedback framework.

4.1.2 Solution Overview

Given data and the output of fusion, we focus on the problem of determining the best data item for the user to validate (Figure 4.1) without using ground truth information.

To generate an ordering in which data items should be validated, we propose two *item-level* ranking strategies that evaluate data items individually based on their local characteristics. We discuss limitations of the item-level ranking approaches, and propose a novel decision-theoretic framework that assesses data items holistically. Our framework uses the concept of *value of perfect information* (VPI) [56] that is based on a *utility function* to measure the desirability of the current state of a system for its users, and selects a claim validating which maximizes gain in the utility function. We show that this procedure leads to a prohibitively expensive computational cost because we need to fuse data each time we wish to compute the utility gain of a data item. To scale up our framework to large-scale datasets, we propose to analytically estimate the impact of a validation on other unvalidated data items, and select a claim that has the maximum utility gain over the estimates.

4.1.3 Summary of Contributions

We address the problem of utilizing user feedback effectively to improve the performance of existing fusion techniques. Our main contributions are:

- We formalize the problem of ordering user feedback for effective conflict resolution in data fusion based on probabilistic graphical models (Section 4.2).
- We propose strategies to generate an effective ordering in which claims should be validated. Our item-level ranking strategies consider data items individually (Section 4.4.1) while our novel decision-theoretic framework, based on the concept of value of perfect information, evaluates data items holistically (Section 4.4.2).
- To scale up the decision-theoretic framework, we derive approximation formulae that quantify the impact of a validation by analytically estimating the change it effects in other claims. (Section 4.4.2)
- We conduct an extensive experimental evaluation on real-world datasets where we demonstrate the efficacy of the proposed methods in improving conflict resolution, and present trade-offs between user involvement and effectiveness of the methods. (Section 4.5)

4.2 Problem Formulation

In this section, we formulate the problem of ordering user feedback for effective conflict resolution in data fusion.

Feedback Solicitation. To improve the effectiveness of a data fusion system, we solicit feedback in the form of validation of a data item, e.g., we ask the user to provide the `true` director of `Zootopia`.

Action. The validation of a data item $\mathcal{O}_i \in \mathcal{O}$ is called an *action* and is denoted by θ_i. The space of possible actions Θ, is determined by the set of data items that have not yet been validated.

Table 4.2.: Output of data fusion for the example in Table 4.1. Value in parenthesis shows the probability that a claim is considered correct.

ID	Probabilities of Claims
$\mathcal{O}_1$	Howard (0), Spencer (1)
$\mathcal{O}_2$	Stevenson (0.015), Nelson (0.985)
$\mathcal{O}_3$	Docter (0.999), leFauve (0.001)
$\mathcal{O}_4$	Stanton (1)
$\mathcal{O}_5$	Coffin (0.921), Renaud (0.079)
$\mathcal{O}_6$	Saldanha (0.985), Jones (0.015)

Problem Statement. Given a data fusion system $\mathcal{F}$ and its output $\langle \mathcal{P}, \mathcal{Q}^{\mathcal{F}} \rangle$, we solve the problem of determining the next action θ_i from the set of possible actions Θ to solicit feedback from a user.

4.3 Data Fusion Model

In this chapter, we deploy the user feedback framework atop the ACCU data fusion model as described in Section 3.2. Table 4.2 shows the output of fusion after ACCU has converged for the example in Table 4.1.

As shown in Figure 4.1, we treat the data fusion model as a black-box and use the output of fusion to determine the next action which is, thus, independent of the convergence of the fusion model. In the next section, we outline ranking algorithms that leverage only the data and the output of fusion to determine the next action.

4.4 Solution

In the present work, we propose two broad ranking approaches to generate the order in which data items should be validated. The *item-level* ranking strategies presented in Section 4.4.1 consider data items individually, while the decision-theoretic feedback framework of Section 4.4.2 evaluates data items based on their ability to impact the performance of fusion on other unvalidated data items.

4.4.1 Item-level Ranking Strategies

This section presents two techniques that assess the local characteristics of data items to determine the next action. The techniques presented are built upon the principle of *uncertainty* inherent in a data item. Intuitively, an item with greater uncertainty offers more information to the system.

We propose using *entropy* [57] to quantify the average information content in a data item. Entropy is a way to measure the level of uncertainty in probabilistic objects. In the context of data fusion, data item $\mathcal{O}_i$ is a probabilistic object whose `true` claim ranges over all of its possible claims $v_i^k \in V_i$. The entropy of $\mathcal{O}_i$ is defined as:

$$\mathcal{H}_i = -\sum_{v_i^k \in \mathcal{V}_i} p_i^k \log p_i^k \tag{4.1}$$

where p_i^k is the probability that claim v_i^k is `true`.

A data item that has a low entropy has a higher degree of certainty, i.e., some claim has a high probability of being `true`, compared to a data item having claims that are almost equally likely. On the contrary, a low entropy means we can be more certain about true/false labels that should be attached to the claims. However, it also encapsulates the case when a false claim is predicted `true` with a high probability.

Using entropy as the uncertainty measure, the next action is determined as validating the data item that has the highest entropy, i.e.,

$$a_i = \underset{\theta_i \in \Theta}{\operatorname{argmax}}\ \mathcal{H}_i \tag{4.2}$$

We now present our item-level ranking algorithms that elaborate on obtaining p_i^k to use in Equation (4.1). In Section 4.4.1, we present an algorithm based on the disagreement of sources over claims of a data item whereas Section 4.4.1 presents an algorithm that ranks data items based on the output of data fusion.

Disagreement-based algorithm

This section presents Query-by-Committe (QBC), a widely used technique in active learning [24] that is based on the disagreement of sources over claims of a data item. QBC is built upon the principle of majority voting where the true claim of a data item is the one supported by most of the sources. The intuition behind QBC is that an item is less likely to be predicted incorrectly by fusion if most of the sources agree upon it while the true claim of an item disputed by many sources may be questionable. In such cases, it might be more beneficial to validate the latter data item.

QBC uses the votes of sources over claims to compute the probability of correctness of a claim $v_i^k \in \mathcal{V}_i$ as the fraction of sources (voting for $\mathcal{O}_i$) that support v_i^k:

$$p_i^k = \frac{\sum_{j=1}^{n} \psi_{j,i,k}}{\sum_{r=1}^{|\mathcal{V}_i|} \sum_{j=1}^{n} \psi_{j,i,r}} \tag{4.3}$$

This definition of p_k^i is used in Equation (4.1) to evaluate the uncertainty intrinsic to a data item and is termed as *vote entropy* [24]. The data item queried by QBC is the one most disagreed upon by sources that vote for it.

Example 4.4.1 *In Table 4.1, the vote entropy of $\mathcal{O}_2$ is computed as $\mathcal{H}_2 = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 0.693$, which is greater than the vote entropy of $\mathcal{O}_1$ ($\mathcal{H}_1 = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.637$). QBC would validate $\mathcal{O}_2$ before it validates $\mathcal{O}_1$.*

QBC has a low computational cost because it does not need to recompute entropies after a validation. However, one major drawbacks of QBC is that it does not take into account the dependencies between data items through sources.

Uncertainty-based algorithm

The first and foremost limitation of QBC is that the choice of the next action is determined solely by distribution of source votes on claims of a data item. It is

agnostic to the output of fusion, i.e., it does not consider (i) accuracy of sources, and (ii) probabilities of correctness of claims. For the example in Table 4.1, QBC may select O_3 for validation although its true claim has already been identified (Table 4.2).

To overcome this weakness, we present an uncertainty-based technique that selects an action the fusion system is less certain about. Uncertainty sampling, denoted by US, uses probabilities of correctness of claims as output by data fusion to compute the entropy in Equation (4.1). Intuitively, data items that the fusion system is least certain about are more suitable for validation, since the more confident predictions are probably correct.

Example 4.4.2 *The entropy of $\mathcal{O}_5$ in Table 4.1 is computed using the probabilities in Table 4.2, is $\mathcal{H}_5 = -(0.079)\log(0.079) - (0.921)\log(0.921) = 0.276$. $\mathcal{H}_5$ is greater than the entropy of all other data items and, therefore, US considers $\mathcal{O}_5$ the most suitable for validation.*

Beacuse of its ease of interpretation and implementation, uncertainty sampling is one of the most commonly used strategies in active learning. However, unlike QBC, US considers the output of fusion, and therefore, takes source accuracies into account. The downside is that we need to run the fusion system each time we validate an action.

One of the major drawbacks of the item-level ranking approaches is that these methods aim to resolve conflicts at the site of a single data item without any regard to the conflicts existing in other data items. In the following section, we present a framework that assesses data items with the objective of resolving conflicts in all unvalidated data items.

4.4.2 Decision-Theoretic Framework

The techniques presented in Section 4.4.1, although computationally inexpensive, determine actions with the local view of resolving individual conflicts. An additional limitation is that none of the methods considers possible interdependence among

data items and, therefore, offers no guarantee on the improvement of fusion over other unvalidated data items.

Our objective is to globally identify the *best* action that would benefit fusion on all unvalidated data items. To this end, we design a decision-theoretic feedback solicitation framework based on the value of perfect information. The framework defines a *utility function* to measure the usefulness of the current state of fusion, and identifies an action that is most likely to improve the utility of data fusion for all unvalidated data items. To the best of our knowledge, none of the earlier works incorporates the value of information for the problem of data fusion.

Background Concepts

We introduce the basic concepts of our framework such as utility and the value of perfect information. We show that in the absence of ground truth, we have to rely on an alternative utility function based on the idea of uncertainty reduction (referred to as the *entropy utility function*).

Utility function. We define the utility function as a function that measures the usefulness of a data fusion system. The utility of a system is higher if it is able to predict a greater number of `true` claims correctly. Let $\mathcal{T} : v_i^k \rightarrow \{\texttt{true, false}\}$ be a truth function that assigns `true` to a correct claim and `false` to an incorrect claim.

Definition 4.4.1 *Given truth function $\mathcal{T}$, database $\mathcal{D}$ and fusion system $\mathcal{F} : \mathcal{D} \rightarrow \langle \mathcal{P}, \mathcal{Q}^{\mathcal{F}} \rangle$, the utility function $\mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T})$ is defined as:*

$$\mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T}) = \frac{1}{|\mathcal{V}|} \left(\sum_{\mathcal{V}_i \in \mathcal{V}} \sum_{v_i^k \in \mathcal{V}_i} \frac{p_i^k \delta(\mathcal{T}(v_i^k))}{\mid \mathcal{V}_i \mid} \right)$$

$$\textit{where } p_i^k \in \mathcal{P} \textit{ and } \delta(v) = \begin{cases} 1 & \textit{if } v = \textit{true} \\ 0 & \textit{otherwise} \end{cases}$$

The utility function can be interpreted as measuring the average probability of `true` claims based on the output of fusion system $\mathcal{F}$. The closer the utility function is to 1, the higher is the effectiveness of $\mathcal{F}$.

Value of Perfect Information. We measure the usefulness of an action θ_i with respect to our utility function by the value of perfect information (VPI). VPI has been used widely in areas such as economics [58], healthcare [59], data cleaning [17, 19, 22, 29, 60] and classification [21].

Definition 4.4.2 *The value of perfect information (VPI) of action θ_i is defined as:*

$$VPI(\theta_i) = \sum_{v_i^k \in \mathcal{V}_i} \mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T} \mid \mathcal{T}(v_i^k) = \texttt{true}) p_i^k - \mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T})$$

The VPI of action θ_i is the expected gain in the utility function earned by validating data item $\mathcal{O}_i$. To compute $\mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T} \mid \mathcal{T}(v_i^k) = \texttt{true})$, the information that $v_i^k =$ `true` is input to the data fusion system as prior knowledge by setting $p_i^k = 1$ and $p_i^f = 0 \; \forall \; v_i^f \in \mathcal{V}_i \setminus \{v_i^k\}$. The fusion model uses this additional information in its computation of correctness of claims and accuracies of sources.

A set of all possible actions, denoted by Θ, consists of an action θ_i for each unvalidated data item $\mathcal{O}_i \in \mathcal{O}$. Our goal is to identify the action that has the highest VPI, i.e.,

$$\theta_i = \underset{\theta_i \in \Theta}{\operatorname{argmax}} \;\; VPI(\theta_i) \tag{4.4}$$

Maximum Expected Utility

Real-world applications prevent us from using the utility function from Definition 4.4.1 because we do not possess the truth function $\mathcal{T}$, i.e., ground truth is not available. To this end, we propose using an *entropy utility* function to identify actions that reduce the uncertainty associated with the output of fusion. This idea, known as *uncertainty reduction*, has been extensively used in the past [17, 29, 61–63].

Definition 4.4.3 *Given database D and data fusion system $\mathcal{F} : \mathcal{D} \rightarrow \langle \mathcal{P}, \mathcal{Q}^{\mathcal{F}} \rangle$, the entropy utility function is defined as the sum of entropies across all data items in $\mathcal{D}$, i.e.,*

$$EU(\mathcal{D}, \mathcal{F}) = -\sum_{\mathcal{O}_i \in \mathcal{O}} \mathcal{H}_i = -\sum_{\mathcal{O}_i \in \mathcal{O}} \sum_{v_i^k \in \mathcal{V}_i} p_i^k \log\ p_i^k$$

where $p_i^k \in \mathcal{P}$ is the probability that claim $v_i^k \in \mathcal{V}_i$ is `true`.

The entropy utility function measures the average uncertainty in the probabilities of claims; the closer the entropy utility is to 0, the higher is the effectiveness of fusion.

We present Maximum Expected Utility (denoted by `MEU`), a framework that integrates the entropy utility function with the concept of VPI. `MEU` uses $EU(\mathcal{D}, \mathcal{F})$ as the utility function in Definition 4.4.2 instead of $\mathcal{U}(\mathcal{D}, \mathcal{F}, \mathcal{T})$ to compute the expected entropy utility gain of action θ_i as:

$$\Delta EU_i = EU(\mathcal{D}, \mathcal{F}) - EU(\mathcal{D}, \mathcal{F} \mid v_i^k = true) p_i^k \tag{4.5}$$

`MEU` considers the one-step lookahead state of fusion after a *potential* action and identifies one that has the highest expected entropy utility gain, i.e.,

$$\theta_i = \underset{\theta_i \in \Theta}{\operatorname{argmax}} \quad \Delta EU_i \tag{4.6}$$

This kind of validation strategy is *myopic* in nature because we look only one step ahead each time we make a decision. It is possible that some action may not lead to the highest VPI at the current step but validating it can result in a higher VPI in subsequent validations. Sequential validations are challenging and often computationally expensive [19]; the present work focuses only on myopic strategies.

Example 4.4.3 *For the example in Table 4.1, we use Table 4.2 to compute $EU(\mathcal{D}, \mathcal{F})$ = 0.437. Considering $\mathcal{O}_1$ for validation, Table 4.3 shows the output of fusion when* `Howard` *is true and Table 4.4 shows the output when* `Spencer` *is true. (For ease of*

Algorithm 1: `MEU` Algorithm

1: **for** each unvalidated data item $\mathcal{O}_i$ **do**
2: **for** each claim $v_i^k \in \mathcal{V}_i$ **do**
3: Compute $EU(\mathcal{D}, \mathcal{F} \mid v_i^k = true)$
4: **end for**
5: Compute ΔEU_i as in Equation (4.5)
6: **end for**
7: Select the action with the maximum ΔEU_i

display, we represent the columns to be claims as they appear in Table 4.2, e.g., for $\mathcal{O}_1$*,* p^0 *represents the probability of claim* **`Howard`** *and* p^1 *the probability of* **`Spencer`***.)*

Table 4.3.: Probabilities when `Howard` is correct.

ID	p^0	p^1
$\mathcal{O}_1$	1	0
$\mathcal{O}_2$	0.082	0.918
$\mathcal{O}_3$	0.045	0.955
$\mathcal{O}_4$	1	
$\mathcal{O}_5$	0.004	0.996
$\mathcal{O}_6$	0.918	0.082

Table 4.4.: Probabilities when `Spencer` is correct.

ID	p^0	p^1
$\mathcal{O}_1$	0	1
$\mathcal{O}_2$	0.004	0.996
$\mathcal{O}_3$	1	0
$\mathcal{O}_4$	1	
$\mathcal{O}_5$	0.944	0.056
$\mathcal{O}_6$	0.996	0.004

Using Tables 4.3 and 4.4, **`MEU`** *computes* $EU(\mathcal{D}, \mathcal{F}|$**`Howard`** $= true) = 0.781$*, and* $EU(\mathcal{D}, \mathcal{F}|$**`Spencer`** $= true) = 0.262$*. The expected utility of* $\mathcal{O}_1 = 0(0.781) + 1(0.262) = 0.262$*.*

Table 4.5.: Expected utility of data items in Table 4.1.

ID	$\mathcal{O}_1$	$\mathcal{O}_2$	$\mathcal{O}_3$	$\mathcal{O}_4$	$\mathcal{O}_5$	$\mathcal{O}_6$
EU*	0.262	0.231	0.258	0.262	0.052	0.231

Table 4.5 shows the expected utility (EU) of all data items.* **`MEU`** *decides to validate* $\mathcal{O}_5$ *because its utility gain (*$(EU(\mathcal{D}, \mathcal{F}) - EU_5^*) = 0.385$*) is the highest among all items.*

In the absence of ground truth, maximum expected utility (MEU) [56] is considered to be the best alternative to ground truth utility. The main drawback of `MEU` is its lack of efficiency. To determine the next action, `MEU` re-runs fusion $\mathcal{F}$ on database $\mathcal{D}$ for each claim of every data item $o \in \mathcal{O}$. The time complexity of `MEU` is $\mathrm{O}(m\kappa t_{\mathcal{F}})$ where m is the number of unvalidated data items in $\mathcal{D}$, κ is the average number of unique claims per data item and $t_{\mathcal{F}}$ is the time needed to run $\mathcal{F}$ on one instance of data. A typical run of fusion iterates over all data items and all sources until convergence. This contributes to an $\mathrm{O}(m\kappa\mathcal{I}(m+n))$ complexity where $\mathcal{I}$ is the average number of iterations to convergence and n is the number of sources. With data items far outnumbering sources, the result is a complexity of $\mathrm{O}(m^2\kappa\mathcal{I})$. Concluding, `MEU` can tackle datasets a few hundred data items in size in a reasonable amount of time. Our objective is to be able to process datasets with at least a few thousands of data items.

Approximate-MEU

`MEU` describes a general decision-theoretic framework for the problem of ordering conflicts for user feedback in data fusion. However, the extreme computational cost of `MEU` makes it infeasible for large-scale datasets.

To this end, we present `Approx-MEU`, a method that leverages the structure of interactions between data items and sources to estimate the impact of a data item on other unvalidated data items. In the next step, it calculates the expected utility of each data item and determines the next action as the one with the maximum expected utility gain.

This approach is built on the intuition that an action would alter the probabilities of claims of the validated data item and its neighbors. The intuition is based on principles inherent in Bayesian network inference methods such as belief propagation [55], variational message passing [64] and incremental expectation-maximization [65]. These methods decompose the computation into local data item

calculations and pass them to other items via messages. In our problem, a validation is considered a local update of the probabilities of claims of a data item.

Consider data items $\mathcal{O}_i$ and $\mathcal{O}_j$. The goal of `Approx-MEU` is to estimate the probabilities of claims of $\mathcal{O}_j$ after $\mathcal{O}_i$ has been validated. This computation involves the following two steps: (i) measuring the change in probabilities of claims of the validated item $\mathcal{O}_i$, and (ii) estimating the change in probabilities of claims of $\mathcal{O}_j$ as a function of the change in probabilities of $\mathcal{O}_i$. We estimate the probabilities of claims of unvalidated data items by the method of linear approximation by differentials in the following steps.

Change in probabilities of claims of o_i

We assume an arbitrary claim $v_i^t \in \mathcal{V}_i$ to be `true`. Upon validating $\mathcal{O}_i$, the change in probability of v_i^t is: $\Delta p_i^t = (1 - p_i^t)$. This validation ensures that the remaining claims in $\mathcal{V}_i$ are `false`. The change in probability of $v_i^f \in \mathcal{V}_i \setminus \{v_i^t\}$ is : $\Delta p_i^f = (0 - p_i^f) = -p_i^f$.

Propagation of changes from $\mathcal{O}_i$ to $\mathcal{O}_j$

Data items $\mathcal{O}_i$ and $\mathcal{O}_j$ could be connected either through a source that votes for both of them or through a path consisting of alternating sources and items. As seen in the graph in Figure 4.2, $\mathcal{O}_1$ and $\mathcal{O}_2$ are connected through source $\mathcal{S}_3$ whereas $\mathcal{O}_2$ and $\mathcal{O}_4$ are connected via the $\langle \mathcal{O}_2, \mathcal{S}_3, \mathcal{O}_1, \mathcal{S}_4, \mathcal{O}_4 \rangle$ path. We present an analysis of both the cases:

1. **$\mathcal{O}_i$ and $\mathcal{O}_j$ have at least one common source.** We first examine how the probabilities of claims in $\mathcal{V}_i$ impacts the accuracies of sources that vote for both $\mathcal{O}_i$ and $\mathcal{O}_j$ (because change is propagated to o_j through these sources).

   Updates in source accuracies. The intuition behind the effect of changes in $\mathcal{O}_i$ to sources that vote on it is straightforward: we reward sources that

support the correct claim $v_i^t \in \mathcal{V}_i$ by trusting it more on information it provides on other data items. Similarly, our model penalizes sources that vote on some other claim v_i^f by discounting its information on other data items as well. From Equation (3.3), the change in accuracy $\mathcal{A}(s)$ of a source s is computed as:

$$\Delta\mathcal{A}(s) = \begin{cases} \Delta p_i^t / N(s) & \text{if s votes for } v_i^t \\ \Delta p_i^f / N(s) & \text{if s votes for } v_i^f \in V_i \setminus \{v_i^t\} \end{cases} \tag{4.7}$$

where $N(s)$ is the number of data items for which s votes.

Propagation of updates in sources to $\mathcal{O}_j$. Our next task is to measure further propagation of changes from the sources to $\mathcal{O}_j$. We compute the change in probability of claim $v_j^r \in \mathcal{V}_j$ attributable to the change in probabilities of claims of $\mathcal{O}_i$ by the method of approximation by differentials. This part of the analysis involves a short sequence of basic calculus over the formulae described in Section 4.3:

We rewrite Equation (3.2) as:

$$\frac{1}{p_j^r} = \sum_{v \in \mathcal{V}_j} \frac{\prod\limits_{s \in \mathcal{S}(v)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}}{\prod\limits_{s \in \mathcal{S}(v_j^r)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}} \tag{4.8}$$

and represent each summation term as a function f:

$$f(v_j^r, v) = \frac{\prod\limits_{s \in \mathcal{S}(v)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}}{\prod\limits_{s \in \mathcal{S}(v_j^r)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}} \tag{4.9}$$

Equation (4.8), therefore, simplifies to:

$$\frac{1}{p_j^r} = \sum_{v \in \mathcal{V}_j} f(v_j^r, v) \tag{4.10}$$

To compute the change in p_j^r, we estimate the approximate change in each $f(v_j^r, v)$ through a series of steps: take the logarithm of $f(v_j^r, v)$ and obtain the derivative with respect to $\mathcal{A}(s)$, thus presenting $\Delta f(v_j^r, v)$ as:

$$\frac{\Delta f(v_j^r, v)}{f(v_j^r, v)} = \sum_{s \in \mathcal{S}(v)} \frac{\Delta \mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))} - \sum_{s \in \mathcal{S}(v_j^r)} \frac{\Delta \mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))} \tag{4.11}$$

For each of the sources s that vote for $\mathcal{O}_j$, the term $\Delta \mathcal{A}(s)$ in Equation (4.11) takes a value as noted in Equation (4.7) depending on whether: (i) s supports v_i^t, (ii) s supports a claim other than v_i^t, or (iii) s does not provide any information on $\wr_i$. Clearly, if s belongs to the third category, it will not be affected by the validation of $\wr_i$.

We compute the change in probability of claim $v_j^r \in \mathcal{V}_j$ attributable to the change in probabilities of claims of o_i by taking the derivative of Equation (4.10):

$$\Delta p_j^r = -(p_j^r)^2 \sum_{v \in \mathcal{V}_j} \Delta f(v_j^r, v) \tag{4.12}$$

The change in probability of claim $v_j^r \in \mathcal{V}_j$ because of the validation of data item $\wr_i$ can, therefore, be expressed as:

$$\Delta p_j^r = -(p_j^r)^2 \sum_{v \in \mathcal{V}_j} \left(\frac{\prod_{s \in \mathcal{S}(v)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}}{\prod_{s \in \mathcal{S}(v_j^r)} \frac{(|\mathcal{V}_j| - 1)\mathcal{A}(s)}{1 - \mathcal{A}(s)}} \right) \cdot \left(\sum_{s \in \mathcal{S}(v)} \frac{\Delta \mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))} - \sum_{s \in \mathcal{S}(v_j^r)} \frac{\Delta \mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))} \right) \tag{4.13}$$

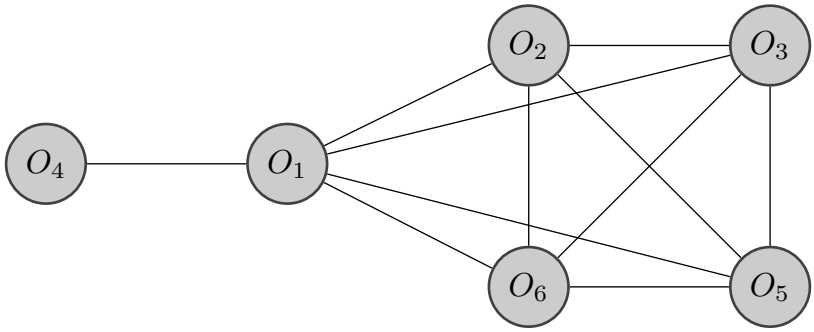


Figure 4.2.: Graph of data items in Table 4.1: An edge implies there is at least one source that provides information for the connecting data items.

With Δp_j^r, the approximate change in probability of claim v_j^r, the updated probability of claim v_j^r is computed as:

$$(p_j^r)' = p_j^r + \Delta p_j^r \tag{4.14}$$

2. $\mathcal{O}_i$ **and** $\mathcal{O}_j$ **have no source in common.** We know that any change in $\mathcal{O}_i$ reaches data items connected to it via at least one source, i.e., through data items that are one-hop away from $\mathcal{O}_i$. The changes in these data items then reach data items one-hop away from them, and so on.

Theorem 4.4.1 *The change in probabilities, Δp_j^r, of claim $v_j^r \in \mathcal{V}_j$ attributable to the change in probabilities, Δp_i^k, of claim $v_i^k \in \mathcal{V}_i$ is inversely proportional to the minimum number of data items a source votes for, raised to the power of d, the number of hops $\wr_j$ is away from $\wr_i$.*

$$\Delta p_j^r \propto \left(\frac{1}{N^d}\right) \Delta p_i^k$$

Proof. Consider data items $\mathcal{O}_i$ and $\mathcal{O}_j$ that are more than one hop away from each other, i.e., they are connected via an alternating path of sources and other

data items. In this section, we compute through a sequence of steps, the change in probabilities of $\mathcal{O}_j$ attributed to the validation of $\mathcal{O}_i$.

First, the change in probabilities of $\mathcal{O}_i$ are propagated to sources that provide claims about it. This changes the accuracies of sources: by boosting the accuracy of those that provide a `true` claim and decreasing the accuracy of those that provide an incorrect claim. From Equation (4.7), if source s provides claim v_i^l about data item $\mathcal{O}_i$, then the accuracy of the source changes as:

$$\Delta\mathcal{A}(s) = \frac{\Delta p_i^l}{N(s)}$$

Change in probabilities of $\mathcal{O}_j$. We represent Equation (3.2) for data item $\mathcal{O}_j$ as $p_j^r = q/t$ to obtain:

$$p_j^r t = q = \prod_{s \in \mathcal{S}(v_j^r)} \frac{(|\mathcal{V}_j - 1|)\mathcal{A}(s)}{1 - \mathcal{A}(s)} \tag{4.15}$$

We apply the logarithm function to both sides of Equation (4.15) to simplify the representation for further computation as:

$$\log q = \sum_{s \in \mathcal{S}(v_j^r)} \log \frac{(|\mathcal{V}_j - 1|)\mathcal{A}(s)}{1 - \mathcal{A}(s)} \tag{4.16}$$

Next, to compute the change in quantity q, we obtain the first derivative of the expressions in Equation (4.16) as:

$$\frac{dq}{q} = \sum_{s \in \mathcal{S}(v_j^r)} d\left(\log \frac{(|\mathcal{V}_j - 1|)\mathcal{A}(s)}{1 - \mathcal{A}(s)}\right) = \sum_{s \in \mathcal{S}(v_j^r)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))}$$

and express dq in a cleaner form as:

$$dq = q\left(\sum_{s \in \mathcal{S}(v_j^r)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1 - \mathcal{A}(s))}\right) \tag{4.17}$$

We express the change in probabilities of $\mathcal{O}_j$ by computing the first derivative of Equation (4.15):

$$p_j^r(dt) + (dp_j^r)t = dq \tag{4.18}$$

where t can be expressed as a sum of terms, t_k, similar to q for each $v_j^k \in \mathcal{V}_j$. Using Equation (4.17), Equation (4.18) can thus be rewritten as:

$$p_j^r \left(\sum_{v_j^k \in \mathcal{V}_j} t_k \sum_{s \in \mathcal{S}(v_j^k)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right) + (dp_j^r)t = q \left(\sum_{s \in \mathcal{S}(v_j^r)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right)$$

We now rearrange the terms appropriately and replace q/t by p_j^r, to express dp_j^r as:

$$dp_j^r = p_j^r \left(\sum_{s \in \mathcal{S}(v_j^r)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right) - p_j^r \left(\sum_{v_j^k \in \mathcal{V}_j} \frac{t_k}{t} \sum_{s \in \mathcal{S}(v_j^k)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right) \tag{4.19}$$

We are interested in analyzing the upper bound on dp_j to get an estimate of the maximum change that $\mathcal{O}_i$ would effect upon $\mathcal{O}_j$. We present a step-by-step conclusion of the same. It follows from Equation (4.19) that:

$$\begin{aligned}
|dp_j^r| &\leq p_j^r \left| \sum_{s \in \mathcal{S}(v_j^r)} \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right| \\
&\leq p_j^r \sum_{s \in \mathcal{S}(v_j^r)} \left| \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right| \\
&\leq p_j^r |\mathcal{S}(v_j^r)| \left| \frac{d\mathcal{A}(s)}{\mathcal{A}(s)(1-\mathcal{A}(s))} \right|_{max} \\
&\leq p_j^r |\mathcal{S}(v_j^r)| \left| \frac{dp_i^t}{N(s)\mathcal{A}(s)(1-\mathcal{A}(s))} \right|_{max} \\
&\leq p_j^r |\mathcal{S}(v_j^r)| \left| \frac{dp_i^t}{N'\mathcal{A}'(1-\mathcal{A}')} \right|_{max}
\end{aligned}$$

where $N' \leq N(s)$ is the least number of data items any source votes for and $\mathcal{A}'$ is the accuracy of a source that yields the minimum for function $\mathcal{A}(s)(1-\mathcal{A}(s))$.

Real datasets are often faced with the situation of few sources providing information about far too many data items. As a result, N' is usually more than half the number of items in the dataset. This, coupled with p_j, dp and $\mathcal{A}'(1-\mathcal{A}')$, contributes to the change in probabilities of a data item one-hop away being much less than the change in the probabilities of the validated data item.

For a data item, $\mathcal{O}_k$, two hops away from the validated node, following similar analysis, if $\mathcal{O}_k$ is reachable from $\mathcal{O}_i$ through $\mathcal{O}_j$, we reach the conclusion that:

$$\begin{aligned} |dp_k^l| &\leq \left(p_k^l |S(v_k^l)| \left| \frac{dp_j^r}{N'\mathcal{A}'(1-\mathcal{A}')} \right|_{max} \right) \\ &\leq \frac{dp_i^t}{N'^2} \left(\left| \frac{p_k^l p_j^r |\mathcal{S}(v_k^l)||\mathcal{S}(v_j^r)|}{(\mathcal{A}'(1-\mathcal{A}'))^2} \right|_{max} \right) \end{aligned}$$

We observe an exponential decay of the changes in probability distributions as we move away from the validated node. More specifically, the changes in probability distributions in the first hop are significantly higher than those from the second hop and so on. This is attributed to the sole reason that a typical source provides information about a large number of data items in the dataset. □

Real-world datasets typically consist of few sources providing claims about a large number of data items, and most of the data items are connected to each other. Through Theorem 4.4.1, we observe an exponential decay in the change in probabilities of claims as we move away from the validated data item.

Deciding the next action. Using Equation (4.14), `Approx-MEU` estimates first-order approximations of probabilities of claims of data items within one hop of $\mathcal{O}_i$

Algorithm 2: `Approx-MEU` Algorithm

```
for each unvalidated data item O_i do
    for each claim v_i^k ∈ V_i do
        Assume v_i^k is true
        for each unvalidated data item O_j ≠ O_i do
            for each claim v ∈ V_j do
                Estimate updated probabilities of v
            end for
        end for
        Compute entropy utility of updated probabilities
    end for
    Compute ΔEU_i as in Equation (4.21)
end for
Select next action according to Equation (4.6)
```

attributable to validating claim $v_i^k \in \mathcal{V}_i$. Entropy of a data item is then computed over the estimated probabilities of its claims, i.e.,

$$\mathcal{H}_i = -\sum_{v_i^k \in \mathcal{V}_i} (p_i^k)' \log (p_i^k)' \tag{4.20}$$

The expected utility gain of action θ_i is expressed as:

$$\Delta EU_i = EU(\mathcal{D}, \mathcal{F}) - \sum_{v_i^k \in \mathcal{V}_i} p_i^k \sum_{\mathcal{O}_j \in \mathcal{O}} \mathcal{H}_j \tag{4.21}$$

and the next action is determined as in Equation (4.6).

Example 4.4.4 *Consider $\mathcal{O}_3$ for validation in Table 4.1. Table 4.6 shows the estimated probabilities of claims obtained using Equation* (4.14) *when* `Docter` *is true and Table 4.7 shows the estimated probabilities when* `leFauve` *is correct.*

The expected utility of $\mathcal{O}_3 = 0.999(0.401) + 0.001(0) = 0.401$.

Table 4.8 shows the expected utility (EU) of all data items using the approximate probabilities of claims.* `Approx-MEU` *validates $\mathcal{O}_2$ because it has the highest expected utility gain.*

Table 4.6.: Probabilities when `Docter` is correct.

ID	p^0	p^1
$\mathcal{O}_1$	0	1
$\mathcal{O}_2$	0.019	0.981
$\mathcal{O}_3$	1	0
$\mathcal{O}_4$	1	
$\mathcal{O}_5$	0.931	0.069
$\mathcal{O}_6$	0.99	0.01

Table 4.7.: Probabilities when `leFauve` is correct.

ID	p^0	p^1
$\mathcal{O}_1$	0	1
$\mathcal{O}_2$	1	0
$\mathcal{O}_3$	0	1
$\mathcal{O}_4$	1	
$\mathcal{O}_5$	1	0
$\mathcal{O}_6$	1	0

Table 4.8.: Expected utility of data items in Table 4.1.

ID	$\mathcal{O}_1$	$\mathcal{O}_2$	$\mathcal{O}_3$	$\mathcal{O}_4$	$\mathcal{O}_5$	$\mathcal{O}_6$
EU*	0.437	0.184	0.401	0.437	0.235	0.313

Complexity. For each unvalidated data item, `Approx-MEU` assumes each of the claims to be `true` (one at a time) and estimates the first-order approximate probabilities of data items one-hop away from it. By eliminating the bottleneck iterative computation in `MEU`, `Approx-MEU` has a complexity of $O(m\kappa d)$ where m is the number of unvalidated data items, d is the average number of data items connected to a data item and k is the number of claims per item. In the worst case, $d = m$, when every data item is directly connected to every other data item through a source.

4.4.3 Further Optimizations

We now describe further optimizations to effectively scale up our ranking strategies. We briefly elaborate on bounding the number of data items to consider for validation and the effect of batch size on the performance of fusion.

1. **Shrinking the search space.** In datasets where all data items are connected to each other through one ore more sources, the complexity of `Approx-MEU` blows up to $O(\kappa m^2)$. To efficiently scale up the approximation formulae for

such dense data, we propose a hybrid approach that takes the best insights from `QBC`, `US` and `MEU`:

(a) Data items with high vote entropy (`QBC`) are the most disputed ones and, therefore, suitable for validation;

(b) Data items with low uncertainty over output of fusion are less suited to validation (similar to `US`);

(c) Among the high-entropy items, our goal (as in `MEU`) is to validate one with a greater expected utility gain.

We denote by $\texttt{Approx-MEU}_k$, the method that ranks unvalidated data items by their vote entropies and considers the top $k\%$ data items for the impact computation step. By tuning the value of k, we improve the complexity of `Approx-MEU` to $O(\kappa k^2)$.

2. **Batch of Actions.** The present work deals with one action at a time. However, if we have a budget of, say, twenty actions in total, one may argue that the most effective method should identify the set of best twenty actions that would result in the maximum expected utility. However, the task of finding an optimal set of twenty actions is not efficient: it is computationally expensive because the algorithm would need to consider all possible subsets of twenty actions. It is also not effective: by soliciting validation of twenty data items at once, we lose out on the opportunity to integrate earlier actions before deciding the next action. Our framework could be easily extended to solicit the top twenty actions that have the highest expected utility. While slashing run-time by reducing the number of iterations, this approach is expected to converge to ground truth slower than when we validate one data item at a time. (We present the results of this approach in Section 4.5.5).

4.4.4 Feedback Errors

So far, we have assumed access to accurate feedback from an expert. Real-world applications, however, are often faced with two major concerns: (1) Experts are expensive and often vary across domains; (2) Users (experts and otherwise) often give erroneous feedback.

To address these issues, in light of the recent advances in crowdsourcing [27], applications often turn to collecting feedback from a crowd of readily-available *workers*. Note that workers add a third dimension to the problem of data fusion previously governed by data items and sources; worker errors are independent of source (extraction) errors. Prior research that deal with non-experts [30, 31] jointly estimate user quality and true labels of data items, and query only the more trustworthy users in subsequent feedback rounds. The present work focuses on true labels of data items and does not address modeling the quality of users in a crowd setting. We assume that the crowd provides us either a single claim considered (partially) correct or probabilities representing correctness of claims of a data item.

Consider the case when a user (or, crowd) provides feedback for the data item that our ranking algorithm has determined to be the most beneficial for fusion. In the best case, all feedback is correct. To integrate erroneous input into our framework, we translate imperfect feedback to correctness of claims and leverage this prior knowledge, along with the observations, to estimate the correctness of claims for rest of the data items.

1. **Feedback confidence.** In some cases, users express confidence in their feedback, e.g., '80% certain that v_i^k is the correct claim for data item $\mathcal{O}_i$'. We incorporate this knowledge into our model by assigning the confidence to correctness of the claim, i.e., $p_i^k = 0.8$ and the rest as 0.

2. **Incorrect feedback.** This case pertains to quality of the user (or, crowd) providing feedback. In case of a crowd, we assume that the crowdsourcing system processes conflicting answers from workers and provides the most accurate la-

bel.Knowing the user's (or, crowd's) error-rate ϵ, e.g., on 4 out of 6 instances, the feedback is incorrect, we compute the expected utility gain over correct and incorrect feedback. If the provided claim v_i^k is correct, we set $p_i^k = 1$ and the rest as 0. Otherwise, we set $p_i^k = 0$ and set a uniform probability distribution for rest of the claims, i.e., $p_i^r = 1/|claims|$ whenever $r \neq k$.

3. **Conflicting feedback.** We also consider the case when, instead of providing a single correct claim for a data item, the crowd simply presents the answers from different workers. For example, say for data item o_i having three claims (v_i^A, v_i^B, v_i^C), 6 workers agree on v_i^A being correct, 3 agree on v_i^B and 1 says claim v_i^C is correct. We summarize this information in the form of probabilities either by counting or some other mechanism, i.e., we conclude that $(p_i^A, p_i^B, p_i^C) = (0.6, 0.3, 0.1)$ and feed this knowledge to the data fusion model.

4.5 Experimental Evaluation

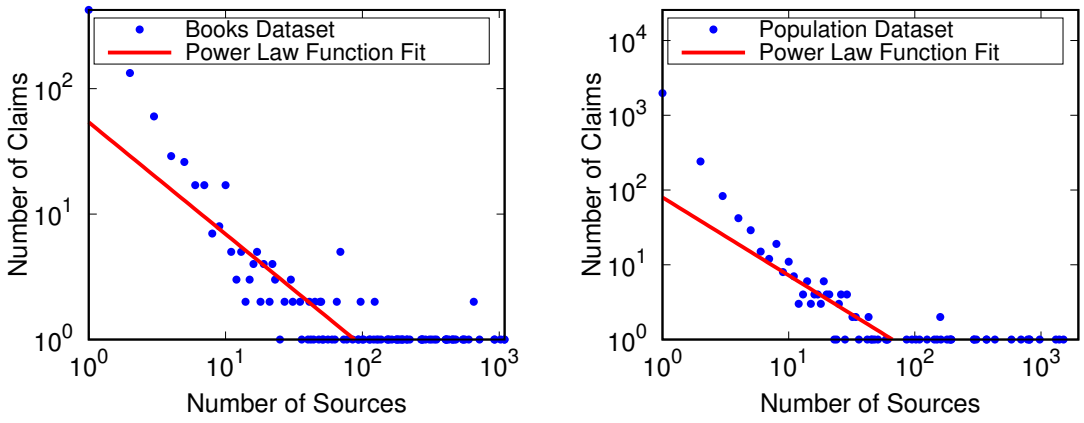
This section presents an empirical evaluation of the proposed solutions on two real-world datasets. Our objectives are: (1) To assess the effect of acquiring feedback in improving the performance of data fusion, (2) To evaluate the proposed ranking algorithms, and (3) To analyze the trade-offs between effectiveness and efficiency offered by the various approaches. Moreover, we study the behavior of the methods on data with different characteristics and with respect to parameters such as batch size and erroneous feedback.

4.5.1 Datasets

To validate the proposed methods, we conducted experiments on the following real-world datasets (Table 4.9):

Table 4.9.: Statistics of real-world datasets.

	Books	FlightsDay	Population	Flights
Items	1,263	5,836	40,696	121,567
Sources	894	38	2,545	38
Claims	24,303	80,452	46,734	1,931,701



(a) Books. (b) Population.

Figure 4.3.: Long-tail characteristics in real data. Most sources provide information on a small fraction of and few provide data about a large number of items.

Books: We used the books dataset from [8] that contains a listing of computer science books and their authors as provided by bookstores registered at *abebooks.com*.

Flights: We used the flights dataset from [1] that contains status information for flights over an entire month as reported by 38 sources. A data item is an attribute (such as scheduled arrival time) of a particular flight. We permit slightly different reported values (to a maximum difference of 10 minutes) in flight times that might have arisen due to slight lag in updates, or error in estimating times.

FlightsDay: We used a one-day snapshot of Flights (for the day of 12/1/2011); this dataset is representative of the Flights dataset that spans over a month's time.

Population: We used the city population dataset from [66] that contains Wikipedia edit histories of the populations of certain cities in a given year. To account for unreasonably large values and to have a source provide a single claim per data item, we adopt preprocessing steps similar to [13].

For simplification, we consider only those flight and population data items that have up to two contesting values. In Books, we consider the top two author sets per book.
Data Characteristics: We notice that our real-world data-sets exhibit interesting properties: (i) Most of the data items in the flights datasets are connected to each other because the small number of sources provide information on almost all data items, (ii) Both Books and Population exhibit long-tail characteristics (Figure 4.3), i.e., the distribution of number of claims per source follows the power law phenomenon where more than 90% sources provide information on fewer than 4% data items. Such varied characteristics of data allow us to evaluate our approaches in different scenarios.
Feedback Simulation. We simulated user feedback for data items by providing feedback as determined by the ground truth. We used the silver standard provided in [8] as the ground truth for Books. For Flights, we considered data provided by each of the carrier websites, *American Airlines, United Airlines* and *Continental*, to be the ground truth. We manually identified the true claim for data items in Populationthat have more than one claim.

4.5.2 Competing Methods

We compared the following ranking approaches:

1. **QBC** (Section 4.4.1): This item-level ranking method uses the distribution of claims to rank data items.

2. **US** (Section 4.4.1): An item-level ranking method that uses fusion probabilities to rank data items.

3. **Greedy Upper Bound (GUB)** (Section 4.4.2): Assuming that ground truth is known, this method selects an action that results in the highest ground truth utility gain according to Definition 4.4.2.

4. **MEU** (Section 4.4.2): In the absence of ground truth, this method selects the action that has the maximum expected utility gain.

5. **Approx-MEU** (Section 4.4.2): A decision-theoretic approach that ranks data items according to their approximate impact on other unvalidated data items.

6. **Random**: This naïve method selects an action at random; all data items are considered equally beneficial.

We implemented all the algorithms in Java, and ran experiments on a Macbook Pro with 8GB RAM, 2.7 GHz Intel Core i5 processor, and OSX El Capitan 10.11.5.

Performance Metrics

Effectiveness: To evaluate the effectiveness of the proposed methods, we conducted a sequential validation of all data items having conflicting claims (in the order determined by a given method) and obtained an assignment of `true` and `false` claims using a truth function $\mathcal{T}$. We report the following metrics on the results:

1. **Distance to ground truth**: We report the improvement in output of data fusion after an action as the reduction in distance of probabilities of claims to ground truth defined as:

$$\texttt{distance\_to\_ground\_truth} = \sum_{i=1}^{|\mathcal{O}|} \sum_{v_i^k \in \mathcal{V}_i} \frac{\delta(\mathcal{T}(v_i^k))(1 - p_i^k)}{|\mathcal{O}|}$$

where $\delta(\mathcal{T}(v_i^k)) = 1$, if $v_i^k = \texttt{true}$. Intuitively, `distance_to_ground_truth` can be seen as the average error of data fusion. The smaller the `distance_to_ground_truth`, the more accurate is the output of fusion.

2. **Uncertainty**: We report the reduction in uncertainty over output of data fusion defined as the entropy across all data items:

$$\texttt{uncertainty} = -\sum_{i=1}^{|\mathcal{O}|}\sum_{k=1}^{|\mathcal{V}_i|} -p_i^k \log(p_i^k)$$

where p_i^k is the probability that claim $v_i^k \in \mathcal{V}_i$ is correct. A higher value of `uncertainty` indicates less confidence in the output of data fusion.

Once a data item is validated, we retain the validation result and therefore, observe a cumulative gain of all validations. Figure 4.4 presents example curves for the effectiveness metrics that start at 0 (when no data item is validated) and gradually approach -100% (when all items are validated). A plot closer to the axes indicates a better method.

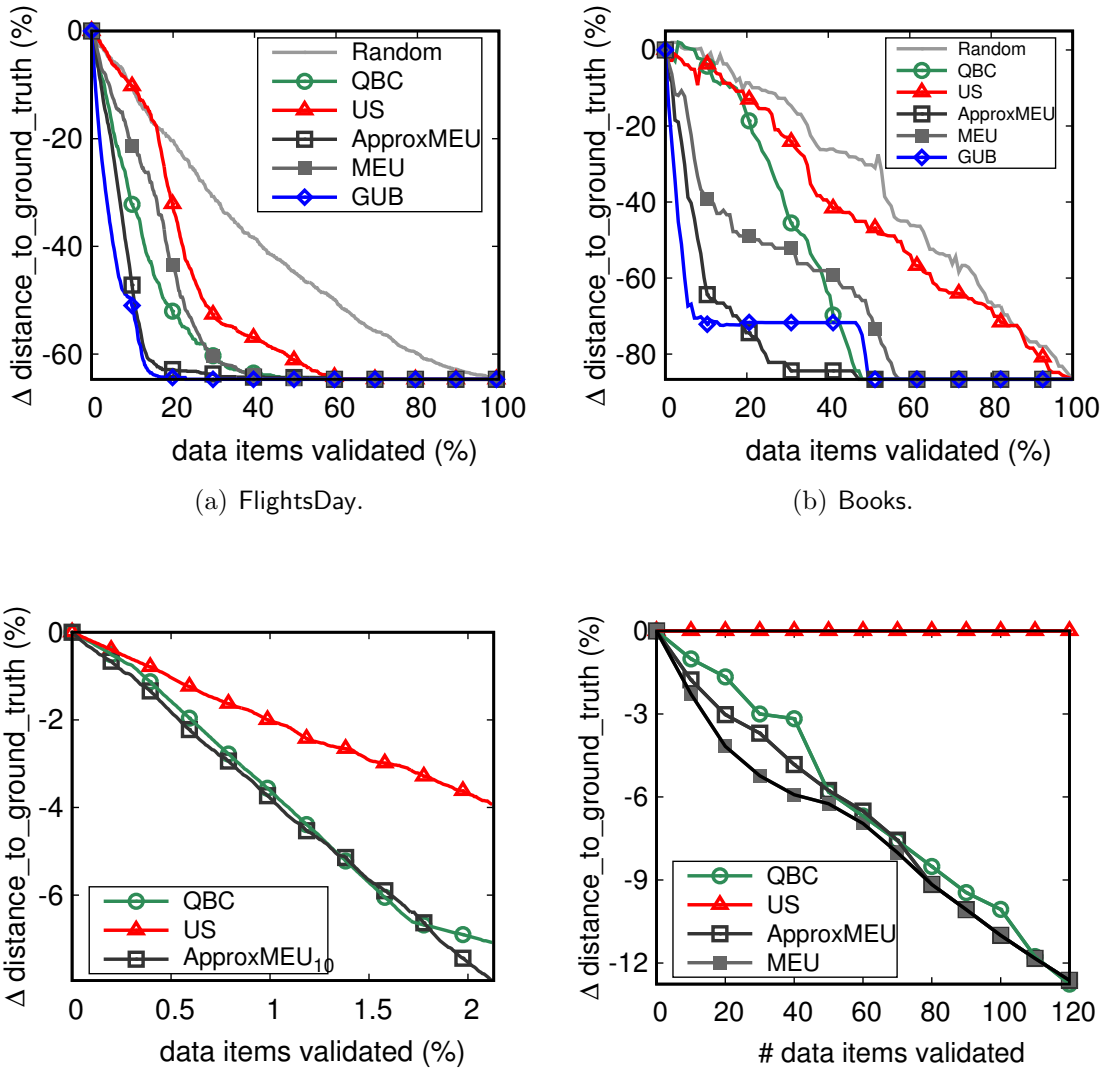
Efficiency: To evaluate the efficiency of an approach, we report the average time it takes to determine the next action.

4.5.3 Evaluation of Ranking Strategies

In this section, we evaluate effectiveness of the item-level ranking strategies (Section 4.4.1) and the decision-theoretic framework (Section 4.4.2) in improving the performance of data fusion. Our best-case decision-theoretic mechanism involves a utility function based on the ground truth.

Effectiveness. Assuming the availability of a ground truth utility function, we demonstrate in Figure 4.4, the gradual improvement in distance to ground truth for increasing number of validated data items for all the validation methods.

As illustrated in Figure 4.4, all the approaches improve the distance of the output of fusion to ground truth, albeit by various degrees. `Random` almost linearly decreases



(a) FlightsDay.

(b) Books.

(c) Flights.

(d) Population.

Figure 4.4.: Effectiveness of different ranking strategies measured as the reduction in distance_to_ground_truth against number of items validated.

the distance to ground truth indicating that only the number of actions determines its effectiveness. `QBC` and `US`, through guided selection of data items, converge to ground

truth faster than Random; QBC consistently performs better than US. Specifically, in the long-tail datasets, because the adopted data fusion model assigns either very high or low probabilities to claims, most of the data items have very low uncertainties and therefore, US is unable to distinguish them. On the other hand, true quality of the sources in dense datasets is aptly reflected in their accuracies and correctness of claims. The data items selected by US are also ones that the data fusion model has not been able to resolve, indicating that these items are probably not well-connected to other data items. Validating these data items, therefore, does not have much impact on the accuracy of other items.

We notice that MEU is consistently superior to US, indicating that we benefit from a method that aims at reducing uncertainty across all data items instead of resolving a single uncertain data item. We also observe that MEU and QBC have contrasting performances in long-tail datasets and in dense data. This behavior is attributable to the structure of the datasets: each source in dense data (e.g., FlightsDay) provides information on a large number of items. The change in accuracy of a source upon a validation is, therefore, not large enough to propagate to other items. It is useful in such cases to validate items with higher vote entropies first.

Not surprisingly, GUB has the steepest initial curve among all the methods. GUB takes advantage of the ground truth information and, therefore, theoretically, has the best performance in reporting the distance_to_ground_truth.

Interestingly, we observe that after GUB, Approx-MEU has the best performance in FlightsDay and Books — the method estimates expected correctness of claims from a validation and aims to reduce uncertainties in the estimates across all data items, thus outperforming both the item-level ranking algorithms (QBC, US). However, in Population, the room between QBC, Approx-MEU and MEU is not very large. This similarity in performance of the methods is due to sparsity of the data ($|\mathcal{V}|/(|\mathcal{O}| \times |\mathcal{S}|) = 0.04\%$) which results in a very small portion of data items ($\sim 2.5\%$) having more than one claim. The idea then is to identify among these items, those that are

Table 4.10.: Time taken to determine the next action.

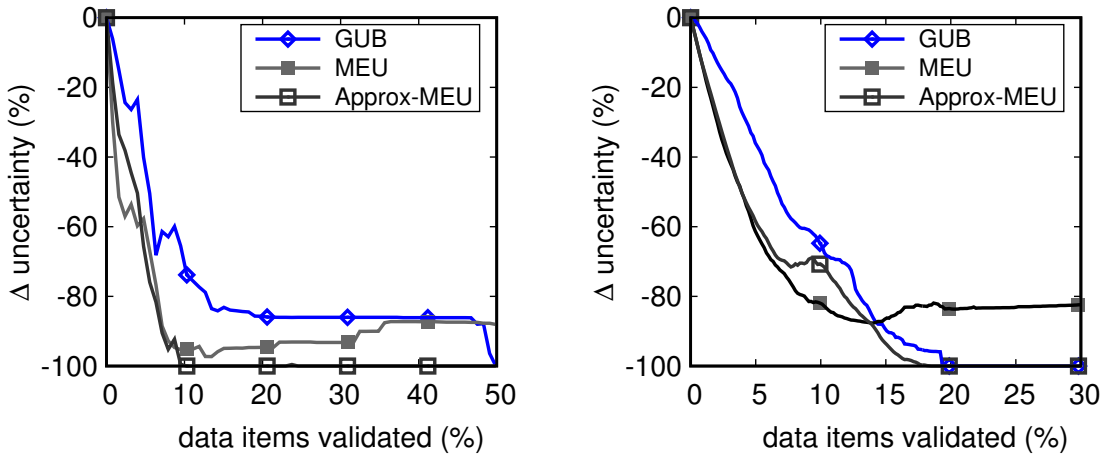
time (sec)	QBC	US	MEU	ApproxMEU
Books	0.01	0.001	11.73	0.231
FlightsDay	0.045	0.002	90.00	4.401
Population	0.14	0.011	> 5 min	9.728

time (sec)	QBC	US	ApproxMEU_5	ApproxMEU_{10}
Flights	7	4	146	348

the most beneficial to others. Both Approx-MEU and MEU, therefore, have an advantage over QBC that does not take into account the holistic impact of an action.

To scale up Approx-MEU to large dense data (Flights), we set $k = 10$ in Approx-MEU_k. With as few as a tenth of the total number of data items considered for validation, Approx-MEU_k is seen to achieve higher quality fusion results than QBC and has significantly better performance than US. Although Approx-MEU and QBC are comparable in early validations, Approx-MEU displays a notably rapid rate of convergence to ground truth as more items are validated. The results further confirms effectiveness of the decision-theoretic framework over item-level ranking methods. However, considering both effectiveness and efficiency, in such large dense data, QBC might be a better choice than Approx-MEU_k if $k << |\mathcal{O}|$.

Efficiency. In Table 4.10, we report the average time taken by the methods for one validation (recall that we cannot compare GUB on real data, and we cannot scale MEU to large dense data). The item-level ranking algorithms (QBC, US) are observed to be significantly faster than the decision-theoretic framework (MEU, Approx-MEU); QBC makes a single pass over all data items and US ranks them after each validation whereas MEU and Approx-MEU fuse data with each claim of an item separately considered as prior knowledge. The high numbers for MEU motivate the need for a cheaper (but effective) alternative. Approx-MEU, while still slower than QBC and US, is faster than MEU by almost two orders of magnitude. Our goal for efficiency is to provide an interactive validation time for users of a data fusion system. We conclude that MEU



(a) Books. (b) FlightsDay.

Figure 4.5.: Comparing methods based on entropy utility function (`MEU`, `Approx-MEU`) against ground-truth-based method (`GUB`).

cannot be used for datasets typical to data fusion. From a theoretical standpoint, the time for `MEU` is based on time for the fusion system since it runs the system for all claims of each data item.

Practicability of Entropy Utility. The strength of `GUB` lies in its access to a ground truth utility function. However, real datasets provide the ground truth for a small subset of data items. In this experiment, we assess the feasibility of entropy utility function as a substitute to the ground truth utility function by comparing the performance of entropy-utility-based methods (`MEU, Approx-MEU`) against that of the ground-truth-based method (`GUB`).

As shown in Figure 4.5, `MEU` and `Approx-MEU` achieve a greater reduction in uncertainty than `GUB`. This mechanism comes at the price of `MEU` converging to ground truth at a rate slower than `GUB` (Figures 4.4(a) and 4.4(b)). Interestingly, the rate of convergence to ground truth of `Approx-MEU` is better than `MEU` and is almost identical to `GUB`. Practically, however, `GUB` is infeasible; `MEU` and `Approx-MEU` are our best viable alternatives.

Relation between performance metrics In this experiment,we notice that the plots representing the distance to ground truth and those representing the reduction in uncertainty follow the same trend, i.e., as the distance to ground truth decreases, the uncertainty is also reduced. Moreover, the rate of reduction in these two metrics appears to be comparable for `GUB` and `MEU`. Theoretically, we can explain this behavior in one direction: as the database gets closer to ground truth, the data fusion system becomes more certain in its predictions. Therefore, the uncertainty of the database is expected to decrease. On the other hand, as uncertainty decreases, there is no guarantee that the fusion system would fare better in predicting correct claims; it simply might be more certain in wrong predictions.

To better understand the relation between the two metrics, we conducted an experimental study of the metrics for the fundamental methods, `GUB` and `MEU` (since these are our gold standards), on synthetic datasets generated using a number of parameters.

Synthetic Data Generation. Our objective in generating synthetic data is to replicate dense real-world data with $|\mathcal{O}| >> |\mathcal{S}|$ (typical datasets for data fusion systems, e.g., see [1]). We model most of the sources to be of good quality with few very good and few poor sources. Source accuracies, $\mathcal{A}(\mathcal{S}_j)$, can therefore be assumed to follow a normal distribution: $\mathcal{A}(\mathcal{S}_j) \sim N(a_{mean}, a_{sd})$ where a_{mean} is the average accuracy and a_{sd} is the standard deviation of the source accuracies. Density of the data, i.e. the probability that a source votes for a data item, is specified by d. The default values for the parameters, $a_{mean} = 0.8$, $a_{sd} = 0.1$ and $d = 0.4$, correspond to the characteristics of real datasets. Source S_j provides a claim for data item o_i with probability d and the claim is correct with a probability $\mathcal{A}(\mathcal{S}_j)$.

Observation. As seen in Figure 4.6, we observe empirically that the distance to ground truth and uncertainty are strongly correlated. This study is further supported by the Pearson's correlation coefficient, $\rho = 0.86$. For FlightsDay, $\rho = 0.71$ and for Books, $\rho = 0.72$, indicating a moderately positive correlation. Specifically, uncertainty in the fusion predictions and their distance to ground truth go hand in hand. This

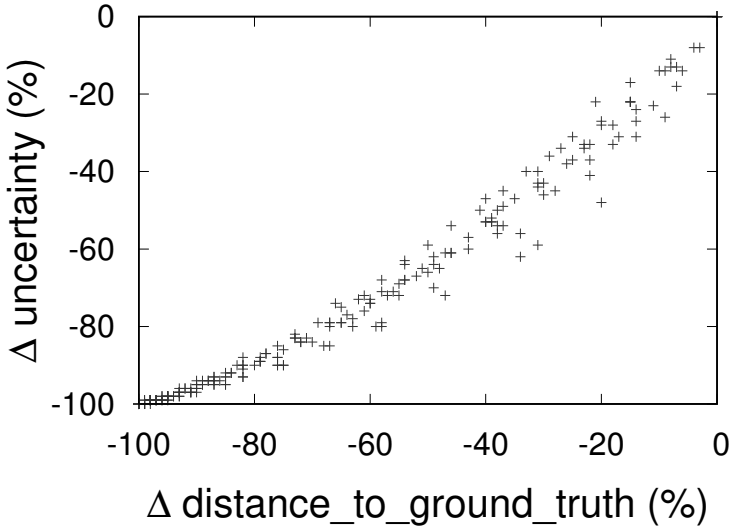


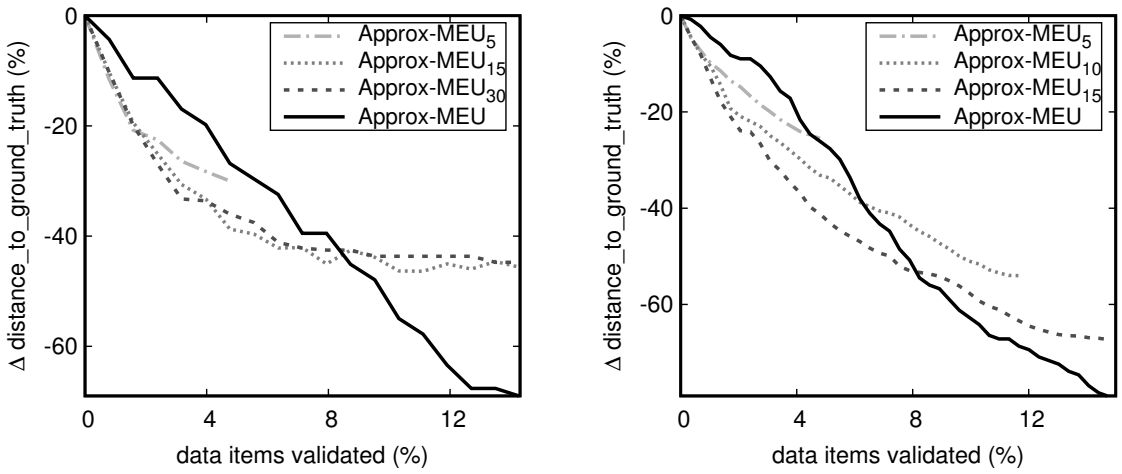
Figure 4.6.: Scatter plot showing the relation between different performance metrics.

additionally confirms the suitability of entropy utility as an alternative to ground truth utility function.

Takeaways. (1) Active feedback improves data fusion better than a passive approach (`Random`). (2) The decision-theoretic framework (`MEU`, `Approx-MEU`) exhibits effectiveness superior to that of the item-level ranking approaches (`QBC, US`); in practice, however, the latter are significantly faster methods. (3) The entropy utility function is a suitable alternative to the ground-truth utility function. (4) `MEU` has an extreme computation cost and cannot be used for validation on large datasets. (5) `Approx-MEU` is a cheaper, and also effective, substitute to `MEU`.

4.5.4 Exploring Approx-MEU

As mentioned in Section 4.4.2, in the worst case, `Approx-MEU` mandates an all-pairs computation of the impact of data items on each other — still expensive in datasets where all data items are connected to each other. In Section 4.4.3, we discussed optimizations to reduce the computation cost by shrinking the search space for the



(a) Books. (b) FlightsDay.

Figure 4.7.: Hybrid approach combining `QBC` and `Approx-MEU`. Figures depict the effect of expanding the set of candidates for validation in `Approx-MEU`.

impact computation step; we now explore the effect of this approach i.e., the role of k in `Approx-MEU`$_k$, on the improvement in data fusion.

Effectiveness. Figure 4.7 demonstrates the various degrees of improvement offered by `Approx-MEU`$_k$ as k varies. Subscript k denotes the fraction of all data items considered for impact computation. When $k = 5$, we consider only the top 5% data items ranked first according to their vote entropies and then, in the order of their entropies over probabilities of claims. We compute only the impact of these 5% data items on each other; evidently, the line ends when 5% of all data items are validated. We observe that as k increases, more data items are considered in the impact computation step and the system converges to ground truth faster. `Approx-MEU`, while less effective in the beginning, gradually surpasses the improvement in fusion achieved by the `Approx-MEU`$_k$ methods. The plots indicate that for early validations (less than 8% of items validated), choosing as small a value as $k = 30$ (Books) or $k = 15$(FlightsDa y) results in better conflict resolution than `Approx-MEU`; by tuning k, we can effectively scale up the decision-theoretic framework with estimated probabilities to large datasets.

Table 4.11.: Time taken (in seconds) by QBC, US and Approx-MEU$_k$ with different values of k.

time(sec)	Books	FlightsDay	Flights
QBC	0.08	0.07	6.0
US	0.09	0.12	1.8
Approx-MEU$_5$	0.04	0.23	156
Approx-MEU$_{10}$	0.09	0.73	323
Approx-MEU$_{15}$	0.15	0.98	475

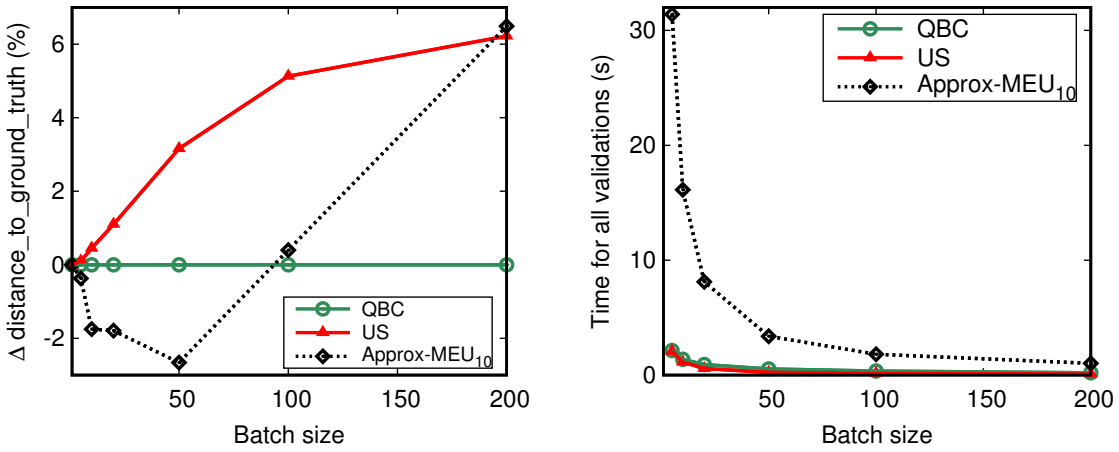
Efficiency. We report in Table 4.11 the time taken for one validation on the three datasets by QBC, US and Approx-MEU$_k$ with different values of k. As expected, with an increase in k, as more data items are considered for impact computation, Approx-MEU$_k$ takes longer to determine the next action. However, for the large-scale Flights data, Approx-MEU has a significantly rapid convergence to ground truth than QBC and US in slightly more than 5 minutes.

Takeaways: (1) By limiting the fraction of data items for the impact computation step, Approx-MEU can be efficiently scaled up to large datasets. (2) Different values of k offer trade-offs between effectiveness and efficiency. Specifically, the smaller the value of k, the faster it takes to determine the next action although a method with a higher k rapidly converges to ground truth.

4.5.5 Effect of Batch Size

Based on our intuitions about batch size (Section 4.4.2), we now study the effect of validating multiple data items simultaneously on the performance of our methods.

Effectiveness. As shown in Figure 4.8(a), performance of QBC is not affected by batch size because by selecting data items based on their vote entropies, at the end of 200 actions, the set of validated data items remains unchanged.



(a) Effectiveness. (b) Efficiency.

Figure 4.8.: Effect of batch size on effectiveness of the methods and time taken to validate 200 claims from FlightsDay.

With an increase in the batch size, the distance to ground truth steadily increases for `US` because by validating multiple data items at once, it loses the opportunity to adaptively integrate the acquired feedback.

`Approx-MEU` displays an interesting behavior: the method converges to ground truth faster with an initial increase in batch size, and after `batchSize`= 50, its performance worsens. The initial improvement is because with smaller batches, the algorithm selects data items having high entropy (e.g., entropy > 0.67); as the batch size increases, the algorithm selects data items with a mix of high and medium entropies (e.g. entropy > 0.6). By not ordering data items with medium entropies correctly, the performance of the method deteriorates with an increase in batch size.

Efficiency. We observe in Figure 4.8(b) that the time taken by `QBC` and `US`, after sorting, is effectively the time taken to fuse the data. As more data items are validated together, the fusion system reaches a steady state faster and the methods have almost flat gain in the time for all validations. Going from a `batchSize` of 1 to 200, the runtime of `Approx-MEU`, however, reduces by more than one order of magnitude. Specifically, for FlightsDay, we observe that a `batchSize`= 50 achieves the best im-

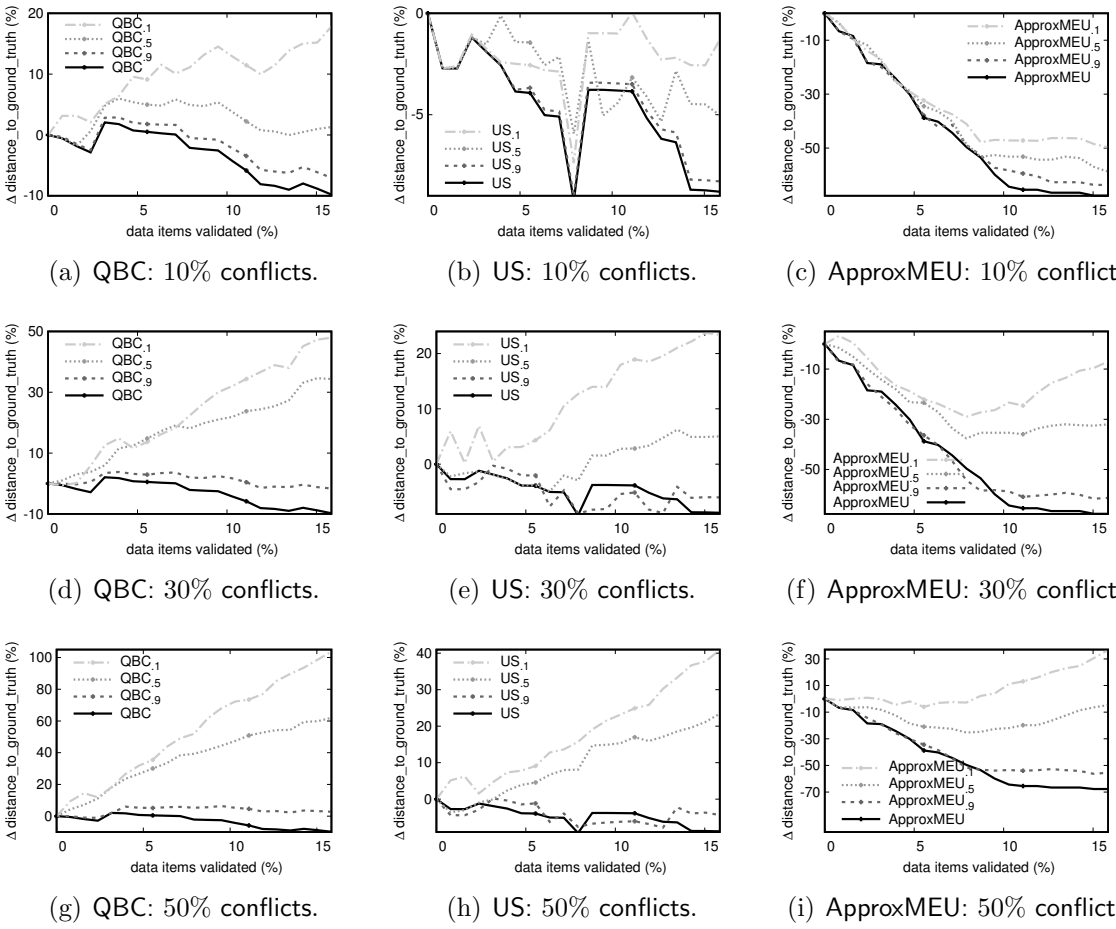
provement in fusion in about one-sixth the time taken by validating individual data items.

Takeaways: Increasing the batch size: (1) has no effect on `QBC` while it typically degrades performance of `US` and `Approx-MEU` (although the latter shows improvement with smaller increase in batch size), and (2) drastically reduces the time taken for validations by `ApproxMEU`.

4.5.6 Feedback Errors

To evaluate our ranking approaches in the presence of imperfect feedback, we perform experiments that study effectiveness of the methods in different error scenarios as discussed in Section 4.4.4. We perform experiments on Books and FlightsDay because results were the most promising for these datasets in the previous experiments. Due to space constraints, we present only few of the experiment results.

Conflicting feedback. In this experiment, we assume access to feedback from a crowd of workers who provide correctness of all claims instead of providing a single correct label. We consolidate conflicts of the crowd by varying (1) the fraction of data items that it disagrees on (i.e., the crowd provides correctness of all claims of say, 5% data items), and (2) their consensus on the correct claim for a data item (i.e., 70% probability that the true claim is indeed correct). We vary the first parameter from $10\% - 50\%$ and the second from $10\% - 90\%$ and report the results of this experiment in Figure 4.9. Lines in the plots compare a method when correctness of the correct claim varies from 0.9 to 0.1. As expected, as the crowd varies its consensus on the correct claim from 90% to 10%, the performance of all the methods consistently deteriorates. `QBC` and `US` start falling apart as the crowd's consensus degrades. The methods with 90% consensus, however, exhibit comparable performance to their no-error counterparts even when the fraction of data items with conflicting feedback increases. On the other hand, `Approx-MEU` demonstrates substantial improvement in fusion even when the consensus goes to 50% on 30% of all data items. It only starts



(a) QBC: 10% conflicts. (b) US: 10% conflicts. (c) ApproxMEU: 10% conflicts.

(d) QBC: 30% conflicts. (e) US: 30% conflicts. (f) ApproxMEU: 30% conflicts.

(g) QBC: 50% conflicts. (h) US: 50% conflicts. (i) ApproxMEU: 50% conflicts.

Figure 4.9.: Conflicting feedback (Books). Each row compares methods when x% of items have conflicting feedback.

to worsen when the crowd assigns really low probability to the correct claim for 50% of all data items.

Feedback Confidence. We simulate the confidence in feedback as a probability attached to it. This could also be likened to *worker* (or, crowd) quality, e.g., there is only 80% probability that any feedback provided by `Worker A` on a data item is correct. We assume the confidence to be varying from $80\% - 100\%$ and report the results of this experiment in Figure 4.10. We notice that performance of the methods consistently deteriorates as confidence decreases from 100% to 80%. While with even 90% conviction in feedback, `QBC` and `US` no longer improve fusion on Books,

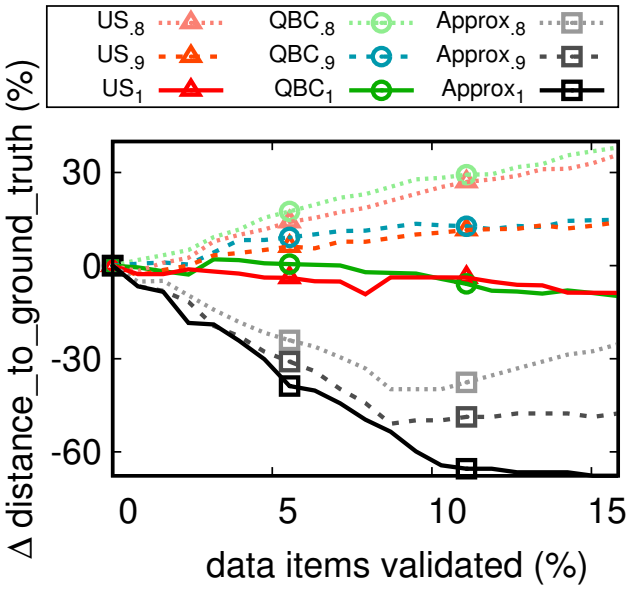


Figure 4.10.: Feedback confidence (Books). Subscript denotes user confidence (or, worker quality).

`Approx-MEU` is the most resilient to such feedback errors. Even at 80% confidence, `Approx-MEU` adaptively integrates erroneous input and continues to improve fusion in initial validations (although with diminished power) before tapering off and worsening after $\sim$ 8% of the data items have been validated. `Approx-MEU`$_{.9}$ almost levels out after 10% items are validated, and with `Approx-MEU`$_{.8}$, soliciting feedback after 5% validations does not boost fusion. The net improvement with `Approx-MEU`$_{.8}$ after 15% of items are validated, however, is comparable to that achieved in `QBC` and `US` without any feedback errors.

Incorrect Feedback. We assume the hypothetical case when we have an ineffective user that (either knowingly or unknowingly) provides incorrect answers. We further consider the user to be wrong on $0\% - 30\%$ of data items and report the results in Figure 4.11. With slight abuse of notation, the subscript with a method is used to represent the fraction of data items that the user is wrong about. We notice that as the fraction of erroneous data items increases, the methods essentially worsen fusion. However, even with 10% of data items judged incorrectly by the user, `QBC`

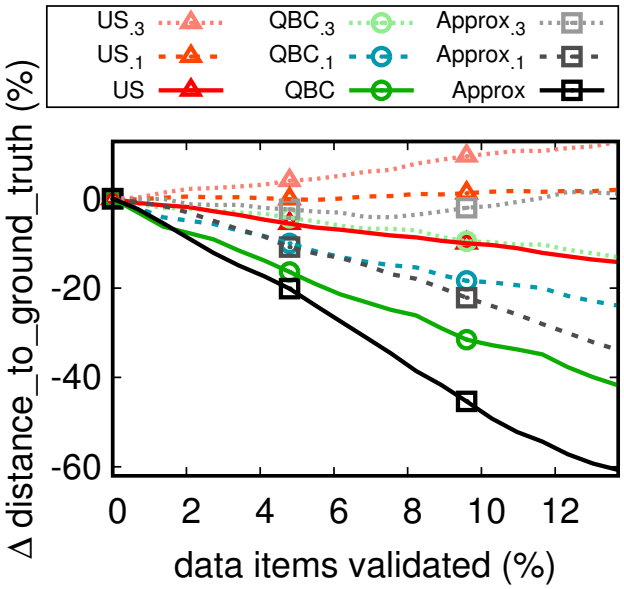


Figure 4.11.: Incorrect Feedback (FlightsDay). Subscript denotes fraction of items with incorrect feedback.

and `approx-MEU` exhibit better performance than `US` without incorrect feedback. This demonstrates that on dense data, identifying items that have high entropy is more beneficial and more resilient to feedback errors than selecting items with `US`.

Takeaways: (1) Among all the approaches, `Approx-MEU` is most robust in the presence of feedback errors. (2) `Approx-MEU` continues to improve fusion even when the feedback is close to incorrect for a small fraction of data items. (3) On dense data, `QBC` is resilient to completely incorrect feedback on a small fraction of all data items.

4.6 Summary

In this chapter, we proposed a pay-as-you-go approach for effectively soliciting feedback from users to resolve conflicts and improve the performance of existing data fusion techqniques.

To judiciously utilize the user, we proposed generating effective ordering of data items for validation. We presented algorithms that assess data items individually

by considering their local characteristics, and also proposed a novel decision-theoretic framework that evaluates data items holistically by their ability to improve the performance of fusion. We further devised approximation formulae to scale up the decision-theoretic framework to large-scale datasets, and also explored scenarios in the presence of imperfect feedback.

The main highlights of the proposed approaches are that they do not assume any domain-knowledge constraints or access to ground truth. Furthermore, in the presence of noisy feedback from a crowd of workers, any of the existing crowdsourcing approaches can be used to obtain the most accurate label for data items and plugged into the user feedback framework.

Our experimental evaluation on real-world datasets confirmed that guided feedback rapidly increases the effectiveness of data fusion. The proposed methods exhibited different behavior for data with different characteristics, and also offered trade-off between effectiveness and efficiency, and the amount of feedback acquired.

Results from this chapter were published in [67, 68].

5 LEVERAGING DATA RELATIONSHIPS TO RESOLVE CONFLICTS

In this chapter, 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.

This chapter is organized as follows: in Section 5.1, we present a motivating example for the problem of integrating entity-relationships among claims of data items, describe the solution overview and outline the summary of our contributions. We formally present our problem in Section 4.2. We explore entity-relationships among claims, describe the relationship model and outline steps to pre-process it in Section 5.3.In Section 5.4, we discuss algorithms to integrate the relationship model with data fusion models and solutions to determine correct claims that are consistent with each other. We conduct an experimental evaluation of our approach on real-world data in Section 5.5, and finally summarize the chapter in Section 5.6.

5.1 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 [5] to resolve data discrepancies from disparate sources and present high-quality integrated data to users. Recently, [8, 12] studied the problem of dependence among sources in the context of data fusion whereas [37,38] studied the interdependence among data items in the fusion of spatial and temporal data. However, the space of existing associations

Table 5.1.: Table shows five websites providing information about music genres of four songs. Correct claims are marked with a (*).

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

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 [69]. The rich space of relationships among claims of data items makes it challenging to distinguish correct from incorrect information as illustrated next.

Example 5.1.1 *Consider an example of information provided by five websites on music genres of certain songs (Table 5.1). Sources provide conflicting information for the same data item, e.g.,* $\mathcal{S}_2$ *provides* Christmas *as the genre for song* Silent Night *whereas* $\mathcal{S}_3$ *claims it to be* Pop *and* $\mathcal{S}_4$ *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 [42], map databases) and general purpose knowledge bases [40, 41]. (The relationships among claims for this example have been obtained from DBpedia [41] and AllMusic* [1]*, the popular online music guide.)*

[1] www.allmusic.com

Single-truth data fusion models [3, 8] 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 [12, 16] 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.

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 $\mathcal{S}_2$ and $\mathcal{S}_4$ 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 chapter, 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 5.3)

- 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 5.4)

- 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 5.5)

5.2 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

Definition 5.2.1 *A binary relation $\mathcal{R} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the entity-relationships among claims $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_{|\mathcal{O}|}\}$ of data items in $\mathcal{O}$.*

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

5.3 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.

5.3.1 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 5.1). 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 $\mathcal{O}_4$, 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.

5.3.2 Relationship Model

Based on these two themes, we define relation $\mathcal{R} \subseteq \mathcal{V} \times \mathcal{V}$ 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 $\mathcal{G} = (V, E)$ where $V = \mathcal{V}$, i.e., vertices in $\mathcal{G}$ represent the set of distinct claims in $\mathcal{V}$ and edges in E represent the relation between claims at the corresponding vertices. $\forall (u, v) \in \mathcal{R}, \exists (u, v) \in E$ denoting the fact that claim represented by vertex u supports that represented by v. In the rest of this chapter, where applicable, we will use claim $v \in \mathcal{V}$ and the vertex represented by claim $v \in V$ interchangeably. Subgraph $\mathcal{G}_i = (V_i, E_i) \subseteq \mathcal{G}$ represents the relations over claims of data item $\mathcal{O}_i$.

Following standard graph notation, if $e = (u, v) \in E$, 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,

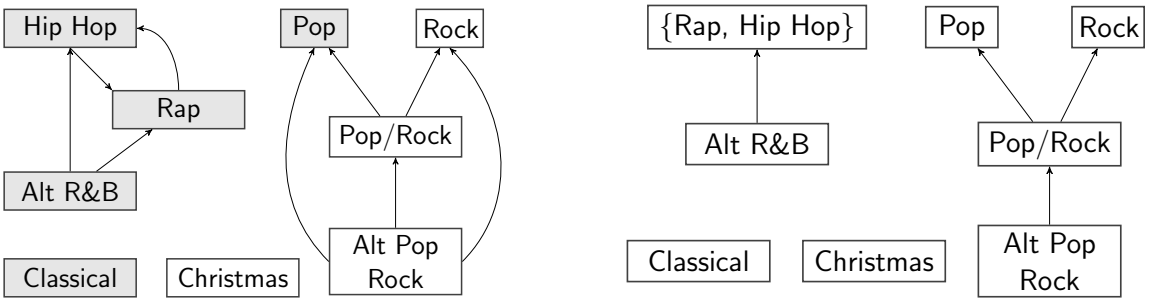
1. v represents *coarser* information than u and encapsulates subsumption.
2. Overlapping claims have a common descendant. Formally, u overlaps with v if there exists w such that $w \rightsquigarrow u$ and $w \rightsquigarrow v$.
3. If $u \rightsquigarrow v$ and $v \rightsquigarrow u$, then u and v represent 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, $\mathcal{G} = (V, E)$ can be seen as a graph with claims as vertices with no edges in between, i.e., $E = \emptyset$.

Example 5.3.1 *Figure 5.1(a) shows the directed graph of relations over claims of data items in Table 5.1. The shaded subgraph denotes relations between claims specific to data item $\mathcal{O}_4$. 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 $\mathcal{G}_2 = (V_2, E_2) \subseteq \mathcal{G}$ consisting of claims of data item $\mathcal{O}_2$. Since edge (Alt



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

Figure 5.1.: Figure 5.1(a) shows the directed graph $\mathcal{G}$ of entity-relationships among claims of data items in Table 5.1. Figure 5.1(b) shows modified graph $\mathcal{G}^{\mathcal{T}}$ obtained as transitive reduction of the equivalent acyclic graph of $\mathcal{G}$.

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 $\mathcal{G}_4 = (V_4, E_4) \subseteq \mathcal{G}$ of claims of data item $\mathcal{O}_4$, claims Hip Hop and Rap are in the same equivalence class and therefore, can be represented by a single vertex.

We process graph $\mathcal{G} = (V, E)$ 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 V by forming the equivalent acyclic graph [70] of $\mathcal{G}$, denoted by $\mathcal{G}^* = (V^*, E^*)$. 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 [71] of $\mathcal{G}$. Consider vertices $u, v \in V$. Let u^* and v^* respectively represent the equivalence classes for claims u and v in $\mathcal{G}^* = (V^*, E^*)$. For $u^* \neq v^*$, if $\exists(u, v) \in E$, then edge $(u^*, v^*) \in E^*$. Note that $\mathcal{G}^*$ may still have redundant edges because of the transitivity property.

2. **Redundant Edges.** We identify the unique transitive reduction [70] of $\mathcal{G}^*$, denoted by $\mathcal{G}^{\mathcal{T}} = (V^*, E^{\mathcal{T}}) \subseteq \mathcal{G}^*$. $\mathcal{G}^{\mathcal{T}}$ has no redundant edge, i.e., for $u, v \in V^*$,

if v is not a parent of u and $u \rightsquigarrow v$, then edge $(u, v) \notin E^{\mathcal{T}}$. 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 $\mathcal{G}_i$. Figure 5.1(b) shows the graph obtained after processing directed graph in Figure 5.1(a).

Complexity Analysis. Equivalent acyclic graph $\mathcal{G}^*$ is obtained in $O(|V| + |E|)$ time [71] whereas transitive reduction $\mathcal{G}^{\mathcal{T}}$ can be derived in $O(|V|^{\beta})$ steps [70] where $\beta \geq 2$. Note that processing directed graph $\mathcal{G}_i$ 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 chapter, we use $\mathcal{G}$ to represent the modified directed graph representation $\mathcal{G}^{\mathcal{T}}$ and $\mathcal{G}_i$ to denote $\mathcal{G}_i^{\mathcal{T}}$.

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 $v \in \mathcal{V}_i$: (a) set of claims in $\mathcal{V}_i$ that support v, denoted by $\delta(v, \mathcal{G}_i)$; and (b) set of claims in $\mathcal{V}_i$ that v supports, denoted by $\alpha(v, \mathcal{G}_i)$. 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 $\mathcal{G}_i$ to $\delta(v, \mathcal{G}_i)$, and add claims in the equivalence class and claims that are its ancestors in $\mathcal{G}_i$ to $\alpha(v, \mathcal{G}_i)$. The notion of supporting and supported claims will be used in Section 5.4.1 to estimate source qualities and correctness of claims.

5.4 Integration with Data Fusion

Given the entity-relationships among claims as described in $\mathcal{G}$ (obtained in Section 5.3), 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.

5.4.1 Revised Data Fusion Methodology

To determine which of the provided claims are correct and which incorrect, state-of-the-art data fusion models [3, 8, 12] 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 $\mathcal{P}$, 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 [8], trustworthiness [3]) or multiple measures (e.g., precision, recall, accuracy, false positive rate [12, 16]). The quality of source $\mathcal{S}_j$, denoted by $\mathcal{Q}_j^{\mathcal{F}}$, is measured based on $\mathcal{V}_i(\mathcal{S}_j)$, the set of claims that $\mathcal{S}_j$ provides for data item $\mathcal{O}_i \in \mathcal{O}$.

  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, $\mathcal{Q}_j^{\mathcal{F}}$ depends on claims in $\mathcal{V}_i(\mathcal{S}_j)$ and claims supported by those in $\mathcal{V}_i(\mathcal{S}_j)$. Given directed graph $\mathcal{G}_i \subseteq \mathcal{G}$ for data item $\mathcal{O}_i$, claim $v \in \mathcal{V}_i(\mathcal{S}_j)$ supports claims in $\alpha(v, \mathcal{G}_i)$ (Section 5.3). Consequently, we replace $\mathcal{V}_i(\mathcal{S}_j)$ by $\overrightarrow{\mathcal{V}}_i(\mathcal{S}_j) = \{\alpha(v, \mathcal{G}_i) \mid v \in \mathcal{V}_i(\mathcal{S}_j)\}$ in the computation of $\mathcal{Q}_j^{\mathcal{F}}$. Clearly, $\mathcal{V}_i(\mathcal{S}_j) \subseteq \overrightarrow{\mathcal{V}}_i(\mathcal{S}_j)$.

  Example 5.4.1 *Consider source $\mathcal{S}_2$ in Table 5.1. Using the modified directed graph in Figure 5.1(b), we observe that $\mathcal{S}_2$ supports claims as shown in Table 5.2.*

  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 $\notin \mathcal{V}_2$*, we do not consider that* Rap *supports* Hip Hop *in the context of data item $\mathcal{O}_2$.*

  Comparing Table 5.2 with Table 5.1, we observe that out of the 11 *claims $\mathcal{S}_2$ supports,* 8 *are correct resulting in a precision (fraction of claims provided that*

Table 5.2.: Claims provided or supported by source $\mathcal{S}_2$.

ID	$\mathcal{V}_i(\mathcal{S}_2)$	$\overrightarrow{\mathcal{V}_i}(\mathcal{S}_2)$	**Correct**
$\mathcal{O}_1$	Christmas	Christmas	Pop, Pop/Rock
$\mathcal{O}_2$	Alt Pop Rock, Rap	Alt Pop Rock, Pop/Rock, Pop, Rock, Rap	Alt Pop Rock, Pop/Rock, Pop, Rock
$\mathcal{O}_3$	Pop	Pop	Pop/Rock, Pop
$\mathcal{O}_4$	Pop, Alt R&B	Pop, Alt R&B, Hip Hop, Rap	Alt R&B, Hip Hop, Rap

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 in the absence of knowledge of relations among the claims of data items, the precision and recall of* $\mathcal{S}_2$ *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, $\mathcal{P}(v)$ is defined for items in the training data and $Q^{\mathcal{F}}$ is computed over those items. Otherwise, $\mathcal{Q}^{\mathcal{F}}$ is initialized to random values, and source quality and claim correctness are estimated iteratively.

- **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 $v \in \mathcal{V}_i$, denoted by $\mathcal{P}(v)$, is computed in terms of the quality measures of sources in $\mathcal{S}^i(v)$, 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 $\mathcal{G}_i \subseteq \mathcal{G}$ for data item $\mathcal{O}_i$, claim v is supported by claims in $\delta(v, \mathcal{G}_i)$. In estimating the correctness of v by a particular data fusion model, we replace $\mathcal{S}^i(v)$ by

Algorithm 3 Estimating quality of sources

procedure EstimateSourceQuality($\mathcal{D}$, $\mathcal{G}$, $\mathcal{F}$, $\mathcal{P}$)
  for $s \in \mathcal{S}$ **do**
    for $O_i \in \mathcal{O}$ **do**
      $\overrightarrow{\mathcal{V}_i}(s) = \{\alpha(v, \mathcal{G}_i) \mid v \in \mathcal{V}_i(s)\}$
      for $v \in \overrightarrow{\mathcal{V}_i}(s))$ **do**
        Compute $\mathcal{Q}^{\mathcal{F}}(s)$ according to $\mathcal{F}$ based on $\mathcal{P}(v)$
      end for
    end for
  end for
  Output $\mathcal{Q}^{\mathcal{F}}$, the quality measures of sources
end procedure

Algorithm 4 Estimating correctness probabilities of claims

procedure EstimateClaimCorrectness($\mathcal{D}$, $\mathcal{G}$, $\mathcal{F}$, $\mathcal{Q}^{\mathcal{F}}$)
  for $\mathcal{O}_i \in \mathcal{O}$ **do**
    for claim $v \in \mathcal{V}_i$ **do**
      $\mathcal{S}^i(\overrightarrow{v}) = \{\mathcal{S}^i(u) \mid u \in \delta(v, \mathcal{G}_i)\}$
      for $s \in \mathcal{S}^i(\overrightarrow{v})$ **do**
        Compute $\mathcal{P}(v)$ according to $\mathcal{F}$ based on $\mathcal{Q}^{\mathcal{F}}(s)$
      end for
    end for
  end for
  Output $\mathcal{P}$, the correctness probability of claims
end procedure

$\mathcal{S}^i(\overrightarrow{v}) = \{\mathcal{S}^i(u) \mid u \in \delta(v, \mathcal{G}_i)\}$. Again, $\mathcal{S}^i(v) \subseteq \mathcal{S}^i(\overrightarrow{v})$. 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 $\mathcal{G}_i$, instead of computing the correctness of each provided claim for data item $\mathcal{O}_i$, we compute the correctness of each vertex in $\mathcal{G}_i$. 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.

Algorithm 5: ModifyDataFusion

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

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 5.

Iterative fusion models [3, 8] randomly initialize source quality estimates and iterate over lines 1 and 2 until $\mathcal{Q}^{\mathcal{F}}$ converges. When ground truth data is available, fusion models [12] utilize it to compute source quality estimates.

5.4.2 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 $\mathcal{P}$ and directed graph $\mathcal{G}$, we describe the steps to determine correct claims for data items in Algorithm 6. Lines 4-6 identify root nodes of the directed graph $\mathcal{G}_i$ over claims of data item $\mathcal{O}_i$. 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 $\mathcal{O}_i$. The algorithm then identifies children nodes of the selected

Algorithm 6: DetermineCorrectClaims

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

vertex for further traversal and repeats lines 8-10 until a leaf node (i.e., vertex with no children) is reached.

5.5 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 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.

5.5.1 Competing Methods

We evaluate effectiveness of using the domain information on entity-relationships among claims on the following single- and multi-truth data fusion models (also described in Section 3.2):

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 [3]: Iteratively computes trustworthiness of sources and confidence in claims, and selects claim with the highest confidence to be correct.

ACCU [8]: 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 [12]: 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 [52](denoted by PO). We implemented all the algorithms in Java.

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

% 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.

5.5.3 Real-World Data

We conducted experiments on the Restaurants dataset in [34] 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 [52]) but can be any arbitrary directed graph with cycles. A partial ordering solution, therefore, will not be directly applicable to resolve such conflicting data.

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

Table 5.3.: Effectiveness of data fusion models on Restaurants. While effective in identifying correct claims, PrecRec outputs inconsistent correct claims.

	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

5.5.4 The Case for Consistency

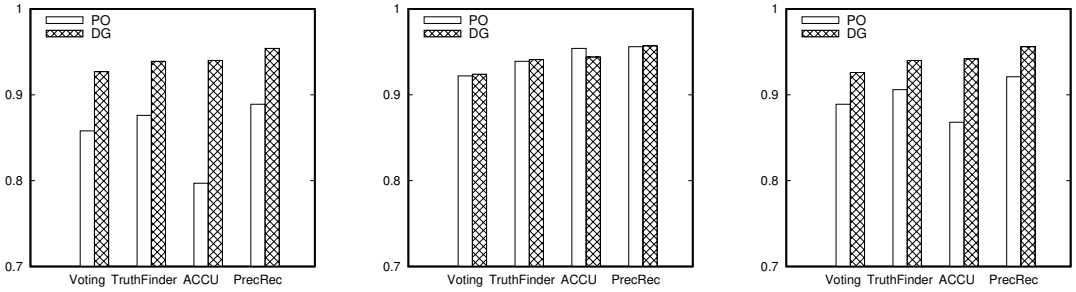
To demonstrate the need for approaches that generate consistent correct claims, we run the described data fusion models (Voting, TruthFinder, ACCU and PrecRec) on Restaurants and report their performance as measured by precision, recall and F1-measure in Table 5.3. We observe that while the multi-truth 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.

5.5.5 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

Table 5.4.: Effect of integrating the entity-relationships among claims on the effectiveness of different fusion models.

	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**



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

Figure 5.2.: Comparing relationship models PO and DG during fusion of Restaurants. For PO, we set probability threshold $\theta = 0.05$.

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.

We present in Table 5.4, the results of using DG and PO, entity-relationships among claims, in conjunction with the data fusion models. Comparing the results with Table 5.3, 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 5.2, 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 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 5.4 vs. PrecRec in Table 5.3).

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.

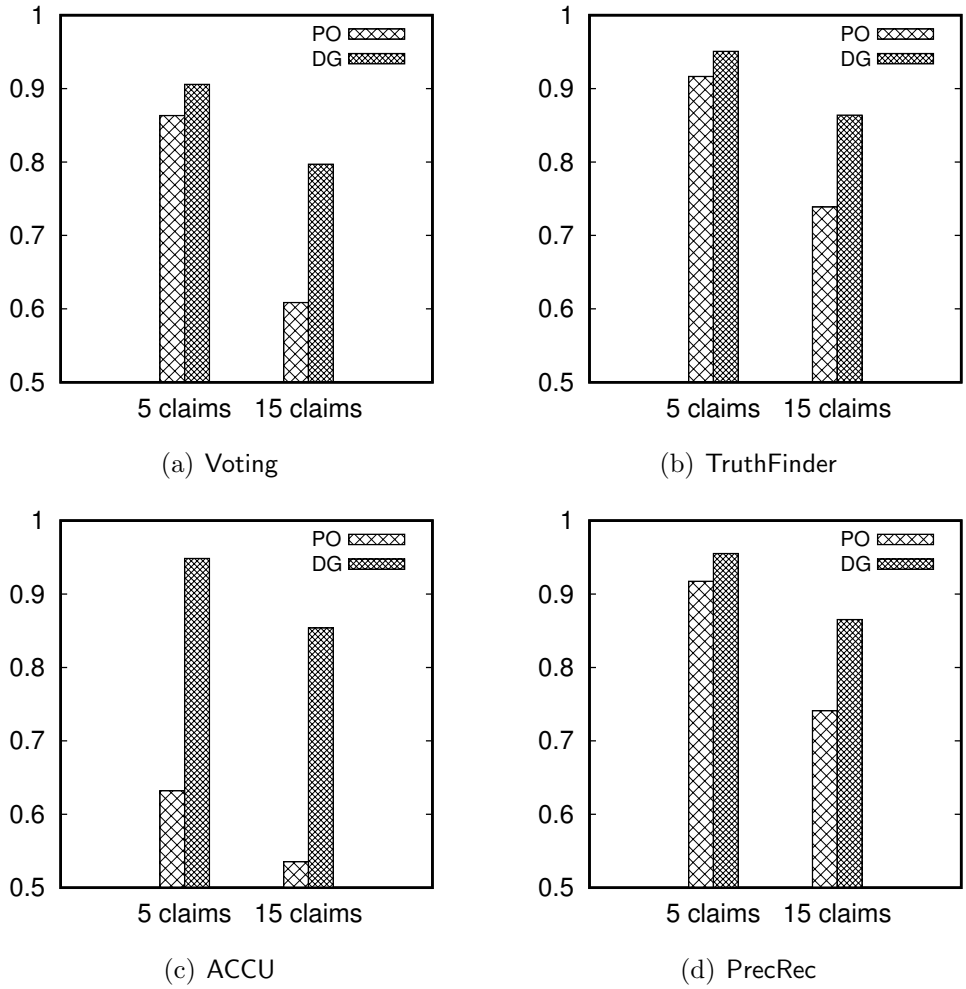
5.5.6 Synthetic Data

To compare our approach (DG) to the partial order algorithm (PO) that is tailored to use hierarchical ontologies, we conducted a set of experiments on synthetic data with acyclic edges in the graphs depicting the relationship among its claims. The prime parameters for data generation were: number of data items (m), number of sources (n), number of distinct claims per item (k) and probability of an edge between two claims of a data item (p_h). We compare the approaches for a number of scenarios that we discuss in the following.

5.5.7 Comparison with Partial Order Algorithm

In the first experiment, we generate data by varying k and present the results in Figure 5.3. Since both the approaches have comparable (and high) precision, we report only the recall of the methods. The partial order algorithm, in an attempt to limit overestimating source trustworthiness, does not update the trustworthiness of sources when correctness of claims are updated during fusion. However, as is evident from the plots, incorporating this information greatly influences effectiveness of the algorithm. We also observe that as the number of claims increases, there is a stark difference between performance of PO and DG. Specifically, we observe at least 10 percentage points of improvement by using the latter over the former when data items have a large number of claims. This behavior can be explained thus: as the number of claims increases, there is the possibility of more complicated edges, e.g., overlaps, existing between claims. DG is designed to address such edges whereas PO is not, thus resulting in lower recall of the latter.

Next, we test how well the approaches perform as claims are more (or less) related to each other, i.e., as claims have more (or less) number of edges between them. We

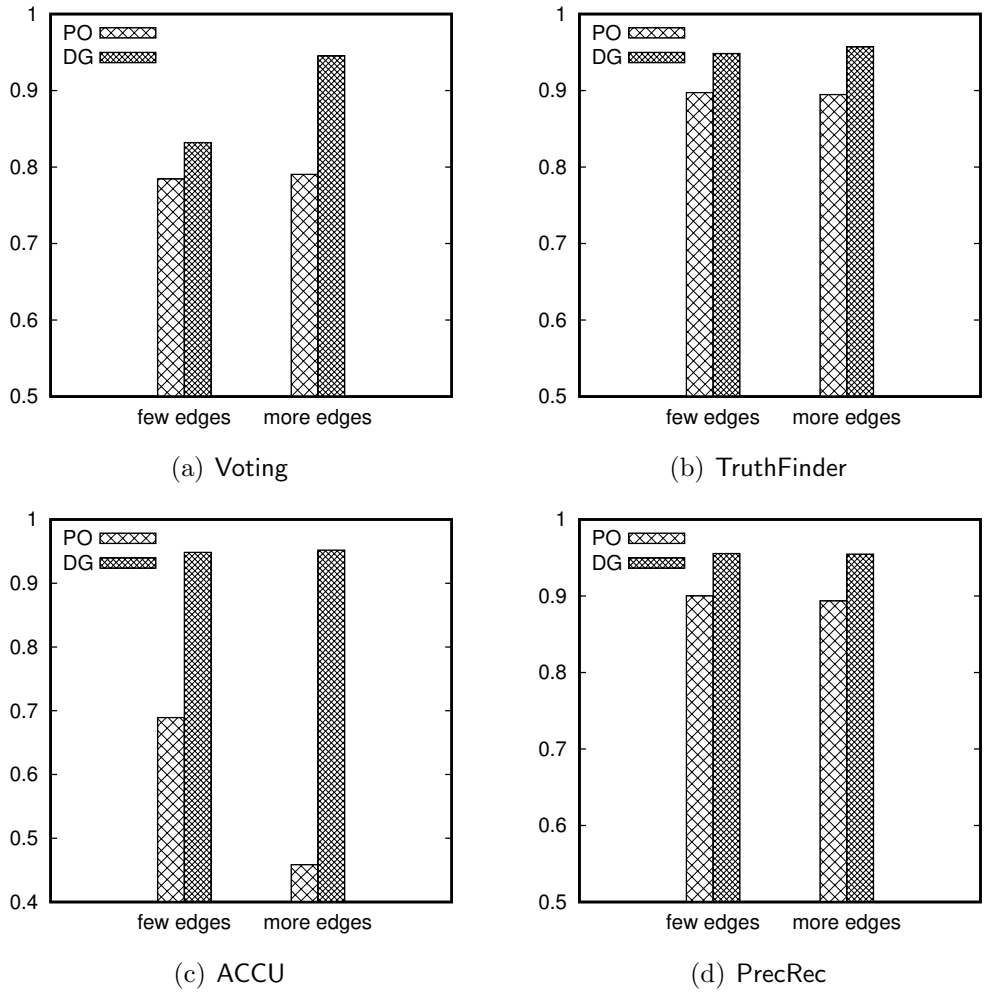


(a) Voting (b) TruthFinder

(c) ACCU (d) PrecRec

Figure 5.3.: Comparing the recall of DG with that of PO on synthetic data with different number of claims per data item.

generate datasets by varying p_h and present the results in Figure 5.4. Again, we report only recall of the methods since their precision is comparable. We observe that as claims are more related to each other, there may exist greater number of overlapping edges. As a result, PO is able to retrieve fewer correct claims than DG, with as much as 45 percentage points between recall of the two approaches for the same fusion model. Interestingly, fusion models TruthFinder, ACCU and PrecRec exhibit similar performance across datasets when used in conjunction with DG. It is surprising because although PrecRec uses training data, TruthFinder and ACCU are



(a) Voting (b) TruthFinder

(c) ACCU (d) PrecRec

Figure 5.4.: Comparing the recall of DG with that of PO when the claims of data items are rarely related to each other vs. when they are related quite often.

automated fusion models. This observation further strengthens our finding from real data that with the knowledge of relations, we may not need advanced fusion models to achieve high effectiveness.

Experiment Takeaways. On data with strict partial orders, DG outperforms PO — both as the number of claims for a data item increases, and as greater number of claims are related.

5.5.8 Discussion on the Efficiency of Directed Graphs

Throughout the course of these experiments, we also kept track of the time taken by the two approaches, PO and DG, on Restaurants when integrated with different fusion models. The time taken by the approaches was broken down into the runtime of their constituents — pre-processing the relationship model, running the data fusion model and identifying correct claims. We observed that the cost of incorporating the knowledge representation in data fusion models is similar irrespective of whether PO or DG is used. The run-time for identifying correct claims is smaller for PO than DG because the former uses a probability threshold that prunes substantial parts of (and, therefore, does not require complete traversal of) the relationship model. In a nutshell, DG has a longer run-time since it takes additional relations into account and performs computations overlooked by PO, e.g., leveraging relations during source quality estimation and navigating multiple access paths for identifying correct claims.

5.6 Summary

In this chapter, 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 and evaluated our approach against other methods that incorporate such relation information in the data. We showed that, compared to other methods, our algorithm achieves significant improvement in fusion results.

Through experimental evaluation on real-world data, we 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 con-

sistent with each other. Our approach outperforms state-of-the-art algorithms that consider the presence of relationships over claims of data items.

Results from this chapter were published in [72].

6 FUTURE WORK

Recent proliferation of "fake" news has resulted in a number of solutions for automated fact-checking that view the problem from a largely linguistic perspective. We observe that the problem of false data detection has roots in several extensively studied research areas in data management and data mining such as data integration, data cleaning, crowdsourcing and machine learning. In this chapter, we present our ongoing and future work aimed at combating false data on the Internet.

False data detection mechanisms have primarily leveraged either linguistic cues or structured conflict resolution approaches to distinguish correct from incorrect information. *Language-based* false data detection approaches heavily rely on different aspects of language (e.g., tone, stance, objectivity, hedges, negation) and structure of community networks (e.g., social media, microblogging websites and e-commerce websites) to fight fabricated information such as hoaxes, rumors, vandalism, fake product reviews, controversies etc. *Data fusion* mechanisms, on the other hand, consider the role of information providers to be vital in determining the correctness of claims provided by them. While the latter has proved quite successful in resolving inconsistencies in structured data, it has not been fully explored for the resolution of unstructured data conflicts.

Furthermore, in the era of "alternative" facts, fact-checking websites, such as Snopes and PolitiFact, have emerged as vanguards having dedicated teams of employees who comb through speeches, news stories, press releases to verify rumors and political claims. We contend that advances made in effectively involving users in data management tasks, along with language-based and structured conflict resolution systems, will benefit the cause of combating misinformation on the Internet.

We propose the architecture of AuthIntegrate, an end-to-end system that ingests (possibly) conflicting data from disparate information providers, curates and

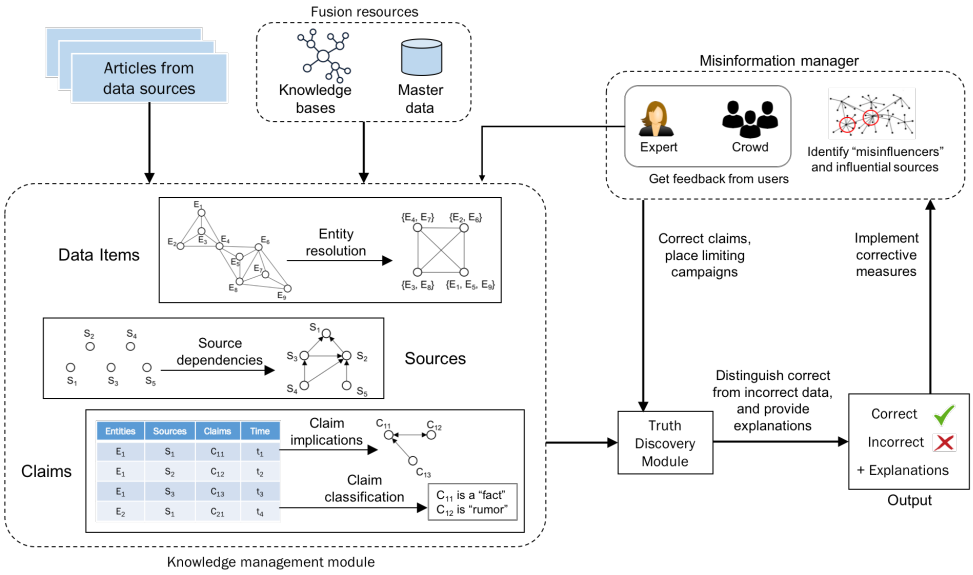


Figure 6.1.: Figure depicts envisioned architecture of the AUTHINTEGRATE system.

presents highly accurate data to end-users. In the following, we address key challenges in building this system and outline an agenda for future research. We focus on (a) detection of false data (Sections 6.1 and 6.2) coupled with combating its spread through identifying mis-influencers and installing corrective measures (Section 6.3).

6.1 AUTHINTEGRATE: Knowledge Management Module

The foremost step in our proposed architecture is information extraction, which focuses on retrieving structured information from the collection of unstructured and noisy textual data provided by disparate data sources such as news agencies and social media. Broadly, information extraction approaches can be categorized as based on knowledge engineering techniques (that leverage expert intervention in the form of rules, examples and domain knowledge), and machine learning techniques (that learn concept-specific mapping from text and generate rules from training data).

6.1.1 Information Extraction

We envision the knowledge management module to take a hybrid approach learning from training data and external resources, such as general-purpose knowledge bases, master data and human input, to extract data items and their relationships. Several data management problems form pivotal building blocks of this module, e.g., entity resolution [73]; learning entity-relationships [43, 44] and establishing source dependencies [8, 12, 36]; and provenance [74], to determine the origin and information on history of the life cycle of data. A comprehensive knowledge of the relationships between data items, claims and sources will prove instrumental in explaining the *plausability* of claims whereas provenance information and metadata associated with claims, such as the context of a claim and fragments that have been used as is or have been altered, will be important in designing algorithms that assess sources and claims in a principled manner.

6.1.2 Leveraging Linguistic Cues

We also intend to understand the complete context of claims — whether they are facts, opinions, rumors, hoaxes, urban legends, vandalisms, joke, advertisement, controversy, their sentiment and establishing their temporal existence (happened in the past or is a prediction) — a natural language processing task made feasible with the help of domain experts, crowdsourcing platforms and knowledge bases. We discuss how these classifications help build the reputation of sources (in Section 6.2) and curb the rise of false data (in Section 6.3).

6.2 AuthIntegrate: Truth Discovery Module

Data fusion models consider source characteristics to play a pivotal role in estimating the correctness of claims — an approach that is in sharp contrast with most false data detection mechanisms that solely exercise natural language techniques to

identify correct information. While the idea of *all important* sources has made tremendous advances in resolving conflicts, data fusion is designed on the single premise that sources are primarily *benevolent*. Current times (of abundant false news), however, bear testimony to the fact that the "honest sources" assumption no longer holds true. Adversarial settings are breeding grounds for false and biased data that have the potential to misguide fusion systems toward incorrect conclusions.

6.2.1 Modeling Distrustful Scenarios

It is imperative to design data fusion models that are guaranteed to be effective even in the face of malicious data sources. Identifying adversarial and colluding data sources during data fusion is challenging primarily because these sources may not behave consistent over time: driven by their interest, data sources may furnish data sporadically or continuously in large amounts. Traditional data integration mechanisms consider data sources to be independent of each other. Recent studies have found data sources to either copy from one another [8,34,35] or be correlated [12] *positively* (when sources follow similar data extraction rules) or *negatively* (when sources provide complementary data or extract different types of data).

Existing data fusion techniques thrive on the principle of *trust* in data sources — information from trusted sources are more likely to be correct and a source is trusted if it provides more accurate information. However, in situations when malicious sources could collude to falsely boost their level of trust, it becomes easy to propagate misinformation and prove detrimental to data fusion. We, therefore, need principled mechanisms to answer the following question: Is it possible to render data fusion systems aware of the presence of collusive relations among data sources?

6.2.2 Broader Characterization of Sources

Data sources are primarily characterized in terms of performance metrics, such as accuracy, precision and recall, that depend on the number of correct and incorrect

claims provided by sources. Counting-based approaches fail to address the quality of sources where claims may span lengthy texts. There is a need to develop source quality measures that encompass wider categories of claims, such as hoax, opinion, fact, prediction etc., and are able to capture evolving language tones and stances. Characterizing sources in this manner helps refine their reputation. For example, speculative facts and opinions make sources less credible than correct facts and may, in fact, damage their credibility.

6.2.3 Leveraging Knowledge Bases and Knowledge Representations

Claims for data items, provided by different sources, are often related to each other. We intend to apply our solutions from Chapter 5 toward integrating the knowledge of claim relationships during conflict resolution. The effectiveness of our solution, however, depends on completeness of the extracted knowledge and can be improved by accounting for ambiguity in relations. For example, depending on the context, *jaguar* could be related to either *cat* or *car*. Currently, we assume the extracted knowledge to be agnostic to the context and thus, devoid of such uncertainties. General-purpose knowledge bases tend to capture contextual information and will be able to address ambiguous relationships inherent in data. To integrate knowledge bases with data fusion models, we will need to devise algorithms that efficiently sift through the volumes of data and identify information pertinent to the data at hand. Moreover, the framework for integrating binary entity-relationships can be further improved by exploring more expressive formalisms such as logic-based knowledge representations and conceptual graphs.

6.3 AuthIntegrate: Misinformation Manager Module

The objective of this module is two-pronged: one, identifying influential data sources that have the potential of inflicting maximum damage, and two, implementing

corrective measures to minimize the damage. Toward this goal, we envision strategies to efficiently utilize human input and to limit the spread of false information.

6.3.1 Human-in-the-Loop Conflict Resolution

Although automated fact-checking systems [75] enable deconstructing vague and countering questionable claims, the undeniable success of fact-checking websites (e.g., Snopes, PolitiFact) has made it clear that verification by experts is a stepping stone in the battle to counter false data. Corrective information published from an authoritative resource has the potential to diffuse enormously and prevent the rapid increase in false data [67, 76]. However, incorporating user input is challenging because there are a large number of claims and few experts with limited budgets to process the claims. This approach of vetting by experts is particularly important in the face of limited information on emerging claims. We intend to build upon strategies proposed in [67] to judiciously leverage user feedback by determining the most beneficial claims to be validated; these strategies can also be utilized for labeling different forms of claims (in Section 6.1) where the challenge is to prioritize labeling tasks for annotators.

Imperfect Feedback. The solutions described in Chapter 4 largely assume access to domain experts; preliminary solutions involving a crowd of workers has also been presented. We believe holistically modeling users would facilitate better judgement over their input. We intend to benefit from the breadth of research in crowdsourcing over the last few years. However, we can readily identify challenges in involving non-experts in resolving conflicts. To aggregate uncertain feedback on the same data item, we need to holistically model users and the quality of their input while also taking into account their cognitive and physiological characteristics. Alongside, there is a need to develop an economical model to incorporate uncertain feedback that addresses the trade-off between time and the cost and quality of improvement in conflict resolution. We also intend to extend the approximate algorithms in the decision-theoretic framework to other data fusion models.

6.3.2 Limiting the Spread of False Information.

False data has the potential to be considered true by a large fraction of consumers; it is, therefore, of utmost importance to identify *misinfluencers* and prevent them from spreading misinformation. [77, 78] demonstrated that by placing limiting campaigns at influential nodes, it is possible to minimize the number of individuals that believe in a particular piece of misinformation and prevent the growth of false data. We propose to extend this idea of identifying misinfluencers to Bayesian networks of data items and sources, which is different from the influence maximization problem that examines the flow of a single propaganda (false data usually spans more than just one claim in a specific community (false data may extend to a multitude of communities such as social media, blogs and the Web).

7 CONCLUSION

Due to the proliferation of data on the Internet, conflict resolution of data integrated from disparate sources has continued to garner interest from research communities over the last few decades. The prime objective of this research is to improve conflict resolution of data integrated from different sources. In this dissertation, we proposed to augment automatic data fusion systems with the knowledge of relationships existing within data provided by different sources.

We proposed a novel user feedback framework that employs active learning techniques to integrate user-provided ground truth labels and rapidly improve the effectiveness of data fusion. To minimize user interaction, we proposed a decision-theoretic approach to determine the data item best suited for validation. The proposed decision-theoretic approach was expensive for large-scale datasets and we proposed approximate algorithms that incorporate relationships among data items and sources to reduce this cost. Through experimental evaluation on real-world data, we demonstrated applicability of the approximation algorithms to large datasets and and existing fusion models, and showed the trade-off between effectiveness and efficiency achieved by the proposed solutions.

We also proposed incorporating entity-relationships among claims during the process of data fusion where data items may have multiple correct claims. Our proposed encoding of entity-relationships in the form of an arbitrary directed graph captures most of the binary relations existing between claims of data items. We outlined steps to pre-process the directed graph for effective representation and efficient navigation during data fusion, and proposed modifications to existing data fusion models for supporting the directed graph representation. We implemented our approach on top of existing fusion models and through experiments on real data, demonstrated its effectiveness in identifying multiple related and consistent truths.

Finally, we presented some of our ongoing and future work aimed at resolving conflicts during data integration tasks. Data integration and conflict resolutions remain longstanding areas of research and significant progress has already been made in these areas. Notwithstanding, there are several interesting problems that remain to be solved. This dissertation is a step forward toward our goal of resolving data conflicts during data integration and presenting end-users with highly accurate data integrated from disparate data sources.

REFERENCES

REFERENCES

[1] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved? In *Proceedings of the VLDB Endowment (PVLDB)*, pages 97–108, 2012.

[2] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN)*, pages 233–244, 2012.

[3] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, pages 796–808, 2008.

[4] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1–41, January 2009.

[5] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web (WWW)*, pages 107–117, 1998.

[7] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, pages 604–632, 1999.

[8] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 2, pages 550–561, 2009.

[9] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)*, pages 131–140, 2010.

[10] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 675–684, 2015.

[11] Jeff Pasternack and Dan Roth. Latent credibility analysis. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 1009–1020, 2013.

[12] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM International Conference on Management of Data (SIGMOD)*, pages 433–444, 2014.

[13] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. In *Proceedings of the VLDB Endowment (PVLDB)*, pages 425–436, 2014.

[14] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 745–754, 2015.

[15] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 217–226, 2011.

[16] Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. A Bayesian approach to discovering truth from conflicting sources for data integration. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 5, pages 550–561, February 2012.

[17] Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Z. Miklos, K. Aberer, A. Gal, and M. Weidlich. Pay-as-you-go reconciliation in schema matching networks. In *Proceedings of the IEEE 30th International Conference on Data Engineering (ICDE)*, pages 220–231, 2014.

[18] AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM International Conference on Management of Data (SIGMOD)*, pages 509–520, 2001.

[19] Shawn R. Jeffery, Michael J. Franklin, and Alon Y. Halevy. Pay-as-you-go user feedback for dataspace systems. In *Proceedings of the 2008 ACM International Conference on Management of Data (SIGMOD)*, pages 847–860, 2008.

[20] Donatella Firmani, Barna Saha, and Divesh Srivastava. Online entity resolution using an oracle. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 9, pages 384–395, 2016.

[21] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 877–882, 2007.

[22] Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F. Ilyas. Guided data repair. In *Proceedings of the VLDB Endowment (PVLDB)*, pages 279–289, 2011.

[23] Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. Conflict resolution with data currency and consistency. *Journal of Data and Information Quality (JDIQ)*, 5(1-2):6:1–6:37, 2014.

[24] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

[25] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 647–653. 2001.

[26] Darius Braziunas. Computational approaches to preference elicitation. Technical report, Department of Computer Science, University of Toronto, 2006.

[27] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. Crowddb: Answering queries with crowdsourcing. In *Proceedings of the 2011 ACM International Conference on Management of Data (SIGMOD)*, pages 61–72, 2011.

[28] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. Scaling up crowd-sourcing to very large datasets: A case for active learning. In *Proceedings of the VLDB Endowment (PVLDB)*, pages 125–136, 2014.

[29] Nguyen Quoc Viet Hung, Duong Chi Thang, Matthias Weidlich, and Karl Aberer. Minimizing efforts in validating crowd answers. In *Proceedings of the 2015 ACM International Conference on Management of Data (SIGMOD)*, pages 999–1014, 2015.

[30] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 614–622, 2008.

[31] Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 259–268, 2009.

[32] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Wenyuan Yu. Towards certain fixes with editing rules and master data. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 3, pages 173–184, 2010.

[33] Wenfei Fan, Shuai Ma, Nan Tang, and Wenyuan Yu. Interaction between record matching and data repairing. *Journal of Data and Information Quality (JDIQ)*, 4(4):16:1–16:38, 2014.

[34] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth discovery and copying detection in a dynamic world. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 2, pages 562–573, 2009.

[35] Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. Global detection of complex copying relationships between sources. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 3, pages 1358–1369, 2010.

[36] Anish Das Sarma, Xin Luna Dong, and Alon Halevy. Data integration with dependent sources. In *Proceedings of the 14th International Conference on Extending Database Technology (EDBT/ICDT)*, pages 401–412, 2011.

[37] Chuishi Meng, Wenjun Jiang, Yaliang Li, Jing Gao, Lu Su, Hu Ding, and Yun Cheng. Truth discovery on crowd sensing of correlated entities. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 169–182, 2015.

[38] Shiguang Wang, Lu Su, Shen Li, Shaohan Hu, Tanvir Amin, Hongwei Wang, Shuochao Yao, Lance Kaplan, and Tarek Abdelzaher. Scalable social sensing of interdependent phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 202–213, 2015.

[39] Shiguang Wang, Dong Wang, Lu Su, Lance Kaplan, and Tarek F Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In *The 2014 IEEE Real-Time Systems Symposium (RTSS)*, pages 74–85. IEEE, 2014.

[40] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 697–706, 2007.

[41] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference (ISWC/ASWC)*, pages 722–735, 2007.

[42] George A. Miller. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[43] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2010.

[44] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL/IJCNLP)*, pages 1003–1011, 2009.

[45] Boris Gelfand, Marilyn Wulfekuhler, and William Punch. Discovering concepts in raw text: Building semantic relationship graphs. Technical report, Michigan State University, East Lansing, MI, USA, 1998.

[46] Stanley Loh, Leandro Krug Wives, and José Palazzo M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explorations Newsletter*, 2:29–39, 2000.

[47] Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[48] Sung Ju Hwang, Fei Sha, and K. Grauman. Sharing features between objects and their attributes. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1761–1768, 2011.

[49] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 8689, 2014.

[50] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 17–24, 2007.

[51] Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 27–34, 2004.

[52] Valentina Beretta, Sébastien Harispe, Sylvie Ranwez, and Isabelle Mougenot. How can ontologies give you clue for truth-discovery? an exploratory study. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 15:1–15:12, 2016.

[53] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Data fusion: Resolving conflicts from multiple sources. In *Web-Age Information Management*, pages 64–76, 2013.

[54] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, and Divesh Srivastava. Online data fusion. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 4, pages 932–943, 2011.

[55] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[56] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2003.

[57] Claude E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 2001.

[58] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[59] Gretchen B. Chapman and Frank A. Sonnenberg (Editors). *Decision Making in Health Care: Theory, Psychology, and Applications.* Cambridge University Press, 2003.

[60] Steven Euijong Whang, David Marmaros, and Hector Garcia-Molina. Pay-as-you-go entity resolution. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, pages 1111–1124, 2013.

[61] Steven Euijong Whang, Peter Lofgren, and Hector Garcia-Molina. Question selection for crowd entity resolution. *Proceedings of the VLDB Endowment (PVLDB)*, 6(6):349–360, 2013.

[62] Rubi Boim, Ohad Greenshpan, Tova Milo, Slava Novgorodov, Neoklis Polyzotis, and Wang Chiew Tan. Asking the right questions in crowd data sourcing. In *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE)*, pages 1261–1264, 2012.

[63] Chen Jason Zhang, Lei Chen, Hosagrahar V. Jagadish, and Caleb Chen Cao. Reducing uncertainty of schema matching via crowdsourcing. *Proceedings of the VLDB Endowment (PVLDB)*, 6(9):757–768, 2013.

[64] John M. Winn and Christopher M. Bishop. Variational message passing. In *Journal of Machine Learning Research (JMLR)*, pages 661–694, 2005.

[65] Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, volume 89. Springer, 1998.

[66] Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 877–885, 2010.

[67] Romila Pradhan, Siarhei Bykau, and Sunil Prabhakar. Staging user feedback toward rapid conflict resolution in data fusion. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD)*, pages 603–618, 2017.

[68] Romila Pradhan, Siarhei Bykau, and Sunil Prabhakar. A framework to integrate user feedback for rapid conflict resolution. In *Proceedings of the IEEE 34th International Conference on Data Engineering (ICDE)*, 2018.

[69] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. In *Proceedings VLDB Endowment*, volume 7, pages 881–892, 2014.

[70] Alfred V. Aho, Michael R. Garey, and Jeffrey D. Ullman. The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1972.

[71] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1972.

[72] Romila Pradhan, Walid G. Aref, and Sunil Prabhakar. Leveraging data relationships to resolve conflicts from disparate data sources. In *Database and Expert Systems Applications – 29th International Conference (DEXA)*, 2018.

[73] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(1):1–16, 2007.

[74] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. *Provenance in Databases: Why, How, and Where*. Now Publishers Inc., 2009.

[75] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. Toward computational fact-checking. In *Proceedings of the VLDB Endowment (PVLDB)*, volume 7, pages 589–600, 2014.

[76] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: A twitter case study. *PLOS ONE*, 10(4):1–18, 04 2015.

[77] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 665–674, 2011.

[78] Marco Amoruso, Daniele Anello, Vincenzo Auletta, and Diodato Ferraioli. Contrasting the spread of misinformation in online social networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1323–1331, 2017.