This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

A Survey on Truth Discovery: Concepts, Methods, Applications, and Opportunities

Shuang Wang, He Zhang, Quan Z. Sheng, Xiaoping Li, Zhu Sun, Taotao Cai, Wei Emma Zhang, Jian Yang, Qing Gao

***Abstract*—In the era of data information explosion, there are different observations on an object (e.g., the height of the Himalayas) from different sources on the web, social sensing, crowd sensing, and data sensing applications. Observations from different sources on an object can conflict with each other due to errors, missing records, typos, outdated data, etc. How to discover truth facts for objects from various sources is essential and urgent. In this paper, we aim to deliver a comprehensive and exhaustive survey on truth discovery problems from the perspectives of concepts, methods, applications, and opportunities. We first systematically review and compare problems from objects, sources, and observations. Based on these problem properties, different methods are analyzed and compared in depth from observation with single or multiple values, independent or dependent sources, static or dynamic sources, and supervised or unsupervised learning, followed by the surveyed applications in various scenarios. For future studies in truth discovery fields, we summarize the code sources and datasets used in above methods. Finally, we point out the potential challenges and opportunities on truth discovery, with the goal of shedding light and promoting further investigation in this area.**

***Index Terms*—Truth Discovery, Source Reliability, Object Confidence, Dependent Sources.**

I. Introduction

In the past few decades, the amount of helpful information available on the Web has been proliferating, which has brought dramatic changes to human society [1]. People are more dependent on the Web to fulfill their information needs than ever [2]. However, a huge amount of disinformation, outdated data, and factual errors are filled on the Web [3]. It is difficult for users to distinguish the truth from various information [4]–[7]. When searching for the birthplace of Adolf Hitler on the web, the answers include "Austria", "Braunau", "Germany". It is difficult for users to choose the correct information among these conflicting answers. Since the collected information may be informed, incomplete, outdated, or existing factual errors, it is crucial to discover truths from various information which improves accuracy in information extraction. To solve the problem, truth discovery has attracted researchers' attention recently [8]–[10] in many real-world application scenarios including web, social sensing, crowd sensing, privacy sensing, and deep neural network applications. For example, true facts are found from a large amount of conflicting information on many objects provided by various websites [11]. The sensory data collected from various mobile devices are usually unreliable in mobile cloud computing. Truthful information is extracted from unreliable sensory data in mobile crowd sensing [12]. It is necessary to aggregate noisy information on the objects, entities, or events collected from various sources.

Shuang Wang, He Zhang, and Xiaoping Li are with the School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China. E-mail: shuangwang@seu.edu.cn, zhanghezhedu@163.com, xpli@seu.edu.cn.

Quan Z. Sheng and Jian Yang are with the School of Computing, Macquarie University, Sydney, NSW 2109, Australia. E-mail: {michael.sheng, jian.yang}@mq.edu.au.

Zhu Sun is with the Centre of Frontier AI Research, Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore. E-mail: sunzhuntu@gmail.com

Taotao Cai is with the Computing at the University of Southern Queensland, Australia. Email: taotao.cai@usq.edu.au.

Wei Emma Zhang is a Senior Lecturer and As- sociate Head of People and Culture at the School of Computer and Mathematical Sciences, and a researcher at the Australian Institute for Machine Learning, The University of Adelaide, Australia. Email:wei.e.zhang@adelaide.edu.au.

Qing Gao is with School of Automation Science and Electrical Engineering, Beihang University, China. Email:gaoqing@buaa.edu.cn.

Manuscript received April 19, 2021; revised August 16, 2021.

A. Truth Discovery Overview

Truth discovery is motivated to resolve conflict information on objects among multiple resources which contains objects (i.e., things of interest), sources (i.e., providing the information about objects), and observations (i.e., the information provided by sources about objects). To make the following description clear and consistent, we introduce the basic definitions on the object, source, and observation that are used in this survey.

- **Object**: The object is a basic conception to describe an interesting thing in truth discovery. An object is defined as a thing of interest [13], [14], a device [15], [16], or an entity [17].
- **Source**: A source describes the place where we can collect information about objects [18].
- **Observation**: An observation, also known as a record, is a 3-tuple that consists of an object, a source, and its provided value.
- **Truth**: The identified truth for an object is the information selected as the most trustworthy one of all possible candidate values related to this object [18].

For example, when seeking information about book authors, the object is the book itself, while the sources of information contains book publishers. The observations are the authors' name provided by the book publishers while the truth is the truth values of authors for the book.

Precisely, objects can be categorical [19]–[21] (i.e., book information) or continuous [14], [22], [23](i.e., weather

information). Sources can be dependent [24]–[28], or independent [11], [19]–[21] in terms of whether there is relationship among sources or not. Moreover, we further subdivided the observations into single [10], [29]–[32], multiple [11], [26], [33], [34], and unknown [35]. For example, if we would like to discover information of a book (e.g., the title, author and published year of a book), the object is categorical, while if we want to know the weather in a day, it is a continuous object. If the problem is to find the title of a book, it is easily to know that the object value is single. However, how to find the authors of a book is a multiple value truth discovery problem [36] since usually, there are more than one author in a book. If the constraints of the book information are known in advance but all the sources which provide information on the book cannot satisfy the constraints, it is an unknown value truth discovery problem [18]. For each object, various sources provide different information for it. If there is no relation with some prior knowledge among different sources, sources are independent [37]. Otherwise, sources are dependent since one source may copy information from another source/sources [36].

B. Challenges in truth discovery problems

For truth discovery problems, it is difficult to discover the truth for an object among various information. It is more challenging when the dependency relationship among sources are uncertain which makes it harder for evaluating source reliability. In this survey, there are two kinds of challenges for truth discovery problems: object-based challenges and source-based challenges. For object-based challenges, the various properties of objects increase the difficulty of verifying the confidence of each observation. For source-based challenges, the relationship among different sources is hard to obtain, resulting in difficulty in estimating source reliability.

Object-based Challenges: Given the fact that the objects have various types of information, including dynamic [38], incomplete [39], unstructured [40], long-tail phenomenon [41], and large-amount properties [14], it is hard for us to validate the truth. In detail, for dynamic information, which implies that information varies with time, the truth of information may change in terms of different timestamps. It is difficult to verify the confidence of object information at different timestamps. For incomplete information, inaccurate extraction is very prevalent that sources only provide information for a subset of attributes about a given object [42]. Enough information for each attribute of the entities cannot be guaranteed. For long-tail information, the objects' information is provided by very few sources which is common in applications. It is difficult to evaluate the confidence of truth for objects with a little information. For objects with amount of properties, it is a challenge to infer the relationship among different attributes.

Source-based Challenges: Truth discovery is complicated for estimating source reliability and finding truth in terms of estimated sources. Since the information of an object collected from different sources may conflicts, it is difficult to determine which source is more reliable. In other words, it is hard for us to estimate the sources' reliability in the truth discovery problems. For sources with single valued objects, it is challenging to corroborate values from different observations. For sources with multiple valued objects, it is even more challenging to get all truths. Since sources and objects interact with each other, it is challenging to measure the relationship between sources and objects. It is more challenging when sources are dependent with each other. Due to erroneous values in the multi-source noisy data, it is hard to correctly link the object information [43]–[46], which results in insufficient evidence for the entities.

C. Existing truth discovery surveys

According to the significance of truth discovery problems, researchers have paid their attention on these problems. There are various existing studies in different scenarios, while there are only a few surveys which focus on truth discovery problems. The first truth discovery survey is introduced in [18], which defines the conceptions on object, source, observation, and truth. Fourteen methods on truth discovery are compared in terms of five aspects including input data, source reliability, object, claimed values, and output for various scenarios in [18] which lacks experimental analysis. To compare the experimental results for truth discovery problems on efficiency, usability, and repeatability, a comprehensive review of 12 state-of-the art algorithms is studied in [47] while the supervised methods are not considered. Some truth discovery methods are reviewed in [48] while it is specific for social sensing scenarios on disinformation detection. In addition, the task formulations, datasets, and natural language processing solutions for the task are reviewed and compared according to which the potentials and limitations are discussed [49] for fake news detection. Considering these limitations, there is an urgent need to review recent developments in truth discovery problems more comprehensively. Therefore, in this paper, we aim to address this gap by classifying the problems, addressing the advantages and disadvantages in different scenarios, comparing different methods from various perspectives, and providing applications for the truth discovery scenarios which are not previously discussed in the existing literature.

D. Main Contributions

The main contributions of this survey are summarized as follows.

- We systematically analyze the existing truth discovery studies and classify them from a well-developed fine-grained taxonomy based on the objects, sources, and observations. We analyzed the challenges and compared the advantages and disadvantages from different perspectives of objects, sources, and observations, respectively.
- The truth discovery methods are classified and compared by whether they have single or multiple values, independent or dependent sources, static or dynamic values, and supervisor information or not. We analyzed the challenges for these methods and compared advantages and disadvantages among these methods.
- We reviewed applications for truth discovery problems including websites, crowdsensing, data sensing, healthcare,

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

and knowledge base. Meanwhile, we analyzed the challenges when discovering truth for different applications. In addition, we listed the source codes and data sets and provided the link for further studies.

- We outline some promising future research directions in truth discovery problems from two different perspectives: problems and methods which will give references to the development of the community. From problem perspectives, we propose future studies from objects, sources, and observation with correlations or constraints. From methods perspectives, the scalability in truth discovery methods will be a hot point.

Papers selection. The papers we reviewed are high-quality papers selected from top journals and conferences, including IEEE Transactions, ACM transactions, ACM conferences, etc. The searching keywords include truth discovery, source reliability, data integration, crowd sensing, object correlation, etc. The published years are constrained to recent 30 years.

The rest of the paper is organized as follows. The fundamental problem classification on truth discovery is described in Section II where different kinds of truth discovery problems are introduced and compared. Truth discovery methods are surveyed and compared in Section III where the advantages and disadvantages are analyzed. Section IV demonstrates various applications with the source codes and datasets. Future directions are discussed in Section V, along with the conclusions in Section VI.

II. Fundamental Problem Description

A. Problem Description

Truth discovery problems try to infer truth labels (e.g., "True" or "False", 1 or 0, the values of objects) for the triples as the output. According to various scenarios, objects (i.e., the claims about things interested in) have different properties such as time property [14], [20]–[22], attribute property [19], [20], [50], and value property [11], [33], [51]. Observations have attributes and value properties, while sources have relations, attributes, and value properties. Therefore, there are lots of combinations for truth discovery problems. In this section, we summarize the general problem statement for truth discovery problems where the input contains some conflicting triples in the form of {source, object, observation}, where source denotes the location that the data originates, the object is an attribute of an entity where users are interested in, and the observation value set depicts the potential value set of an object claimed by a source. Sources can be regarded as views [26], web pages [52], and so on. Objects can be entities [53], facts [26], etc. And observations are the same as claims. Due to the dynamic, incomplete, and large amount properties on objects [19], [20], [22], sources [19], [20], [24], and observations [11], [29], [30], [35], object confidence and source reliability are affected with each other in truth discovery. The influence among objects and sources is evaluated by observations. In this paper, we define truth discovery problems from objects, sources, and observations. The relationship among them is described in Figure 1. In Figure 1, S_i^{tn} denote the source i at timestamp tn. O_j denotes the j^{th} object and O_{ij}^{tn} denotes the observation for object O_j from source S_i at timestamp tn. The number of sources and objects is denoted as a and b, respectively. A source can declare many object values and the object value can be declared by many sources at different timestamps. Among objects, they may have correlations. Therefore, these special characteristics on objects and sources generate different kinds of truth discovery problems.

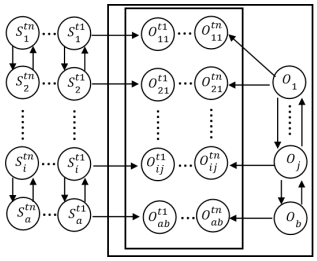


Fig. 1: Relationship among objects, sources, and observations

B. Object

According to different scenarios such as data integration, social media, and information retrieval, the descriptions of objects are various. For example, an object is a question for quiz answers in big data integration [54] while in mobile crowd sensing, an object is a report made by a smart phone. For any object, it is categorical or continuous in terms of different scenarios which contain three different values (single value, multiple values, and unknown values). Object correlations exist in many scenarios which can verify the attribute values of the same object. A probabilistic truth discovery model is constructed that considers not only source reliability but also object correlations which can increase the efficiency of truth discovery [38]. To propagate trustworthy information from crowd sensing, correlated objects have been observed by reliable users. Except object correlation, in real scenarios, constraints for objects can also help improve the efficiency for truth discovery methods.

1) Categorical Objects: Categorical objects exist in real scenarios which implies the values of attributes are discrete [11], [19]–[21], [33], [50], [51], [55]–[62]. A decision-theoretical framework is proposed to resolve numerical discrete value conflicts (various sources provide different values for the same object that should have a unique, specific numerical value) in a systematic manner [63]. The discrete nature of categorical objects confounds the direct application of existing multidimensional visualization techniques. Entropy, mutual information, and joint entropy are measured as a means of harnessing this discreteness to generate more effective visualizations [64]. The proposed system uses the heterogeneous data in both continuous data and categorical data given in [65] to overcome the scalable issues on applications for truth discovery [66]. For categorical objects, the challenge exists in the conflict values for the same object. It is more difficult when only few information is provided for the same object.

2) Continuous Object: The values of attributes are continuous on the time property for continuous objects [14], [22], [23], [62], [67]. For continuous objects, the information

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

may come sequentially, resulting in the truth objects and the reliability of sources evolving dynamically. The advantage for the continuous objects is that the truth can be updated. On the other hand, the drawback is that the truth should be evaluated at each timestamp which consumes more processing time.

3) Correlations: Correlations among entities are commonly observed in many applications [35], [38], [68], [69]. For example, nearby segments of the same road may have similar traffic conditions, and the weather conditions are similar with nearby locations. For objects with single attribute, an object (usually fact or view) may only contain an attribute in terms of different scenarios. How to find the truth values of attributes for objects is important in truth discovery problems.

For objects with multiple attributes, it can provide more information for objects. The advantage is that it can provide more information for objects with multiple attributes from different perspectives. However, there may be relations among different attributes which makes it challenging to evaluate the confidence of objects. In addition, the efficiency for truth discovery may be affected since it consumes more time to match attributes with the corresponding observation values.

Nowadays, attributes with multiple values have attracted researchers' attention. Since leveraging the unique features of the multi-truth problem that sources may provide partially correct values of a data item, more reasonable confidence scores to value sets were assigned. The object attributes with multiple values are well fitted to be used in many real-world application scenarios. However, it may have drawbacks to the number and quality of the values. Object correlations are usually ignored while they exist in many applications which could improve the performance in truth discovery problems.

4) Constraints: Constraints imply the range of the truth for objects. It is important to study constraints for various objects to infer truth while existing survey papers seldom consider it. For different scenarios in truth discovery, constraints are various [9], [34], [39], [70]–[77]. When there is a little information for an object or the objects can not satisfy the constraints, the output of truth discovery is unknown. The advantage for attributes with unknown values is that it provides more information for objects, while the disadvantage is that it is much more challenging to evaluate the truth of these attributes. The advantage of constraints for objects is that it improves the efficiency when discovering the truth among various values.

5) Comparison: There are different truth discovery problems with various objects, correlations, and constraints in different scenarios. Both categorical data [19], [20] and continuous data [14], [22], [62], [67] have been studied with single attribute while continuous data with multiple values is only studied in [23]. Correlations exist ubiquitously among objects. Especially, correlations are more helpful to discovery truth with only a few sources for objects, i.e., observations are only provided for a small portion of the objects. With some constraints in truth discovery problems, the range of truth is decreased by improving the accuracy and efficiency.

C. Sources

A source describes the place where the information about objects can be collected from [18]. According to the relationship among sources, sources are classified into dependent and independent sources. Sources are independent if they provide values of objects independently. On the contrary, if a source copy information from another source, they are dependent. For a given object, sources may contain the information about the object or not. Sources offer various data information on attributes and value properties for objects. When the sources only provide single type data information, it implies that objects included by the sources have an attribute. Otherwise, objects have multiple attributes with heterogeneous types.

Independent sources have been studied in [11], [19]–[21], [33], [41], [50], [51], [55]–[62]. Truth discovery problems with independent sources are more easily than dependent sources but are only suitable for limited scenarios. However, in real scenarios, dependent sources are very common which indicates that there is relationship among sources which usually copy data from other sources. Truth discovery problems with dependent sources have been studied in [24]–[28], [78]. For truth discovery problems with dependent sources, it is more complex than independent ones but suitable for more realistic scenarios such as crowd sensing, web, and social sensing. Sources with single data type have been studied in [11], [19]–[21], [33], [57]–[59], [61]. The advantage for sources with single data is that it is easy to match values with objects. However, it is limited to a few scenarios. Sources with heterogeneous data types have attracted researchers' attention [50], [51], [55], [56], [60], [62]. There are many practical scenarios where sources have heterogeneous data. But however, it is difficult to match heterogeneous data to specific objects.

Comparison: Different types of sources are compared on relation property, attribute property, time property, and value property. For independent sources with single data, views were estimated [11], [19]–[21], [33], [57]–[59], [61]. For independent sources, there are more studies for categorical objects with single attributes. Only a few studies focus on continuous objects since the truth discovery problems would be much more difficult by considering the time property because the truth may change at different timestamps. In real scenarios, it is very common that some information of sources is copied from other sources while the dependence relation among sources was only studied in [24], [25], [78]. The relation among sources is independent or dependent. The challenge for independent sources is how to evaluate the reliability of sources. In real scenarios, it's common for sources to copy information from other sources. But it is difficult to measure the copy relation among sources. Similar to objects with single attribute and multiple attributes, sources with single data imply that objects in these sources have single attribute and sources with heterogenous data indicate that objects have multiple attributes. The challenge for sources with single data is to measure the dependency relationship among sources. It is even more difficult for sources with heterogenous data.

D. Observation

An observation, also known as a record, is a 3-tuple that consists of an object, a source, and its provided value [18]. According to the provided values, there are two different kinds of observations: binary and multi-array observations. For a binary observation, the provided value for the object is either true or false, whereas for a multi-array observation, the provided value for the object may be wrong or partial right. There are two kinds of relationships among the observations: correlation and constraints. When there is correlation among objects, it implies that the observations of different objects are related. In some scenarios, there are some constraints for objects where the provided values of the observations should be satisfied. A binary observation is used for binary claims [10], [29]–[32]. Scalar reliability of objects are used from multi-array observations in [11], [26], [33], [34].

Comparison: For binary observations, Bayesian Truth Discovery [31] and MLE [31] in Social Sensing [30] and Crowd Sourcing [10] are studied which consider the values either True or False. For multi-array observations, the provided values of objects may have different probabilities [26], various answers [33], conflict values [11], and partial true values [34]. For observations with correlation, different kinds of object correlation is considered such as [35], [38], [68], [69]. For the observations with constraints, there are different types such as attribute constraints [9], physical-constraint [34], spatio-temporal constraint [74], privacy constraints [75]–[77], data attack constraints [72], [73], mood sensitivity constraints [70], theme relevance constraints [71], and pattern discovery constraints [39]. The challenge for observations is how to use the conflict values to discovery the latent truth for the object with constraint information.

E. Problem classification

According to subsections II-B, II-C, and II-D, for objects, with different sources, the observations are various. If observation values are single, it is a single truth discovery problem. Otherwise, if observation values are multiple, it belongs to multiple truth discovery problems. When all the sources with observations are independent, it is an independent truth discovery problem. Otherwise, it is a dependent truth discovery problem. For the latent truth, if it does not change with time, it is a static truth discovery problem. Otherwise, it is a dynamic truth discovery problem. Combined with above classifications, if the observation values have supervised information, it is a supervised truth discovery problem. According to above classifications, there are different combinations for truth discovery problems.

III. Truth discovery methods and comparison

A variety of methodologies have been proposed in the literature based on the distinct characteristics of objects, sources, and observations. In this section, we categorize these methodologies on four dimensions according to problem classifications in subsection II-E: single/multi-truth, independent/dependent source, static/dynamic data, and unsupervised/supervised learning.

A. Single truth discovery methods for independent source

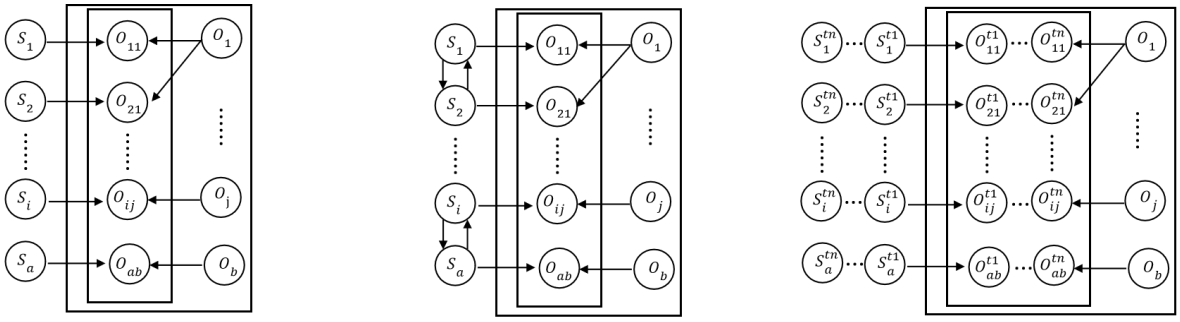
In real scenarios, to discover single truth for independent source with unsupervised learning, as Figure 2(a) shows where S_a represents the a^{th} source, O_b represents the b^{th} object and O_{ab} represents the observation from S_a of the O_b. Sources are independent. For each object, observations can be obtained from different sources which may conflict with each other, but there is only one truth from these observations. A commonly used multi-source aggregation strategy is voting or median. For the **voting method**, if multiple claims are mutually exclusive with each other, the one asserted by the most sources is selected. It may result in low accuracy when most claims are wrong. For the **median method**, the median value is selected as the truth value while for the **average method**, the average value is selected as the truth value. For **median and average methods** [79], it is difficult to measure when the values are not numerical.

A better approach to truth discovery is to determine truth in terms of the source quality. Advancements in the field have focused on differentiating sources based on their trustworthiness and proposing methodologies for assessing the quality of data sources [20], [31], [52], [80]–[82]. A commonly used principle is that sources which provide trustworthy information are more reliable, and the information from reliable sources is more trustworthy. In these approaches, the most trustworthy fact, i.e., the truth, is computed as a weighted voting or averaging among sources where more reliable ones have higher weights. For example, **Hub** method studies hubs and authorities in [80]. A mutually reinforcing relationship is exhibited by hubs and authorities for web (source) where a good hub is a web page that points to many influential authorities and a good authority is a web page that is pointed to by many influential hubs [52]. Each web page is given a hub and authority score. The hub score is calculated by the sum of the authority of linked pages and authority score is computed by the sum of the hub scores which link to the pages.

Iteration based methods are used to discover the latent truth. **TruthFinder** [81] designs iterative procedures to compute the confidence of facts and the trustworthiness of sources, utilizing the relationships between sources and their claims. Pasternack et al. [16] introduce **Average-Log, Investment**, and **PooledInvestment** approaches to prevent the overestimation of trustworthiness for sources that make numerous claims. Galland et al. [20] propose the **Cosine, 2-Estimates** and **3-Estimates** methods corresponding to different complexity of an underlying probabilistic model. According to the experiments on real-world data sets [20], we have observed that 3-Estimates outperforms 2-Estimates and Cosine. However, none of these three methods perform significantly better than random guessing when the data set has few conflicts per data item and a large number of non reliable sources (pessimistic scenarios).

The methods above are classical and heuristic, and further, probability graph methods are used. D. Wang et al. [31] offer the first optimal solution **Regular EM** (Expectation Maximum) to truth discovery problem. In maximum likelihood estimation, truth is attained by solving an expectation

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677



(a) single truth for independent source **(b) single truth for dependent source** **(c) single truth for dynamic source**

Fig. 2: Single truth discovery for independent source, dependent source, and dynamic source

maximization problem that returns the best guess regarding the correctness of each measurement. Based on the probability graph, **SUTD** (Scalable Uncertainty-Aware Truth Discovery), an analytical framework is developed which uses EM algorithm in social sensing applications [82]. **GTM** (Gaussian Truth Model) [83] is specially designed for handling numerical data. Based on Bayesian probabilistic models, which leverages the characteristics of numerical data in a principled way. **HA-EM** (Hardness-Aware Expectation Maximization) is the first proposed to derive an optimal solution for hardness-aware truth discovery problems in [84]. Neither of these methods takes into account the interrelationship of data. However, GTM and HA-EM conduct a more in-depth analysis of the data's characteristics with complex data structures.

The truth discovery problem has further refinement based on neural network models and limited resources. Jermaine Marshall et al. [85] first introduce **NN** (neural network) to truth discovery problem for accurate capture of the complex source-claim relational dependency with interpretability requirements. **ETCIBoot** (Estimating Truth and Confidence Interval via Bootstrapping) is proposed in [86] for long-tail phenomenon with highly imbalanced datasets. **FTS** [87] is proposed for multi-source sparse data which utilizes the False Rate, True Rate, and Silent Rate to measure source quality and the Probability Graphical Model is used to model truth and source quality which is measured through null and real data while these methods may encounter difficulties for sources with unknown or fluctuating reliability metrics. **CTD** (Constrained Truth Discovery) is proposed to incorporate denial constraints into the process of truth discovery [9] for large-scale data sets. A **STDM** (Seeking the Truth in a Decentralized Manner) is studied to offer a decentralized design with limited resources [88] while the absence of centralized oversight could potentially make the system more susceptible to targeted manipulations.

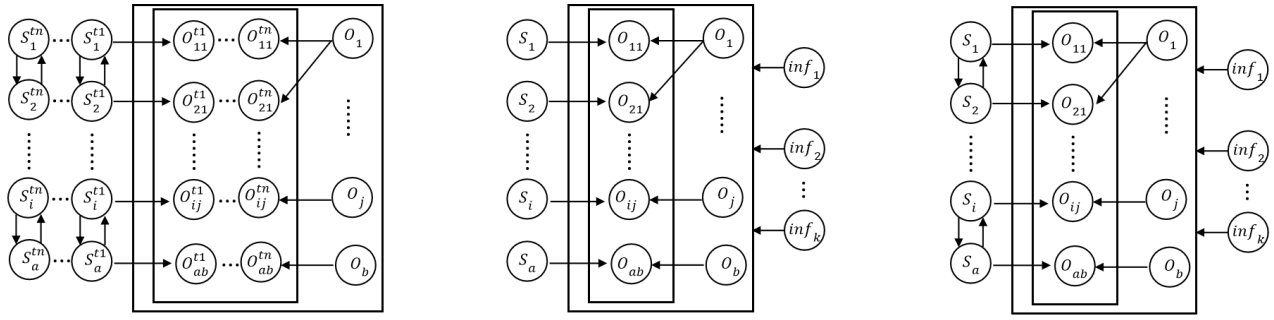
B. Single truth discovery methods for dependent source

The behavior of copying between sources is common in practice [24], especially in social sensing, where participants know each other's outputs and may occasionally copy from others which results in the spread of data contamination, as bad data (e.g., rumors) can be copied from one source to another [89]. The relation among source, object and observation is shown in Figure 2(b), where the arrows between S_a and S_i represents there is a dependence between S_a and S_i and the dependence is bidirectional. The main principle for copy detection is that if some sources make many common mistakes, they are not likely to be independent with each other. However, this principle proves to be ineffective in scenarios where certain sources replicate information from reliable sources, thereby posing a significant challenge in copy detection [18].

Several studies have been proposed to address the issue of source dependence with single-truth, static-data and unsupervised assumption [26], [90]–[92]. AccuSim (Accuracy similarity) [26] addresses the challenge of identifying true values from conflicting information with numerous sources, some of which may engage in copying. Wang et al. [91] later introduced a source dependency model and embedded it into a tool called **Apollo** that improved the estimation of source reliability and the veracity of assertions by accounting for correlated errors (i.e., rumors). The EM algorithm is utilized in dependence detection. A novel **CEM-MutiF** [92] (Constrained Expectation Maximum likelihood with Multiple Features) is proposed to evaluate the veracity of observations in social sensing applications. Ma et al. [93] proposed an iterative EM algorithm for Truth Discovery, called **IEMTD** (Iterative Expectation Maximization algorithm for Truth Discovery), that jointly referred the reliability of agents and truth of events with dependent agents.

However, it is important to note that besides direct copying, there are more intricate copying relationships [18], including co-copying (where multiple sources copy from a single source) and transitive copying (where a source may copy from other sources indirectly). In order to identify such complex global copying relationships, **GLOBAL** [94] considers both the completeness and accuracy of sources, the existence and direction of direct copying relationship is detected. Source dependence can also occur within a group setting. **MSS** [90] (Multi-Source Sensing) reveals the latent group structure among dependent sources, and aggregate the information at the group level rather than from individual sources directly. This can prevent the collective intelligence from being inappropriately dominated by dependent sources.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677



(a) single truth for dependent dynamic source **(b) single truth for independent source with supervised information** **(c) single truth for dependent source with supervised information**

Fig. 3: Truth discovery for dependent dynamic source, independent source, and dependent source with supervised learning

C. Single truth discovery methods for dynamic source

In real-word applications, sources are always dynamic (i.e., their content changes over time), as Figure 2(c) shows, raising the following challenges [95]. The first challenge arises from that the observation may not be effective in capturing changes in a timely manner when sources update their data frequently. The second challenge involves the dynamic nature of source quality, as it may change over time. Furthermore, the integration quality of sources with the specific objects vary over time. In Figure 2(c), S_a^{t1} represents the source a at $t1$ timestamp, O_b represents the b^{th} object and O_{ab}^{t1} represents the observation from S_a of the O_b at $t1$ timestamp. The observation O_{ij}^{tn} changes dynamically at different timestamps.

To solve these problems, several methods [14], [62], [95]–[98] are proposed based on the single-truth, independent-source, dynamic-data, and unsupervised learning on these objects. **Gain-Cost Models** [95] study the problem of source selection considering dynamic data sources whose content changes over time and define a set of time-dependent metrics, including coverage, freshness, and accuracy, to characterize the quality of integrated data. For sequential information,**DynaTD** [14] (Dynamic Truth Discovery) investigates the temporal relations among both object truths and source reliability, and proposes an incremental truth discovery framework that can dynamically update object truths and source weights upon the arrival of new data. Zhao et al. introduce the probabilistic model **StreamTD** [62] (Stream data Truth Discovery), which addresses the challenge of truth discovery over data streams while the sequential Bayesian approach may not optimize the likelihood function due to model assumption discrepancies.

The EM algorithm has proven to be an effective approach for addressing dynamic truth discovery problems [96]–[98]. By utilizing the EM algorithm, researchers can iteratively estimate the latent variables and parameters of the truth discovery model. A streaming fact-finder **Recursive EM** is proposed in [96] that recursively updates previous estimates based on new data where the variables are presumed to be dichotomous in nature. Pal et al. [97] propose a formal approach that models the historical updates of the real-world entity as a hidden semi-Markovian process (HSMM). Yao et al. introduced a recursive estimator **Recursive** for handling streaming social media data [98]. The online recursive estimator leveraged a batch EM framework by transferring posterior beliefs across time windows. The inputs are a binary while in many real applications, the types of inputs exceeds two. The dynamic truth discovery problem can be further refined with additional constraints. **TDCE** (Truth Discovery on Correlated Entities) [53] formulates the task of truth discovery on correlated entities in which both truths and user reliability are modeled as variables and propose both sequential and parallel solutions. Analogously, **CA-DTD** (Constraint-Aware Dynamic Truth Discovery) method is proposed in [34] which develops a constraint-aware Hidden Markov Model to predict the evolving truth of measured variables in terms of physical constraints. Moreover, it fuses observations between online social media and traditional news media to clean rumors, misinformation, and incomplete information. However, this method is highly dependent on physical constraints, necessitating that the data distribution be congruent with these physical limitations.

Several truth discovery algorithms have been proposed to ensure efficient and accurate real-time truth discovery [32], [99], [100]. Li et al. present a novel truth discovery framework named **ASRA** (Adaptive Source Reliability Assessment) [100] specifically designed for data streams. However, ASRA only computes source weights at certain timestamps, causing update delays. To solve the problem, Yang et al. [99] introduce an iterative-based truth discovery method called **DSWC** (Dynamic Source Weight Computation), which has a more flexible source weight evolution condition to limit the unit error. However, the scalability of dynamic truth discovery methods has been overlooked in aforementioned algorithms. In order to address this issue, several studies [101]–[103] have specifically focused on developing scalable algorithms for dynamic truth discovery. A distributed framework, **SSTD** (Scalable Streaming Truth Discovery) [101] is implemented by Work Queue in HTCondor and incorporated the Hidden Markov Model (HMM) to effectively address the dynamic truth discovery challenge. Instead of categorical data, a new **POLICE** (Probabilistic model for real valued sensing data on Correlated Entities) method is proposed to study the data trend for a period in [102]. Nevertheless, the efficiency of this method may be compromised when dealing with large-scale datasets. In the context of quantitative crowdsourcing applications that deal with big or streaming data, Ouyang et

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

al. [103] put forward parallel and streaming truth discovery algorithms These algorithms aim to achieve efficient and scalable truth discovery by decomposing large-scale truth discovery problems and leveraging the online EM algorithm. The proposed approaches allow for effective handling of large volumes of data and enables real-time truth discovery while for further studies, it is necessary to explore alternative parallel processing frameworks.

D. Single truth discovery methods for dependent dynamic source

Recently, researchers have become increasingly interested in the dynamic nature of source dependence, and the relationship among source, object and observation is shown in Figure 3(a), where the dependence can occur in any timestamp instead of remaining unchanged. In **COPYCEF** [24], the authors firstly explore the problem of finding true values and determining the copying relationship between sources. To increase the efficiency of dynamic dependence detection, **EvolvT** [78] is a dynamic truth discovery method designed for numerical data which incorporates three crucial aspects of dynamic truth discovery into a unified model: truth transition regularity, source quality, and source dependency. However, the scalability of this method requires further investigation. To decrease the misinformation spread, solve the data sparsity, and increase the scalability, a **SRTD** (Scalable Robust Truth Discovery) method was proposed in [104]. It considers various source behaviors, claims, and source dependency relationship. Similarly, a distributed framework, **SSTD** (Scalable Streaming Truth Discovery) is implemented by Work Queue in [101].

E. Single truth discovery methods for independent source with supervised learning

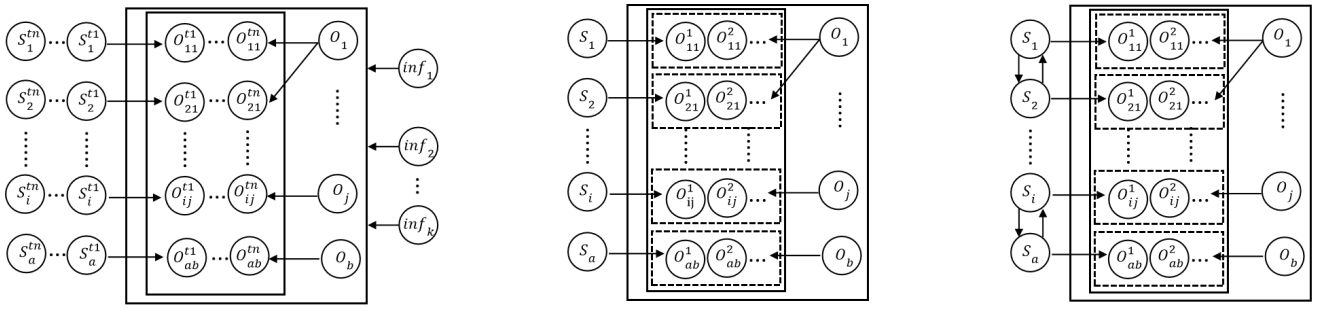
Truth discovery from arbitrary open online sources is a complicated problem due to the uncertainty regarding to source reliability and object confidence. If there is only a little information for truth objects, it is essential to fully exploit the information to find high accurate truth for all objects. The relationship among source, object, observation, and supervised information is shown in Figure 3(b) where the inf_k represents the k^{th} supervised information with dependent sources. Some methods are iterated with supervised information where the objects are divided according to whether the values of objects are known or unknown [105], [106]. To find true values with ground truth data, **SSTF** (Semi-Supervised Truth Finder) is proposed where the unlabeled set is updated until the condition is satisfied. However, this method is predicated among three fundamental relationships (facts similarity, mutual exclusivity, consistency), but its efficacy may be less pronounced with more complex datasets. By considering an electronic medical record set and a question answer pair set, a **MedTruth** method was proposed in [106]. For the medical knowledge condition discovery task, electronic medical record data and question answer data are leveraged to enrich the knowledge graph with knowledge triple condition information. Apart from these, some utilize a subset of labeled truth to semi-supervise the process of source reliability estimation and truth computation. For instance, **SOLARIS** [107], an online data fusion system addresses challenges in its development include maintaining vote counts for each value, computing expected probabilities, maximum and minimum probabilities of a value being true, determining termination conditions, and ordering sources for early termination and output of accurate answers. In [108], a **BCCTD** (Bayesian co-clustering truth discovery) approach is studied to utilize a small portion of ground truth data to aggregate user-contributed observations. However, this method relies on the reliability matrix for data modeling with an accurate supervised dataset.

In [106], [108]–[112], different scenarios using supervised learning for truth discovery are studied. In [106], the reference sources are regarded as the supervised information. In [108], [109], the labeled objects are jointly used to estimate the resource ability and correct claims. For the specific **Twitter platform**, Castillo et al. [110] employ a supervised learning approach, where it constructed a dataset specifically designed to investigate credibility. Similarly, for spammer detection on **Twitter**, Benevenuto et al. [111] set up a substantial dataset encompassing over 54 million users, 1.9 billion links, and nearly 1.8 billion tweets are collected. By leveraging tweets associated with three prominent trending topics from 2009, a sizable labeled dataset comprising users categorized as spammers and non-spammers is curated manually. The aforementioned research primarily focuses on Twitter, while Yang et al. [112] direct their attention towards Sina Weibo. Features aimed at client-side programs and event localization are analyzed which ignore geolocation information.

F. Single truth discovery methods for dependent source with supervised learning

In recent years, there has been a growing trend among researchers to tackle the issue of source dependence in supervised learning frameworks, as Figure 3(c) shows where the dependent sources have supervised information. The **SSEM** (Semi-supervised EM) algorithm is proposed to update the source-claim matrix and estimates the truth in [109]. However, this model enables the greedy algorithm to perform well while it does not work well when networks are sparse. Some researchers use training (from labeled data) to understand how content features correlate with veracity on social media. In [113], two variations of the algorithm, namely **CEM** (constrained expectation maximization algorithm) and **CEM-Jaccard** (constrained expectation maximization algorithm with Jaccard distance), are evaluated, both incorporating prior information on the number of independent sources to refine the probability estimates of latent truth variables. In addition, the CEM-Jaccard algorithm is more robust than the CEM algorithm. Some studies tend to use semi-supervised graph neural networks for truth discovery in social sensing [114], [115]. For instance, in [115], a novel automatic fake news detection model was constructed based on geometric deep learning. Furthermore, this method necessitates further exploration in elucidating the decision-making processes of Graph Neural Networks.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677



(a) single truth for dynamic source with supervised information **(b) multiple truth for independent source** **(c) multiple truth for dependent source**

Fig. 4: Truth discovery for dynamic source with supervised learning, independent source, and dependent source

G. Single truth discovery for dynamic source with supervised learning

In practical applications, prior knowledge of dynamic data can indeed be valuable for truth discovery. The relationship for source, object, observation, and prior knowledge is shown in Figure 4(a), where the observation is dynamic at different timestamps with supervised learning. Several researchers [116]–[120] have focused on the problem of dynamic truth discovery within a supervised framework in a single-truth and independent-source scenario. These studies aim to leverage supervised learning techniques and incorporate dynamic aspects to improve the accuracy and effectiveness of truth discovery methods in real-world applications. An **OPSTD** (Optimization based Semi-supervised Truth Discovery) method in [22] is proposed in semi-supervised framework to estimate continuous object truths. **TweetCred** (Twitter Credibility Assessment) [116], is an online system designed to assess the credibility of content on Twitter in real-time. This web-based platform assigns a credibility rating ranging from 1 (indicating low credibility) to 7 (indicating high credibility) for each tweet within a user's Twitter timeline. The scoring process is facilitated by a semi-supervised automated ranking algorithm, being trained on human labels obtained through crowdsourcing.

However, the aforementioned truth detection work in real-time social media primarily revolves around the manual extraction of features or rules, necessitating a laborious and time-consuming manual effort. In order to address this issue, several approaches have been proposed [117], [118]. In [117], the authors presented a novel method that learns continuous representations of microblog events for identifying rumors based on recurrent neural networks (**RNN**) for learning the hidden representations. In [118], the authors identify characteristics of rumors by investigating three aspects of diffusion: temporal, structural, and linguistic. For many mobile applications, it is hard to find a distribution that exactly describes the noise in practice. **DeepSense** [119], a deep learning framework whose framework combines convolutional neural networks (CNNs) and RNNs to effectively utilize local interactions among similar mobile sensors. Nevertheless, if there is drastic change in the physical environment, DeepSense might need to be re-trained with new data. In the context of truth discovery in social sensing, Gupta et al. [120] present a decision tree classifier to investigate the temporal aspects, social reputation, and influence patterns associated with the dissemination of fake images on Twitter. The method holds further developmental potential in terms of expanding the scope of research.

H. Multiple truth discovery for independent source

It is common in real world that an object has one or more truths, such as the authors of a book or the side effects of a medicine. The relationship for source, object and observation in multiple truth discovery is shown in Figure4(b) where the observation values of an object is multiple with $O_{11}^1, O_{11}^2, \cdots$. In [51], **LTM** (Latent Truth Model) is the first approach designed to merge multi-valued attribute types which is a probabilistic graphical model to automatically infer true values and source quality without any supervision. It should be noted that, LTM's loss is either 0 (no error) or 1 (error), but in practice loss can be real-valued. In addition, a **MCQA** (Multi-Choice Question Aggregation) [121] system is presented to solve crowd-sourced problem by lightweight machining learning techniques for answers derived from participants' confidence. However, the inter-object relationships are often overlooked in the aforementioned methods for discovering multiple true values. **PRECRECCORR** [69] first combines object correction and multiply truths which uses Bayesian analysis to derive the truthfulness of a fact based on the quality of sources. Apart from this, a **probabilistic** [122] approach is proposed with improvement measures that incorporate the three implications in all stages of truth discovery process. **DTQ** (Deduce Tuples' Quality vectors) [123] proposes the concept of quality predicates to differentiate multiple true values from false values. However, source reliability may varies among different domains. **DART** (Domain-Aware Multi-Truth Discovery) [124] addresses the problem of discovering multi-truth on data provided by multiple sources in various domains and derives the domain expertise of each source based on the information richness of the sources.

I. Multiple truth discovery for dependent source

The multi-truth discovery problem has more complicated features, such as the involvement of finer-grained copy relationship among sources. It is possible for sources to engage in partial copy of claims from other sources, since sources always

provide more than one data for objects in multi-truth problems, leading to a heightened difficulty of detecting the inter-source dependencies, as Figure4 (c) shows where sources are dependent with multiple truth values for objects. Studies [36], [89] have investigated on the multi-truth and dependent source truth discovery problem without prior knowledge. In [36], the authors propose an integrated Bayesian model **MBM** (Multi-Truth Bayesian method), which comprehensively incorporates novel methods on source/value grouping, source dependency, and inter-value mutual exclusion. Furthermore, **DATE** [89] (Dependence and Accuracy based Truth Estimates) explores the propagation direction of the plagiarism phenomenon where truth values can have multiple representations. In **MTD-CC** (Multi-Truth Discovery with Candidate Correlations) method [125], a more general relationship among sources is considered where the sources correlation network is constructed. However, the computation of correlation for every pair data sources is needed, which reduces efficiency on large-scale datasets.

IV. Applications

With the growing importance, truth discovery has been used in many applications such as website [126]–[128], data sensing [72], [129], [130], security sensing [131], [132], social sensing [27], [34], [70], [84], [85], [133]–[135], healthcare [136]–[138] and knowledge based [126], [139]–[143] applications. For example, there are lots of misunderstanding information on the web. How to find the truth for the same object is necessary to users with multiple conflicting information from different sources [11]. In big data, information may come sequentially, which results in the truth of objects as well as the source reliability evolving dynamically [14]. Truth discovery in crowd sourced detection of spatial events with such ambiguous missing reports was studied in mobile computing [144]. An integrated solution is proposed to stimulate the strategic users to contribute more to truth discovery in the edge-assisted mobile crowd sensing [145].

A. Web related Applications

Since everyone can answer and comment on web, there are numerous but different claims to the same object. More and more people use the web to search for information, while most of them are not professional. It is difficult for users to distinguish truth from various information. To find trustworthiness of news content from multiple information sources with minimum misclassification error and retrieval time, the truth content discovery algorithm is proposed to produce trustworthy information with minimal time and multiple domain [127]. For a web application, the unstructured data is a big challenge for truth discovery problems. Because most claims on the web are not structured data, the text data is explored which has unique multifactorial and the diversity of word usages characteristics on the website in [126]. Similarly, to address the novel task of assessing the credibility of arbitrary claims made in natural-language text with unstructured claim, the sources are found automatically in news and social media, and then fed into a distantly supervised classifier for assessing the credibility of a claim [128]. How to implement systems with existing methods is vital for web users to find more reliable information with less time. To make information extraction more easily for web users, AllegatorTrack [146] and VERA [147] systems are constructed to provide more reliable information by combining truth discovery methods.

B. Crowdsensing Related Applications

Truth discovery is usually used to analyze conflicting data, and it traditionally estimates source quality only from the current task. Due to the openness of Mobile Crowdsensing (MCS), workers and sensors are of different qualities. Low quality sensors and workers may yield noisy data or even inaccurate data. Due to the lack of prior information about the quality of workers and the ground truth, how to select most suitable workers and sensors remains a great challenge to guarantee the quality of the sensing tasks. An outlier detection technique is studied to filter out anomalous data items [148]. A novel framework is proposed to choose the most reliable workers among available workers [15], [16]. The quality of workers is analyzed through two factors, i.e., bias and variance, which describe the continuous feature of sensing tasks. In MCS systems, the existing crowdsensing quality methods are mostly based on a central platform, which is not completely trusted in reality and results in fraud. To tackle this issue, crowdsensing quality methods are proposed [134], [135], [149]–[152]. Reliability Adaptive Truth Discovery method is proposed in [15]. A truthful incentive mechanism is proposed which pays for the workers by the workers' performance in the task just completed and the reputation [153].

One of the greatest challenges in spatial crowdsourcing is determined by the veracity of reports from multiple users about a particular event or phenomenon [72]–[76]. The collected data in MCS is usually sparsely distributed among a large sensing area, where each point of interest (PoI) may receive only a few sensing reports [154]. In this case, traditional truth discovery algorithms may not provide an accurate truth estimation for each PoI. To tackle this challenge, Holmes, is proposed to take advantage of the spatial correlations of the monitored phenomena by reusing each contributor's data for multiple nearby PoIs. An efficient truth discovery mechanism is proposed for crowdsensing tasks with temporal and spatial correlations [68]. A new method based on recursive Bayesian estimation from multiple reports of users is proposed to solve the difficulties of truth discovery in spatial-temporal tasks [74].

Vehicles in vehicle Crowdsensing (VCS) systems are equipped with the latest sensors, which work simultaneously for processing at end-user utility applications e.g., navigation, predictions, and traffic monitoring. Due to different source quality, the sensed data of vehicles may vary from the ground truth. Because of the difference in driving behaviors and vehicle suspension systems, a major challenge in building such a system is how to aggregate conflicting sensory reports from multiple participating vehicles. All vehicle nodes upload their sole respective data to the cloud for computing which raises the need for reliability and privacy [76]. Truthfulness of sensing data is very important, as malicious vehicles may create inaccuracy in sensing results. RTSense is studied to

enable trust-based crowdsensing services in [148]. To deal with conflicting estimation results generated from different drivers, novel aggregation methods are proposed in terms of VCS scenarios in [155]–[159].

C. Data sensing related Applications

In MCS, data may be poisoned by attacks which makes truth discovery problems much more challenging. Two types of data poisoning attacks, i.e., the availability attack and the target attack, are studied against a crowd sensing system empowered with the truth discovery mechanism [72]. A mobile crowdsensing system is subject to collusion attacks where a group of malicious participants collaboratively send fake information to mislead the system. Novel methods are proposed to improve data credibility in MCS to alleviate the attacks [72], [73], [130], [160]–[163]. In [129], the partially observable data poisoning attacks in crowdsensing systems are studied which show that even if the malicious workers only have access to local information, they can find effective data poisoning attack strategies to interfere with crowdsensing systems with the **TF** method. An efficient attack method is proposed to maximize the attack utility [164].

To ensure the authenticity and privacy of data, privacy-preserving truth discovery has attracted much attention since it can find reliable information among uneven quality of data collected from mobile users, while protecting both the confidentiality of users' raw sensory data and reliability. A novel cloud-enabled privacy-preserving truth discovery framework for crowd sensing systems is proposed, which can protect not only users' sensory data but also their reliability scores derived by the truth discovery approaches [75]. An efficient and privacy-preserving truth discovery approach is proposed in mobile crowd sensing systems, which can tolerate users offline at any stage, while guaranteeing practical efficiency and accuracy under working process [132]. Although truth discovery has been widely explored to boost aggregation accuracy, numerous security and privacy issues still need to be addressed. Existing schemes either do not guarantee the privacy of each participating user or fail to consider practical needs in crowdsensing systems. Two reliable and privacy-preserving truth discovery schemes are proposed for the above scenarios [131]. Privacy-preserving truth discovery methods are proposed to achieve the reliability and privacy of data [162], [165]–[169].

Social sensing has gradually become a new paradigm of crowd sourcing applications, due to the tremendous data from social media. However, most data collected from social media is imprecise and unreliable. To address the problem, the hardness of claims [84] and uncertainty of claims [82] are explored. Physical constraint awareness [34] and mood sensitive [70] constraints are exploited. The neural network approach [85] is used in the truth discovery about social sensing. Apart from this, there are also numerous data sources in social sensing which make truth discovery problems more challenging. In [170], the authors try to find critical sources to reduce the computational complexity, and in [133] unmanned aerial vehicles are integrated with social media for reliable disaster response.

D. Healthcare Applications

Health is always a critical topic in daily life, and various truth discovery methods have already been used in different healthcare systems. There are different claims about the similar health problems. In [136], the authors focus on the crowdsourced question answering website to help patients to extract medical knowledge. The challenge for truth discovery problems in healthcare systems is that various data types have different formats [171]. For patient healthcare monitoring systems, by applying body sensor networks, a data dependability verification framework is proposed by making decisions in three layers [172]. Since people always actively discuss about medial news online, some specific medical hot spots may conflict with each other, which can also be researched by truth discovery methods, such as the drug side-effects [137], and the pandemic of COVID-19 [138]. The privacy of patient details is improved for cancer prediction [173].

E. Knowledge Based Applications

Knowledge bases, such as DBpedia, Google Knowledge Graph, YAGO have attracted extensive interest over the last years. Large-scale knowledge base is crucial especially when the large language models have been a hot issue currently [174] since they enable the development of more sophisticated AI models and serve as "golden" benchmarks for evaluating their performance [175]–[177]. Truth discovery plays an important role in finding the truth among numerous and noisy data for the knowledge bases. Although some encouraging progress [139]–[141] have been made on truth discovery methods, the content of such knowledge bases still cannot distinguish truth from various information [142] because the ground truth is constantly being changing dynamically. According to multi-layer deep linguistic analysis, knowledge graphs are constructed to incorporate signals from multiple sources [143]. The major challenges of inferring true information on text data stem from the multifactorial property of text answers and the diversity of word usages (i.e., different words may have the same semantic meaning). To tackle these challenges, a novel truth discovery method is proposed, named "TextTruth", which jointly groups the keywords extracted from the answers of a specific question into multiple interpretable factors, and infers the trustworthiness of both answer factors and answer providers. After that, the answers to each question can be ranked based on the estimated trustworthiness of factors [126]. A resource description framework is combined with truth discovery methods in knowledge base in the form of rules which support a claim in [178].

F. Datasets

To study further in truth discovery problems easily, we summarize the source codes and datasets in terms of the methods mentioned in Section III in Table I. According to

https://wiki.dbpedia.org/.
http://www.google.com/insidesearch/ features/search/knowledge.html.
http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

TABLE I: Public code sources and datasets

Algorithm	Ref.	Code Link	Dataset	Link
TF	Yin et al. [11](2008)	https://github.com/IshitaTakeshi/TruthFinder	Book-Authors	×
AccuSim	Dong et al. [26](2009)	https://github.com/daqcri/DAFNA-EA	Book-Authors	×
2,3-estimates	Galland et al. [20](2010)	https://github.com/daqcri/DAFNA-EA	Book-Authors	×
Investment	Pasternack et al. [179](2010)	×	City-Population, Biographi, American-British-spelling	http://cogcomp.cs.illinois.edu/page/resource_view/16
LTM	Zhao et al. [51](2012)	https://github.com/yishangru/TruthDiscovery/tree/master/LTM	Book-Authors, Movie-Director	×
GTM	Zhao et al. [83](2012)	×	City-Population, Biographic	http://cogcomp.cs.illinois.edu/page/resource_view/16
Regular EM	Wang et al. [31](2012)	×	Twitter-feeds	http://apollo.cse.nd.edu
LCA	Pasternack et al. [180](2013)	https://github.com/daqcri/DAFNA-EA	Book-Authors, City-Population	http://cogcomp.cs.illinois.edu/page/resource_view/16
CRH	Li et al. [65](2014)	×	UCI-Adult, Bank	http://archive.ics.uci.edu/ml/datasets/Adult http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
CATD	Li et al. [33](2014)	×	City-Population,Biographic,Indoor-Floorplan,Game	http://cogcomp.cs.illinois.edu/page/resource_view/16
IB	Wang et al. [36]	×	Book-Authors, Movie-Director, Parent-Children	http://cogcomp.cs.illinois.edu/page/resource_view/16
HA-EM	Marshall et al. [84](2016)	×	Twitter-Feeds	http://apollo.cse.nd.edu
ETCIBoot	Xiao et al. [86](2016)	×	Indoor-Floorplan, Flight, SFV, Game	http://lunadong.com/fusionDataSets.htm http://www.nist.gov/tac/2011/
IATD	Zhang et al. [28](2016)	×	Flight, Stock	http://lunadong.com/fusionDataSets.htm.
Probabilistic	Wang et al. [122](2016)	×	Author, Biography, Movie	https://cogcomp.cs.illinois.edu/page/resource_view/16
SSTD	Zhang et al. [101](2017)	×	Twitter-Feeds	http://apollo.cse.nd.edu
NN	marshall et al. [85](2017)	×	Twitter-Feeds	http://apollo.cse.nd.edu
SRTD	Zhang et al. [34](2017)	×	Twitter-Feeds	http://apollo.cse.nd.edu
SUTD	Huang et al. [82](2017)	×	Twitter-Feeds	http://apollo.cse.nd.edu
DTQ	Xie et al. [123](2017)	×	Book, Flight	http://lunadong.com/fusionDataSets.htm
HLCR	Nakhaei et al. [181](2017)	×	Book	×
LTD-RBM	Broelemann et al. [37](2017)	×	Flight, Weather	http://lunadong.com/fusionDataSets.htm
MN	Li et al. [182](2017)	×	Stock, Flight	http://lunadong.com/fusionDataSets.htm
OPSTD	Yang et al. [22](2018)	×	Weather, Gas Price, Stock	http://lunadong.com/fusionDataSets.htm
CASE	Lyu et al. [183](2019)	http://github.com/Sunshine1007472173/CASE	Weather, Biographies	http://cogcomp.cs.illinois.edu/page/resource_view/16
DATE	Jiang et al. [89](2019)	×	Qatar-Living-Forum, Auction	http://alt.qcri.org/semeval2015/task3 http://www.modelingonlineauctions.com/datasets
PTDCorr	Yang et al. [38](2019)	×	Gas Price, Weather	×
CTD	Ye et al. [9](2022)	×	Restaurant, Flight	http://lunadong.com/fusionDataSets.htm
BCCTD	Du et al. [184](2019)	×	BlueBird, Natural-Languag-Processing	http://ir.ischool.utexas.edu/square/index.html
STDM	Fu et al. [88](2021)	×	Flight, Stock	http://snap.stanford.edu/data/ http://lunadong.com/fusionDataSets.htm
HM	Ye et al. [185](2021)	https://github.com/lwb515/Deep-Truth-Discovery-for-Pattern-Based-Fact-Extraction	Drug, Article	http://curtis.ml.cmu.edu/gnat/biomed/. https://github.com/RaRe-Technologies/gensim
BM	Yang et al. [186](2021)	https://github.com/yitianhoulai/ART	IMDB, Sentiment-Polarity, Weather-Sentiment	https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/L5TTRW https://github.com/yitianhoulai/ART

× implies the link is not provided.

Table I, for **TF, AccuSim, 2,3-estimates, Investment, LTM, LCA, HM**, and **BM** algorithms, the code sources are provided which makes it easy to compare further studies in future. For datasets, we tried our best to collect the dataset URL information. Twitter-Feeds, Book-Authors, and City-Population are common datasets used for truth discovery problems. The detail for the dataset information is described in Table I.

V. Future directions

Our comprehensive review of the literature has revealed the significant challenges and developments on truth discovery research so far. We observe that truth discovery has been combined with crowd sensing, security related applications, while only few studies focus on the scalability of truth discovery methods. In this section, we summarize solved and unsolved challenges, and outline several promising prospective research directions, which we believe are critical to the further development of the filed.

A. Solved Challenges and Unsolved Challenges

In truth discovery, significant improvement has been made in addressing a number of challenges that were previously considered intractable. The development of advanced algorithms has facilitated the accurate identification of true assertions from conflicting information sources due to the vast and heterogeneous nature of modern data. Researchers have successfully tackled the issue of data sparsity [33], [34], [86], [87] that can infer truth from limited evidence. Furthermore, the integration of Bayesian networks [29], [36], [51], [108], [184], [186] has allowed for the robust estimation of source reliability.

However, there are still a multitude of challenges which remain to be addressed in truth discovery. These challenges encompass the scalability of truth discovery methodologies, the interconnectedness of attribute value assertions, and the complexities with multiple truths. Furthermore, the exploration of truth discovery methods, particularly those predicated on distributed algorithms and neural network algorithms can be studied in further.

B. Problem Level

In the problem level, there are still many directions for future research. First, objects with multiple values are very common in web, crowd sensing, data sensing, healthcare systems, and knowledge base. Therefore, more studies should be paid attention to objects with multiple values in future. When discussing truth discovery problems, the data format has usually been processed with formal formats while the raw data is usually unstructured from real-world applications. Therefore, how to process unstructured data plays a vital role in truth discovery problems. However, most of the existing studies ignore it. For observations, most of the existing studies ignore the constraints for objects while the constraints of objects could improve the efficiency on the procedure for truth discovery problems.

Objects: Although single attribute with multiple values for objects has been studied [51], [57]–[59], there has been much space to improve. When considering objects with multiple values, the objects only have an attribute. However, for objects in real social applications or web, they usually have different attributes which have single or multiple values. How to construct a general model for objects with different attributes is vital but difficult in real scenarios. Since most studies focus on categorical objects [11], [19]–[21], [33], [50], [51], [55]–[62], The truth discovery problems related to continuous objects are rare but important for weather scenarios and social media applications. In weather prediction applications, continuous objects such as the temperature is dynamic. In addition, in social media applications, the values of objects may change with time. Even continuous objects with multiple values are urgent in future while no existing studies consider this kind of truth discovery problems.

Correlation and constraints among attributes could improve accuracy for truth discovery problems. However, it is difficult to measure that whether these attributes are correlated or not. According to the correlations among attributes for objects, constraints can be obtained. In real scenarios, correlation and constraints among multiple attributes for objects are very common while only a few studies focus on the property. Therefore, in future, more attention should be paid to correlation and constraints among multiple attributes for objects.

Sources: Unstructured data is very common in various scenarios, such as the data on the social media and web applications. Current truth discovery problems mostly consider structured data as input for the objects and source information. However, in the common scenarios, data is usually unstructured which will make the truth discovery problems much more difficult. With unstructured data, the attribute information for objects are varied. How to transform these unstructured data into structured data is important in future. And how to analyze unstructured data is urgent in future work. The information from various sources may change with time going by which results in dynamic truth for the same object. When it comes to sources, existing studies focus on independent relationship among sources. However, in real scenarios such as social applications and crowd sensing applications, sources are usually dependent with each other. Therefore, more attention should be paid for dependent sources in future.

Observations: In different scenarios, the attributes of observations may be single value or multi-value. Most of existing studies focus on single attribute with single value. For attributes of observations with multi-values, the relationship among these values is ignored in recent studies. For attributes with multi-values, the complexity of the truth discovery problems is increased. The confidence value for objects and source reliability will be more complex in future studies to increase the accuracy of truth discovery. For observations with different attributes, existing studies usually ignore the correlation among attributes. However, the correlation matters in truth discovery problems which can improve the efficiency. For observations with constraints, existing studies usually ignore the constraints while they can provide more information for objects. By full use of the provided information, the efficiency for truth discovery problems can be improved.

C. Method Level

For the future directions of the method, according to Section III, there are many truth discovery methods for scenarios with single-valued, independent, or unsupervised learning, . Dynamic and dependency methods for truth discovery problems are popular in current research but remain relatively scarce, particularly for scenarios with dynamic and supervised learning. Additionally, dynamic truth discovery methods for multi-valued scenarios are also limited. In future, research including exploring dynamic dependencies and advancing supervised methods can be studied in dynamic contexts. By addressing these research gaps, we can enhance the applicability and effectiveness of truth discovery methods, catering to the requirements of dynamic, dependency-driven, and multi-valued scenarios commonly encountered in practical applications.

For iteration methods, the confidence of objects and source reliability are updated jointly. Although most existing methods for truth discovery are inspired by some heuristic ideas, local optimal solutions are achieved which have no quality guarantee on global optimality. Therefore, in further studies, more complex heuristic ideas can be used to avoid local optima and improve the global optima. To jump from local optima, there are lots of strategies such as simulated annealing [187], genetic algorithm [188], variable neighborhood search [189] and so on. How to combine these complex heuristic frameworks with truth discovery procedures is important to improve the accuracy and efficiency for truth discovery problems.

For probabilistic graphical model (PGM) based methods, the prior knowledge is used to improve the accuracy of truth discovery problems. The joint probability functions are calculated when there is no supervised learning. For truth discovery problems with supervised learning, how to combine the supervised learning to a probability graph is challenging but necessary. With the complete likelihood functions, existing methods focus on Markov Chain Monte Carlo and Expected Maximization framework to obtain the maximum probability of the truth. For various sources, the source reliability for different objects is different. In future studies, how to evaluate the bias information is important to improve the accuracy.

In the previous truth discovery methods, the relationship between source reliability and claim truthfulness can be represented by simplified functions (e.g., linear, quadratic and binomial). This assumption will result in local optima of truth discovery results because the extracted relational dependency between sources and claims is often unknown a prior. However, a neural network approach can learn the complex relational dependency better than the previous truth discovery methods. How to use neural networks in truth discovery methods to estimate the reliability of the values of objects is important which can use different representations to improve the reliability and dependence relationship among sources. Neural networks are often regarded as "black boxes", making it difficult to interpret their decision-making processes. This lack of transparency can be a significant limitation, especially in applications where explainability is critical. Additionally, the training of these models requires substantial computational resources, which may not be readily available. There is also a risk of overfitting, where models may become too tailored to the training data, potentially reducing their effectiveness on new, unseen data.

Since existing truth discovery methods are designed as sequential algorithms which is suitable for large-scale social sensing events and large language models, some distributed frameworks are used in truth discovery to solve these problems. However, it is urgent to discover the truth information quickly in these applications. Distributed frameworks must manage the communication overhead between nodes, which can impact performance, and address privacy concerns, as data is often distributed across different locations. The characteristics require specialized knowledge to build and train effectively. Therefore, it is important to combine the existing distributed framework with truth discovery methods to improve the efficiency. The scalability of these methods would influence the realistic applications.

The application of Graph Signal Processing (GSP) and Graph Neural Networks (GNNs) in truth discovery is an emerging area with significant potential. GSP offers a structured approach to analyze and process signals across graph-structured data [190], [191], which can be instrumental in identifying the most accurate information by leveraging the relational context within datasets. GNNs, with their ability to learn from graph-structured data, can effectively capture complex patterns and dependencies, enhancing the accuracy of truth discovery tasks [192], [193]. In the future, the integration of these methodologies may focus on developing models resilient to noise and adversarial influences, thereby fortifying the reliability of truth discovery for complex and dynamic data landscapes.

VI. Conclusion

In this paper, we reviewed the problems, methods, applications, code sources, datasets, and opportunities for truth discovery problems. We provide the general statement of truth discovery problems from objects, sources, and observations. According to these perspectives, different kinds of truth discovery problems are classified and the interaction among these is discussed. The challenges from these three perspectives are discussed in terms of the accuracy, efficiency, and scalability metrics. Different methods are compared based on the problem classification. According to the above methods, different applications such as the Web, crowdsensing, data sensing, healthcare, and knowledge base are reviewed. The code sources and datasets are provided based on these mentioned algorithms. The opportunities are discussed in terms of the problem and method level. In the future, more work should be studied to improve accuracy, efficiency, and scalability.

Acknowledgement

This work is supported by the National Key Research and Development Program of China (No.2022YFF0902800), the Natural Science Foundation of Jiangsu Province (No.BK20220803), and the National Natural Science Foundation of China (No.62302095). Michael Sheng's work is partially supported by the Australian Research Council Discovery Projects DP200102298 and DP230100233.

References

[1] T.-M. Choi, H. K. Chan, and X. Yue, "Recent development in big data analytics for business operations and risk management," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 81–92, 2017.

[2] T. Li, G. Kou, Y. Peng, and P. S. Yu, "An integrated cluster detection, optimization, and interpretation approach for financial data," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 848–13 861, 2022.

[3] Z. Yuan, H. Chen, T. Li, B. Sang, and S. Wang, "Outlier detection based on fuzzy rough granules in mixed attribute data," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8399–8412, 2022.

[4] N. D. Doulamis, A. D. Doulamis, P. Kokkinos, and E. M. Varvarigos, "Event detection in twitter microblogging," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 2810–2824, 2016.

[5] D. Wu and J. R. Birge, "Risk intelligence in big data era: A review and introduction to special issue," *IEEE Transactions on Cybernetics*, vol. 46, no. 8, pp. 1718–1720, 2016.

[6] K. Schouten, O. van der Weijde, F. Frasincar, and R. Dekker, "Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data," *IEEE Transactions on Cybernetics*, vol. 48, no. 4, pp. 1263–1275, 2018.

[7] X. Gong, T. Zhang, C. L. P. Chen, and Z. Liu, "Research review for broad learning system: Algorithms, theory, and applications," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 8922–8950, 2022.

[8] X. S. Fang, Q. Z. Sheng, X. Wang, W. E. Zhang, A. H. Ngu, and J. Yang, "From appearance to essence: Comparing truth discovery methods without using ground truth," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 6, pp. 1–24, 2020.

[9] C. Ye, H. Wang, K. Zheng, Y. Kong, R. Zhu, J. Gao, and J. Li, "Constrained truth discovery," *IEEE Trans. on Knowl. and Data Eng.*, vol. 34, no. 1, p. 205–218, 2022.

[10] P. Sun, Z. Wang, L. Wu, Y. Feng, X. Pang, H. Qi, and Z. Wang, "Towards personalized privacy-preserving incentive for truth discovery in mobile crowdsensing systems," *IEEE Transactions on Mobile Computing*, vol. 21, no. 1, pp. 352–365, 2022.

[11] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[12] Y. Zheng, H. Duan, and C. Wang, "Learning the truth privately and confidently: Encrypted confidence-aware truth discovery in mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2475–2489, 2018.

[13] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" *Proc. VLDB Endow.*, vol. 6, no. 2, p. 97–108, 2012.

[14] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han, "On the discovery of evolving truth," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 675–684.

[15] X. Gao, S. Chen, and G. Chen, "Mab-based reinforced worker selection framework for budgeted spatial crowdsensing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1303–1316, 2022.

[16] J. An, J. Cheng, X. Gui, W. Zhang, D. Liang, R. Gui, L. Jiang, and D. Liao, "A lightweight blockchain-based model for data quality assessment in crowdsensing," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 84–97, 2020.

[17] F. Li, M. L. Lee, and W. Hsu, "Entity profiling with varying source reliabilities," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1146–1155.

[18] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *ACM Sigkdd Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

[19] A. Marian and M. Wu, "Corroborating information from web sources." *IEEE Data Eng. Bull.*, vol. 34, no. 3, pp. 11–17, 2011.

[20] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 131–140.

[21] J. Pasternack and D. Roth, "Making better informed trust decisions with generalized fact-finding," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, ser. IJCAI'11. AAAI Press, 2011, p. 2324–2329.

[22] Y. Yang, Q. Bai, and Q. Liu, "On the discovery of continuous truth: A semi-supervised approach with partial ground truths," in *International Conference on Web Information Systems Engineering*. Springer, 2018, pp. 424–438.

[23] F. Shi, Z. Qin, D. Wu, and J. A. McCann, "Effective truth discovery and fair reward distribution for mobile crowdsensing," *Pervasive and Mobile Computing*, vol. 51, pp. 88–103, 2018.

[24] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 562–573, 2009.

[25] D. Salah, B. Raddaoui, and M. Othmani, "Argumentative approach for the discovery truth: The role of source dependence," in *2019 International Conference on Internet of Things, Embedded Systems and Communications (IINTEC)*. IEEE, 2019, pp. 192–197.

[26] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: the role of source dependence," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 550–561, 2009.

[27] D. Wang, N. Vance, and C. Huang, "Who to select: Identifying critical sources in social sensing," *Knowledge-Based Systems*, vol. 145, pp. 98–108, 2018.

[28] H. Zhang, Q. Li, F. Ma, H. Xiao, Y. Li, J. Gao, and L. Su, "Influence-aware truth discovery," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 851–860.

[29] J. Yang, J. Wang, and W. P. Tay, "Using social network information in community-based bayesian truth discovery," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 3, pp. 525–537, 2019.

[30] C. Huang, D. Wang, and N. Chawla, "Towards time-sensitive truth discovery in social sensing applications," in *2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 2015, pp. 154–162.

[31] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 2012, pp. 233–244.

[32] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han, "Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 1986–1999, 2016.

[33] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *Proceedings of the VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.

[34] D. Y. Zhang, D. Wang, and Y. Zhang, "Constraint-aware dynamic truth discovery in big data social media sensing," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 57–66.

[35] D. Y. Zhang, J. Badilla, Y. Zhang, and D. Wang, "Towards reliable missing truth discovery in online social media sensing applications," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 143–150.

[36] X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, and X. Li, "An integrated bayesian approach for effective multi-truth discovery," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 493–502.

[37] K. Broelemann, T. Gottron, and G. Kasneci, "Ltd-rbm: Robust and fast latent truth discovery using restricted boltzmann machines," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 2017, pp. 143–146.

[38] Y. Yang, Q. Bai, and Q. Liu, "A probabilistic model for truth discovery with object correlations," *Knowledge-Based Systems*, vol. 165, pp. 360–373, 2019.

[39] C. Ye, H. Wang, T. Ma, J. Gao, H. Zhang, and J. Li, "Patternfinder: Pattern discovery for truth discovery," *Knowledge-Based Systems*, vol. 176, pp. 97–109, 2019.

[40] Z. Chen, J. Li, T. Li, T. Fan, C. Meng, C. Li, J. Kang, L. Chai, Y. Hao, Y. Tang, O. A. Al-Hartomy, S. Wageh, A. G. Al-sehemi, Z. Luo, J. Yu, Y. Shao, D. Li, S. Feng, W. J. Liu, Y. He, X.-P. Ma, Z. Xie, and H. Zhang, "A crispr/cas12a-empowered surface plasmon resonance platform for rapid and specific diagnosis of the omicron variant of sars-cov-2," *National Science Review*, vol. 9, 2022.

[41] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang, "Towards confidence interval estimation in truth discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 575–588, 2018.

[42] C. Ye, H. Wang, J. Li, H. Gao, and S. Cheng, "Crowdsourcing-enhanced missing values imputation based on bayesian network," in *International Conference on Database Systems for Advanced Applications*. Springer, 2016, pp. 67–81.

[43] E. Laxmi Lydia, T. Madhusudhana Rao, K. Vijaya Kumar, A. Krishna Mohan, and S. Lingamgunta, "Challenging data models and data confidentiality through "pay-as-you-go" approach entity resolution," in *Computer Networks, Big Data and IoT*. Springer, 2021, pp. 469–482.

[44] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 9, pp. 1537–1555, 2011.

[45] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 1, pp. 1–16, 2006.

[46] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," vol. 3, no. 1–2, 2010.

[47] D. A. Waguih and L. Berti-Equille, "Truth discovery algorithms: An experimental evaluation," *arXiv preprint arXiv:1409.6428*, 2014.

[48] F. Xu, V. S. Sheng, and M. Wang, "A unified perspective for disinformation detection and truth discovery in social sensing: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–33, 2021.

[49] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6086–6093.

[50] X. Xu, X. Liu, X. Liu, and Y. Sun, "Truth finder algorithm based on entity attributes for data conflict solution," *Journal of Systems Engineering and Electronics*, vol. 28, no. 3, pp. 617–626, 2017.

[51] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration," *Proc. VLDB Endow.*, vol. 5, no. 6, p. 550–561, 2012.

[52] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[53] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng, "Truth discovery on crowd sensing of correlated entities," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 169–182.

[54] H. Tian, W. Sheng, H. Shen, and C. Wang, "Truth finding by reliability estimation on inconsistent entities for heterogeneous data sets," *Knowledge-Based Systems*, vol. 187, p. 104828, 2020.

[55] Y. Cao, W. Fan, and W. Yu, "Determining the relative accuracy of attributes," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 565–576.

[56] V. Beretta, S. Harispe, S. Ranwez, and I. Mougenot, "Truth selection for truth discovery models exploiting ordering relationship among values," *Knowledge-Based Systems*, vol. 159, pp. 298–308, 2018.

[57] X. S. Fang, "Truth discovery from conflicting multi-valued objects," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 711–715.

[58] X. S. Fang, Q. Z. Sheng, X. Wang, D. Chu, and A. H. Ngu, "Smartvote: a full-fledged graph-based model for multi-valued truth discovery," *World Wide Web*, vol. 22, no. 4, pp. 1855–1885, 2019.

[59] J. Feng, J. Chen, and J. Lu, "Novel approach for multi-valued truth discovery," in *International Conference on Ubiquitous Information Management and Communication*. Springer, 2019, pp. 1015–1028.

[60] H. Homayouni, S. Ghosh, I. Ray, and M. G. Kahn, "An interactive data quality test approach for constraint discovery and fault detection," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 200–205.

[61] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han, "Modeling truth existence in truth discovery," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1543–1552.

[62] Z. Zhao, J. Cheng, and W. Ng, "Truth discovery in data streams: A single-pass probabilistic approach," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 1589–1598.

[63] Z. Jiang, "A decision-theoretic framework for numerical attribute value reconciliation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 7, pp. 1153–1169, 2011.

[64] J. Alsakran, X. Huang, Y. Zhao, J. Yang, and K. Fast, "Using entropy-related measures in categorical data visualization," in *2014 IEEE Pacific Visualization Symposium*. IEEE, 2014, pp. 81–88.

[65] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 1187–1198.

[66] P. B. Thiyagaraj and A. Aloysius, "An approach for truth discovery by resolving the conflicts on categorical data," in *Journal of Physics: Conference Series*, vol. 1142, no. 1. IOP Publishing, 2018, p. 012014.

[67] L. Yao, L. Su, Q. Li, Y. Li, F. Ma, J. Gao, and A. Zhang, "Online truth discovery on time series data," in *Proceedings of the 2018 siam international conference on data mining*. SIAM, 2018, pp. 162–170.

[68] R. Wang, Y.-E. Sun, H. Huang, L. Lu, Y. Du, and D. Huang, "An efficient truth discovery mechanism for crowdsensing tasks with temporal and spatial correlations," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2019, pp. 501–508.

[69] R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava, "Fusing data with correlations," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 433–444.

[70] J. Marshall and D. Wang, "Mood-sensitive truth discovery for reliable recommendation systems in social sensing," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016, pp. 167–174.

[71] D. Wang, J. Marshall, and C. Huang, "Theme-relevant truth discovery on twitter: An estimation theoretic approach," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, 2016.

[72] J. Lin, D. Yang, K. Wu, J. Tang, and G. Xue, "A sybil-resistant truth discovery framework for mobile crowdsensing," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, pp. 871–880.

[73] Z. Huang, M. Pan, and Y. Gong, "Robust truth discovery against data poisoning in mobile crowdsensing," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[74] D. A. Garcia-Ulloa, L. Xiong, and V. Sunderam, "Truth discovery for spatio-temporal events from crowdsourced data," *Proceedings of the VLDB Endowment*, vol. 10, no. 11, pp. 1562–1573, 2017.

[75] C. Miao, W. Jiang, L. Su, Y. Li, S. Guo, Z. Qin, H. Xiao, J. Gao, and K. Ren, "Cloud-enabled privacy-preserving truth discovery in crowd sensing systems," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015, pp. 183–196.

[76] G. Xu, H. Li, S. Liu, M. Wen, and R. Lu, "Efficient and privacy-preserving truth discovery in mobile crowd sensing systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3854–3865, 2019.

[77] Y. Li, C. Miao, L. Su, J. Gao, Q. Li, B. Ding, Z. Qin, and K. Ren, "An efficient two-layer mechanism for privacy-preserving truth discovery," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1705–1714.

[78] S. Zhi, F. Yang, Z. Zhu, Q. Li, Z. Wang, and J. Han, "Dynamic truth discovery on numerical data," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 817–826.

[79] S. Ye, J. Wang, H. Fan, and Z. Zhang, "Probabilistic model for truth discovery with mean and median check framework," *Knowledge-Based Systems*, vol. 233, p. 107482, 2021.

[80] J. M. Kleinberg *et al.*, "Authoritative sources in a hyperlinked environment." in *SODA*, vol. 98. Citeseer, 1998, pp. 668–677.

[81] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2007, p. 1048–1052.

[82] C. Huang, D. Wang, and N. Chawla, "Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems," *IEEE Transactions on Big Data*, pp. 1–1, 2017.

[83] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," *Proc. of QDB*, 2012.

[84] J. Marshall, M. Syed, and D. Wang, "Hardness-aware truth discovery in social sensing applications," in *2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2016, pp. 143–152.

[85] J. Marshall, A. Argueta, and D. Wang, "A neural network approach for truth discovery in social sensing," in *2017 IEEE 14th international conference on mobile Ad Hoc and sensor systems (MASS)*. IEEE, 2017, pp. 343–347.

[86] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang, "Towards confidence in the truth: A bootstrapping based truth discovery approach," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1935–1944.

[87] J. Zhang, S. Wang, G. Wu, and L. Zhang, "A effective truth discovery algorithm with multi-source sparse data," in *International Conference on Computational Science*. Springer, 2018, pp. 434–442.

[88] L. Fu, J. Xu, S. Qu, Z. Xu, X. Wang, and G. Chen, "Seeking the truth in a decentralized manner," *IEEE/ACM Transactions on Networking*, vol. 29, no. 5, pp. 2296–2312, 2021.

[89] L. Jiang, X. Niu, J. Xu, D. Yang, and L. Xu, "Incentivizing the workers for truth discovery in crowdsourcing with copiers," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1286–1295.

[90] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang, "Mining collective intelligence in diverse groups," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 1041–1052.

[91] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti *et al.*, "Using humans as sensors: an estimation-theoretic perspective," in *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*. IEEE, 2014, pp. 35–46.

[92] H. Shao, D. Sun, S. Yao, L. Su, Z. Wang, D. Liu, S. Liu, L. Kaplan, and T. Abdelzaher, "Truth discovery with multi-modal data in social sensing," *IEEE Transactions on Computers*, vol. 70, no. 9, pp. 1325–1337, 2020.

[93] L. Ma, W. P. Tay, and G. Xiao, "Iterative expectation maximization for reliable social sensing with information flows," *Inf. Sci.*, vol. 501, no. C, p. 621–634, oct 2019.

[94] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *Proc. VLDB Endow.*, vol. 3, no. 1–2, p. 1358–1369, sep 2010.

[95] T. Rekatsinas, X. L. Dong, and D. Srivastava, "Characterizing and selecting fresh data sources," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 919–930.

[96] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems*, ser. ICDCS '13. USA: IEEE Computer Society, 2013, p. 530–539.

[97] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon, "Information integration over time in unreliable and uncertain environments," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 789–798.

[98] S. Yao, M. T. Amin, L. Su, S. Hu, S. Li, S. Wang, Y. Zhao, T. Abdelzaher, L. Kaplan, C. Aggarwal, and A. Yener, "Recursive ground truth estimator for social data streams," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, ser. IPSN '16. IEEE Press, 2016.

[99] Y. Yang, Q. Bai, and Q. Liu, "Dynamic source weight computation for truth inference over data streams," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2019, p. 277–285.

[100] T. Li, Y. Gu, X. Zhou, Q. Ma, and G. Yu, "An effective and efficient truth discovery framework over data streams," in *Proceedings of the 20th International Conference on Extending Database Technology*. Springer, 2017, pp. 180–191.

[101] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 966–976.

[102] M. Zhao and J. Jiao, "Police: An effective truth discovery method in intelligent crowd sensing," in *International Conference on Artificial Intelligence and Security*. Springer, 2020, pp. 384–398.

[103] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. J. Norman, "Parallel and streaming truth discovery in large-scale quantitative crowdsourcing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 10, p. 2984–2997, oct 2016.

[104] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Transactions on Big Data*, vol. 5, no. 2, pp. 195–208, 2018.

[105] X. Yin and W. Tan, "Semi-supervised truth discovery," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 217–226.

[106] Y. Deng, Y. Li, Y. Shen, N. Du, W. Fan, M. Yang, and K. Lei, "Medtruth: a semi-supervised approach to discovering knowledge condition information from multi-source medical data," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 719–728.

[107] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava, "Online data fusion," *Proc. VLDB Endow.*, vol. 4, no. 11, p. 932–943, aug 2011.

[108] Y. Du, Y.-E. Sun, H. Huang, L. Huang, H. Xu, Y. Bao, and H. Guo, "Bayesian co-clustering truth discovery for mobile crowd sensing systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1045–1057, 2020.

[109] H. Cui, T. Abdelzaher, and L. Kaplan, "A semi-supervised active-learning truth estimator for social networks," in *The World Wide Web Conference*, 2019, pp. 296–306.

[110] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 675–684.

[111] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.

[112] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ser. MDS '12. New York, NY, USA: Association for Computing Machinery, 2012.

[113] H. Shao, S. Yao, Y. Zhao, C. Zhang, J. Han, L. Kaplan, L. Su, and T. Abdelzaher, "A constrained maximum likelihood estimator for unguided social sensing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 2429–2437.

[114] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake news detection," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019, pp. 568–569.

[115] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *ArXiv*, vol. abs/1902.06673, 2019.

[116] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, *TweetCred: Real-Time Credibility Assessment of Content on Twitter*. Cham: Springer International Publishing, 2014, pp. 228–243.

[117] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 3818–3824.

[118] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 1103–1108.

[119] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 351–360.

[120] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13 Companion. New York, NY, USA: Association for Computing Machinery, 2013, p. 729–736.

[121] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, ser. AAAI'14. AAAI Press, 2014, p. 2946–2953.

[122] X. Wang, Q. Z. Sheng, L. Yao, X. Li, X. S. Fang, X. Xu, and B. Benatallah, "Truth discovery via exploiting implications from multi-source data," in *Proceedings of the 25th acm international on conference on information and knowledge management*, 2016, pp. 861–870.

[123] Z. Xie, Q. Liu, and Z. Bao, "Sifting truths from multiple low-quality data sources," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*. Springer, 2017, pp. 74–81.

[124] X. Lin and L. Chen, "Domain-aware multi-truth discovery from conflicting sources," *Proceedings of the VLDB Endowment*, vol. 11, no. 5, pp. 635–647, 2018.

[125] H. Huang, G. Fan, Y. Li, and N. Mu, "Multi-truth discovery with correlations of candidates in crowdsourcing systems," in *Collaborative Computing: Networking, Applications and Worksharing*, H. Gao and X. Wang, Eds. Cham: Springer International Publishing, 2021, pp. 18–32.

[126] H. Zhang, Y. Li, F. Ma, J. Gao, and L. Su, "Texttruth: an unsupervised approach to discover trustworthy information from multi-sourced text data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2729–2737.

[127] P. Solainayagi and R. Ponnusamy, "Trustworthy media news content retrieval from web using truth content discovery algorithm," *Cognitive Systems Research*, vol. 56, pp. 26–35, 2019.

[128] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 2173–2178.

[129] M. Li, Y. Sun, H. Lu, S. Maharjan, and Z. Tian, "Deep reinforcement learning for partially observable data poisoning attack in crowdsensing systems," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6266–6278, 2020.

[130] "Fidc: A framework for improving data credibility in mobile crowdsensing," *Computer Networks*, vol. 120, pp. 157–169, 2017.

[131] C. Zhang, L. Zhu, C. Xu, X. Liu, and K. Sharif, "Reliable and privacy-preserving truth discovery for mobile crowdsensing systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1245–1260, 2021.

[132] G. Xu, H. Li, S. Liu, M. Wen, and R. Lu, "Efficient and privacy-preserving truth discovery in mobile crowd sensing systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3854–3865, 2019.

[133] M. T. Rashid, D. Y. Zhang, and D. Wang, "Socialdrone: An integrated social media and drone sensing system for reliable disaster response," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020.

[134] F. Wu, S. Yang, Z. Zheng, S. Tang, and G. Chen, "Fine-grained user profiling for personalized task matching in mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 10, pp. 2961–2976, 2021.

[135] F. Restuccia, P. Ferraro, T. S. Sanders, S. Silvestri, S. K. Das, and G. L. Re, "First: A framework for optimizing information quality in mobile crowdsensing systems," *ACM Trans. Sen. Netw.*, vol. 15, no. 1, dec 2018.

[136] Y. Li, C. Liu, N. Du, W. Fan, Q. Li, J. Gao, C. Zhang, and H. Wu, "Extracting medical knowledge from crowdsourced question answering website," *IEEE Transactions on Big Data*, vol. 6, no. 2, pp. 309–321, 2016.

[137] F. Ma, C. Meng, H. Xiao, Q. Li, J. Gao, L. Su, and A. Zhang, "Unsupervised discovery of drug side-effects from heterogeneous data sources," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 967–976.

[138] M. T. Rashid and D. Wang, "Covidsens: a vision on reliable social sensing for covid-19," *Artificial intelligence review*, vol. 54, no. 1, pp. 1–25, 2021.

[139] E. Cao, D. Wang, J. Huang, and W. Hu, "Open knowledge enrichment for long-tail entities," in *Proceedings of The Web Conference 2020*, 2020, pp. 384–394.

[140] C. Lockard, X. L. Dong, A. Einolghozati, and P. Shiralkar, "Ceres: Distantly supervised relation extraction from the semi-structured web," *arXiv preprint arXiv:1804.04635*, 2018.

[141] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-based trust: Estimating the trustworthiness of web sources," *arXiv preprint arXiv:1502.03519*, 2015.

[142] D. Ritze, O. Lehmberg, Y. Oulabi, and C. Bizer, "Profiling the potential of web tables for augmenting cross-domain knowledge bases," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 251–261.

[143] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail, "The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1567–1578.

[144] R. W. Ouyang, M. Srivastava, A. Toniolo, and T. J. Norman, "Truth discovery in crowdsourced detection of spatial events," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 4, pp. 1047–1060, 2015.

[145] J. Xu, S. Yang, W. Lu, L. Xu, and D. Yang, "Incentivizing for truth discovery in edge-assisted large-scale mobile crowdsensing," *Sensors*, vol. 20, no. 3, 2020.

[146] D. A. Waguih, N. Goel, H. M. Hammady, and L. Berti-Equille, "Allegatortrack: Combining and reporting results of truth discovery from multi-source data," in *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 1440–1443.

[147] M. L. Ba, L. Berti-Equille, K. Shah, and H. M. Hammady, "Vera: A platform for veracity estimation over web data," in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 159–162.

[148] L. Zhu, C. Zhang, C. Xu, and K. Sharif, "Rtsense: Providing reliable trust-based crowdsensing services in cvcc," *IEEE Network*, vol. 32, no. 3, pp. 20–26, 2018.

[149] T. Luo, J. Huang, S. S. Kanhere, J. Zhang, and S. K. Das, "Improving iot data quality in mobile crowd sensing: A cross validation approach," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5651–5664, 2019.

[150] X. Gao, H. Huang, C. Liu, F. Wu, and G. Chen, "Quality inference based task assignment in mobile crowdsensing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3410–3423, 2021.

[151] S. Yang, F. Wu, S. Tang, X. Gao, B. Yang, and G. Chen, "On designing data quality-aware truth estimation and surplus sharing method for mobile crowdsensing," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 4, pp. 832–847, 2017.

[152] L. Yan and S. Yang, "Trust-aware truth discovery with long-term vehicle reputation for internet of vehicles crowdsensing," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, 2021, pp. 558–563.

[153] C. Xu, Y. Si, L. Zhu, C. Zhang, K. Sharif, and C. Zhang, "Pay as how you behave: A truthful incentive mechanism for mobile crowdsensing," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 053–10 063, 2019.

[154] Z. Li, S. Yang, F. Wu, X. Gao, and G. Chen, "Holmes: Tackling data sparsity for truth discovery in location-aware mobile crowdsensing," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2018, pp. 424–432.

[155] Y. Zhu, A. Gupta, S. Hu, W. Zhong, L. Su, and C. Qiao, "Driver behavior-aware parking availability crowdsensing system using truth discovery," *ACM Transactions on Sensor Networks (TOSN)*, vol. 17, no. 4, pp. 1–26, 2021.

[156] C. Zhang, L. Zhu, C. Xu, and K. Sharif, "Prvb: Achieving privacy-preserving and reliable vehicular crowdsensing via blockchain oracle," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 831–843, 2020.

[157] Z. Xu, W. Yang, Z. Xiong, J. Wang, and G. Liu, "Tpsense: a framework for event-reports trustworthiness evaluation in privacy-preserving vehicular crowdsensing systems," *Journal of Signal Processing Systems*, vol. 93, no. 2, pp. 209–219, 2021.

[158] F. Shi, D. Wu, D. I. Arkhipov, Q. Liu, A. C. Regan, and J. A. McCann, "Parkcrowd: Reliable crowdsensing for aggregation and dissemination of parking space information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4032–4044, 2018.

[159] C. Zhang, L. Zhu, C. Xu, K. Sharif, K. Ding, X. Liu, X. Du, and M. Guizani, "Tppr: A trust-based and privacy-preserving platoon recommendation scheme in vanet," *IEEE Transactions on Services Computing*, vol. 15, no. 2, pp. 806–818, 2022.

[160] X.-D. Wang, W.-Z. Meng, and Y.-N. Liu, "Lightweight privacy-preserving data aggregation protocol against internal attacks in smart grid," *Journal of Information Security and Applications*, vol. 55, p. 102628, 2020.

[161] Y. Zhao, X. Gong, F. Lin, and X. Chen, "Data poisoning attacks and defenses in dynamic crowdsourcing with online data quality learning," *IEEE Transactions on Mobile Computing*, 2021.

[162] H. Shen, Y. Liu, Z. Xia, and M. Zhang, "An efficient aggregation scheme resisting on malicious data mining attacks for smart grid," *Information Sciences*, vol. 526, pp. 289–300, 2020.

[163] I.-C. Hsieh and C.-T. Li, "Netfense: Adversarial defenses against privacy attacks on neural networks for graph data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[164] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, "Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 13–22.

[165] X. Tang, C. Wang, X. Yuan, and Q. Wang, "Non-interactive privacy-preserving truth discovery in crowd sensing applications," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1988–1996.

[166] C. Lv, T. Wang, C. Wang, F. Chen, and C. Zhao, "Espptd: An efficient slicing-based privacy-preserving truth discovery in mobile crowd sensing," *Knowledge-Based Systems*, vol. 229, p. 107349, 2021.

[167] X. Pang, Z. Wang, D. Liu, J. C. Lui, Q. Wang, and J. Ren, "Towards personalized privacy-preserving truth discovery over crowdsourced data streams," *IEEE/ACM Transactions on Networking*, 2021.

[168] H. Wu, L. Wang, K. Cheng, D. Yang, J. Tang, and G. Xue, "Privacy-enhanced and practical truth discovery in two-server mobile crowdsensing," *IEEE Transactions on Network Science and Engineering*, pp. 1–1, 2022.

[169] L. Jiang, X. Niu, J. Xu, D. Yang, and L. Xu, "Incentive mechanism design for truth discovery in crowdsourcing with copiers," *IEEE Transactions on Services Computing*, pp. 1–1, 2021.

[170] D. Wang, N. Vance, and C. Huang, "Who to select: Identifying critical sources in social sensing," *Knowledge-Based Systems*, vol. 145, no. APR.1, pp. 98–108, 2018.

[171] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1312–1320.

[172] "Dependdata: Data collection dependability through three-layer decision-making in bsns for healthcare monitoring," *Information Fusion*, vol. 62, pp. 32–46, 2020.

[173] A. Vadavalli and R. Subhashini, "An improved differential privacy-preserving truth discovery approach in healthcare," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 1031–1037.

[174] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, and A. Fawzi, "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, no. 7995, p. 468 – 475, 2024, cited by: 2; All Open Access, Green Open Access, Hybrid Gold Open Access.

[175] R. Zhao, Z. Xie, Y. Zhuang, and P. L. H. Yu, "Automated quality evaluation of large-scale benchmark datasets for vision-language tasks," *INTERNATIONAL JOURNAL OF NEURAL SYSTEMS*, vol. 34, no. 03, 2024 MAR 2024.

[176] J. Suzuki, H. Zen, and H. Kazawa, "Extracting representative subset from extensive text data for training pre-trained language models," *INFORMATION PROCESSING & MANAGEMENT*, vol. 60, no. 3, MAY 2023.

[177] S. Thapa and S. Adhikari, "Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls," *ANNALS OF BIOMEDICAL ENGINEERING*, vol. 51, no. 12, pp. 2647–2651, 2023 DEC 2023.

[178] V. Beretta, S. Harispe, S. Ranwez, and I. Mougenot, "Combining truth discovery and rdf knowledge bases to their mutual advantage," in *International Semantic Web Conference*. Springer, 2018, pp. 652–668.

[179] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 877–885.

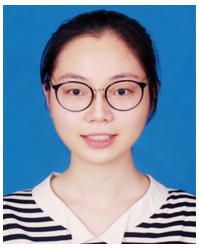
[180] ——, "Latent credibility analysis," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1009–1020.

[181] Z. Nakhaei and A. Ahmadi, "Toward high level data fusion for conflict resolution," in *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1. IEEE, 2017, pp. 91–97.

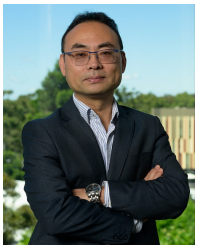
This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677

[182] L. Li, B. Qin, W. Ren, and T. Liu, "Truth discovery with memory network," *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 609–618, 2017.
[183] S. Lyu, W. Ouyang, Y. Wang, H. Shen, and X. Cheng, "Truth discovery by claim and source embedding," *IEEE Transactions on Knowledge and Data Engineering*, 2019.
[184] Y. Du, Y.-E. Sun, H. Huang, L. Huang, H. Xu, Y. Bao, and H. Guo, "Bayesian co-clustering truth discovery for mobile crowd sensing systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1045–1057, 2019.
[185] C. Ye, H. Wang, W. Lu, J. Gao, and G. Dai, "Deep truth discovery for pattern-based fact extraction," *Information Sciences*, vol. 580, pp. 478–494, 2021.
[186] J. Yang and W. P. Tay, "An unsupervised bayesian neural network for truth discovery in social networks," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[187] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.
[188] T. Hiroyasu, M. Miki, M. Sano, H. Shimosaka, S. Tsutsui, and J. Dongarra, "Distributed probabilistic model-building genetic algorithm," in *Genetic and Evolutionary Computation Conference*. Springer, 2003, pp. 1015–1028.
[189] P. Hansen, C. Oğuz, and N. Mladenović, "Variable neighborhood search for minimum cost berth allocation," *European journal of operational research*, vol. 191, no. 3, pp. 636–649, 2008.
[190] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.
[191] G. Leus, A. G. Marques, J. M. Moura, A. Ortega, and D. I. Shuman, "Graph signal processing: History, development, impact, and outlook," *IEEE Signal Processing Magazine*, vol. 40, no. 4, pp. 49–60, 2023.
[192] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
[193] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

Shuang Wang received her B.Sc. in the College of Sciences from Nanjing Agricultural University in 2015. She completed her PhD in the School of Computer Science and Engineering, Southeast University, Nanjing, China in 2020. She was a visiting PhD student at the School of Computing, Macquarie University, Sydney, Australia, from 2019 to 2020, and a postdoctoral research fellow from 2020 to 2021. She is currently a lecturer at Southeast University. Her research has been published in international journals and conferences such as *IEEE Transactions on Computers, IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Services Computing, IEEE Transactions on Network and Service Management* and ICSOC. Her main research interests focus on Service Computing, Big Data Analytics, and Truth Discovery.



He Zhang received her B.Sc. in Chien-Shiung WU College, Southeast University in 2022. She is currently a Master student in the School of Computer Science and Engineering, Southeast University, Nanjing, China. Her main research interests focus on Truth Discovery and Cloud Computing.

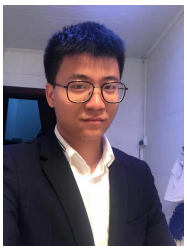


Quan Z. Sheng is a Distinguished Professor and Head of the School of Computing at Macquarie University, Sydney, Australia. His research interests include Service-oriented Computing, Distributed Computing, Internet Computing, and Internet of Things (IoT). Michael holds a PhD degree in computer science from the University of New South Wales (UNSW) and did his postdoc as a research scientist at CSIRO ICT Centre. Prof Michael Sheng is the recipient of AMiner Most Influential Scholar Award in IoT in 2019, ARC (Australian Research Council) Future Fellowship in 2014, Chris Wallace Award for Outstanding Research Contribution in 2012, and the Microsoft Research Fellowship in 2003. He is ranked by Microsoft Academic as one of the Most Impactful Authors in Services Computing (ranked the 4th all-time). Prof Sheng is the Vice Chair of the Executive Committee of the IEEE Technical Community on Services Computing (IEEE TCSVC), the Associate Director (Smart Technologies) of Macquarie University Smart Green Cities Research Centre, and a member of the ACS (Australian Computer Society) Technical Advisory Board on IoT.

Xiaoping Li (Senior Member, IEEE) received his B.Sc. and M.Sc. degrees in Applied Computer Science from the Harbin University of Science and Technology, Harbin, China, in 1993 and 1999, respectively, and the Ph.D. degree in Applied Computer Science from Harbin Institute of Technology, Harbin, China, in 2002. He is currently a Distinguished Professor at the School of Computer Science and Engineering, Southeast University, Nanjing, China. He is the author or co-author over 100 academic papers, some of which have been published in international journals such as *IEEE Transactions on Computers*; *IEEE Transactions on Parallel and Distributed Systems*; *IEEE Transactions on Services Computing*; *IEEE Transactions on Cybernetics*; *IEEE Transactions on Automation Science and Engineering*; *IEEE Transactions on Cloud Computing*; *IEEE Transactions on Systems, Man and Cybernetics: Systems*; *Information Sciences*; *Omega*, *European Journal of Operational Research*. His research interests include Scheduling in Cloud Computing, Scheduling in Cloud Manufacturing, Service Computing, Big Data and Machine Learning.

Zhu Sun is a Senior Scientist at the Centre for Frontier AI Research (CFAR), Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore. Dr. Sun Zhu received her Ph.D. degree from School of Computer Science and Engineering, Nanyang Technological University, Singapore, in 2018. Her main research topic is recommender systems. She has published papers in many leading conferences and journals (e.g SIGIR, SIGKDD, IJCAI, AAAI, CIKM, RecSys, TPAMI, TKDE, and TNNLS). She is the AE with ECRA journal, and PC Member for KDD, SIGIR, IJCAI, AAAI, SDM, CIKM, RecSys, IUI and UMAP, etc.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3423677



Taotao Cai is a Lecturer in Computing at the University of Southern Queensland (UniSQ). His primary focus is on research and teaching in the field of Data Science, including graph data processing, social network analytics, recommendation systems, and complexity science. He completed his Ph.D. degree from Deakin University in 2020, after spending two years at the University of Western Australia (Jan 2017 - Mar 2019) and over one year at Deakin University (Mar 2019 - July 2020) during his Ph.D. studies. Prior to joining the faculty at UniSQ, Taotao held positions as a PostDoctoral Research Fellow at Macquarie University (Mar 2021 - Jan 2023) and an Associate Research Fellow at Deakin University (July 2020 - Feb 2021). During this time, he made significant research contributions, which have been published in leading international conferences and journals such as IEEE ICDE, Information Systems, and IEEE TKDE.

Wei Emma Zhang is a Senior Lecturer and Associate Head of People and Culture at the School of Computer and Mathematical Sciences, and a researcher at the Australian Institute for Machine Learning, The University of Adelaide. She is also an Honorary Lecturer at the School of Computing, Macquarie University. Dr Zhang is the ARC Industry Early Career Research Fellow 2023-2026. She received her PhD degree from the University of Adelaide in Computer Science. Dr Zhang's research interests include Document Summarization, Adversarial attacks, and Artificial Intelligence of Things. She has more than 100 publications as edited books and proceedings, refereed book chapters, and refereed technical papers in journals and conferences including ACM CSUR, IEEE COMST, ACM TIST, ACM TOIT, ACM TOSN, WWWJ, CACM, ACL, ACM SIGIR, WWW, CIKM, ECCV, and ICSOC. Her PhD thesis had been published by Springer as a monograph. Wei is the top 100 authors worldwide, ranked by field-weighted citation impact, in the SciVal topic for Network Security. She is a member of ACM and IEEE.

Jian Yang is a full professor at the School of Computing, Macquarie University. She received her PhD in Data Integration from the Australian National University in 1995. Her main research interests are business process management, data science, and social networks. Prof. Yang has published more than 200 journal and conference papers in international journals and conferences such as IEEE Transactions, Information Systems, Data and Knowledge Engineering, VLDB, ICDE, ICDM, CIKM, etc. She is currently serving as an Executive Committee for the Computing Research and Education Association of Australasia.

Qing Gao received his Ph.D. degree in Mechatronics Engineering from the City University of Hong Kong, Kowloon, Hong Kong in 2014. From 2014 to 2016, he was with the School of Engineering and Information Technology, University of New South Wales, Canberra at the Australian Defence Force Academy, as a postdoctoral research associate. Since 2018, he has joined the School of Automation Science and Electrical Engineering, Beihang University as a full professor. His research interests include intelligent control and quantum control. Dr. Gao is the recipient of the Alexander von Humboldt Fellowship of Germany and the 21st Guan Zhao-Zhi Award.