Truth Discovery

Laure Berti-Equille

Chapter 1
Truth Discovery: A Survey

Laure Berti-Equille, Aix Marseille University, France

1.1 Introduction

In the era of Big Data, *volume*, *velocity*, and *variety* are commonly used to characterize the salient features of Big Data. However, the importance of *veracity*, the fourth "V" of Big Data is now well-recognized as a critical dimension that needs to be assessed by joint solutions coming from various research communities such as Natural Language Processing (NLP), Database (DB), and Machine Learning (ML), as well as from data science practitioners, and journalists [3, 5]. The problem of estimating veracity of on-line information in presence of multiple conflicting data is very challenging: information extraction suffers from uncertainties and errors; information sources may be dependent or colluded; misinformation is evolving and spreading fast in complex social networks. All these aspects have to be well-understood to be properly modeled in order to detect and combat effectively fake news and misinformation campaigns.

Rumor detection, misinformation spreading truth discovery, and fact-checking have been the subjects of much attention recently (See recent surveys [3, 4, 13] and comparative studies [18]). Typically, the main goal of truth finding is to infer the veracity of on-line (conflicting) information being claimed and/or (repeatedly) amplified by some sources on the Web, the blogosphere, and social media.

More and more sophisticated truth discovery models algorithms have been designed to capture the richness and complexity of real-world fact-checking scenarios, both for generic and specific cases and monitoring and tracking systems have been developed and tested in operational contexts[1] for fact-checking (e.g., ClaimBuster [11]), or tracking the social dynamics of on-line news sharing (e.g., Hoaxy [17]).

[1] e.g., Fact Check in Google News `https://blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/`

In this chapter, we will review these lines of work and also touch upon some cutting-edge problems for discovering truth in settings where information from multiple sources is conflicting and rapidly evolving. We discuss how close we are to meeting these challenges and identify many open problems for future research.

1.2 Definitions

Reputation, trust, and trustworthiness are concepts closely related to truth discovery.

Trustworthiness was originally measured by checking whether the contributor (i.e., the source) was contained in a list of trusted providers, since there exists an interdependency between the source (as data provider) and the data itself. On the one hand, data is likely to be accepted as true if it is provided by a trustworthy provider. On the other hand, the source is trustworthy if it provides true data. Thus, both data and data source can be checked to measure their trustworthiness [9]. However, the assumption that the value confidence can be measured using only its source trustworthiness has some limitations: for example, non-authoritative sources may provide some true and relevant information whereas reputable sources may provide false or erroneous information. Therefore the users have to make decisions based on factors such as the source, their prior knowledge about the subject, the reputation of the source and their prior experience.

As we will see in the next section, the conjecture of SourceRank [1] has influenced many research work in truth discovery: typically, in SourceRank, the more true relevant tuples a source returns, the more likely that other independent sources agree with its results. Inversely, independent sources are not very likely to agree on the same false tuples. Various methods have been then proposed to compute :

- source trustworthiness as a score that quantifies how reliable the source is, as a function of the confidence of its claims (also referred as source accuracy, quality or reliability) and
- value confidence as a score that quantifies the veracity of the claim, as a function of the trustworthiness of the sources claiming it.

1.3 Overview

Current approaches for evaluating the veracity of data are iterative methods that compute and update each source trustworthiness score, and then, update the belief score of each claimed value.

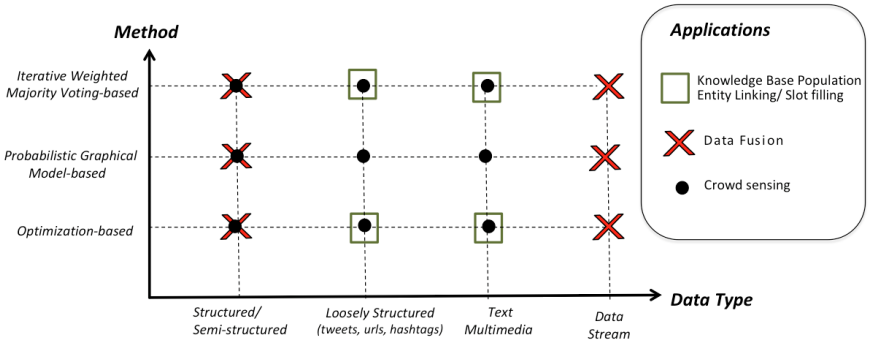

Fig. 1.1 Classification of the main approaches for truth discovery with typical application contexts, underlying computation model and data type.

1.3.1 Classification and evolution of truth discovery methods

We propose a classification existing approaches across a 2-dimensional space illustrated in Figure 1.1. Figure 1.1 presents a high-level view of the coverage of current methods in terms of applications such as data fusion, knowledge base population, and social sensing with crowded data. The X-axis presents the considered data types in four categories based on the data structure: from structured and semi-structured data, loosely structured and microtexts, free text or multimedia content, and data streams. The Y-axis defines the main principle of the underlying truth discovery methods such as weighted majority voting, probabilistic methods, or optimization-based methods as we will detail hereafter.

Truth discovery methods can be classified into the following categories.

- **Weighted voting-based methods** compute the value labels (i.e., true or false) and confidence scores using some variant of Majority Voting (MV) [10, 15, 22]. Traditional MV regards the value claimed by the majority of sources as the true value and randomly selects a value in case of tie with $1/|V_{o_a}|$ chance to infer the truth wrongly (with $|V_{o_a}|$, the number of distinct, conflicting values for the considered attribute a of object o). The reason is that MV assumes that each source has the same quality, but in reality, sources have different qualities, coverage, and scope. Various methods have adapted MV such that truths can be inferred through weighted voting to follow the principle that the information from reliable sources will be counted more in the aggregation.
- **Bayesian and Graphical Probabilistic Model-based methods** were proposed to precisely address the issues of MV mainly related to the latent (unknown) properties of the sources and of the claims. TruthFinder

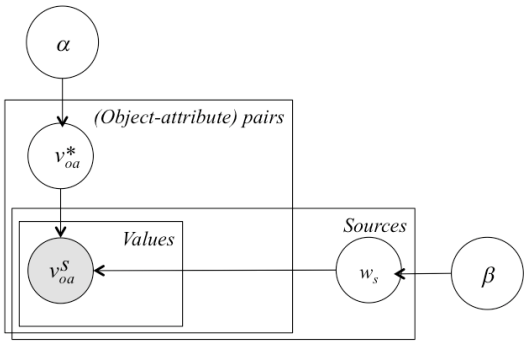

Fig. 1.2 Generic Plate diagram for GPM-based truth discovery methods

[22] was the first method that applies Bayesian analysis to estimate source trustworthiness and identify true claims with taking value similarity into consideration. DEPEN AND ACCU models [6, 7] were also the first Bayesian models that incorporate source copying detection techniques, highlighting the importance of source dependence in truth discovery. Other methods [16, 20, 23] modeled a truth discovery scenario as a Probabilistic Graphical Model (PGM) which expresses the conditional dependency structure (represented by edges) between random variables (represented by nodes). Figure 1.2 shows the generic PGM for truth discovery. In these approaches, the unknown identified truth noted $\{v^*_{o_a}\}$ and weights of the sources $\{w_s\}$ are the latent variables with a prior about the truth, denoted by α and a prior about the source reliability, denoted by β. The main principle is the following: if claimed values are close to the identified truth, the sources supporting them will gain weights. The goal is to maximize the likelihood and estimate high source weights when claims are close to the ground truth. To infer the two latent variables, techniques such as Expectation Maximization (EM) are adopted [16, 20] and the corresponding likelihood of a value being true is defined as:

$$\prod_{s\in S} p(w_s|\beta) \prod_{o\in O} \Big(p(v^*_{o_a}|\alpha) \prod_{s\in S} p(v^s_{o_a}|v^*_{o_a}, w_s) \Big). \tag{1.1}$$

.

- **Optimization-based methods** rely on setting an optimization function that can capture the relations between sources qualities and claims truth with an iterative method for computing these two sets of parameters jointly. The optimization function is generally formulated as:

$$\arg \min_{\{w_s\},\{v^*_{o_a}\}} \sum_{o\in O} \sum_{s\in S} w_s d(v^s_{o_a}, v^*_{o_a}) \tag{1.2}$$

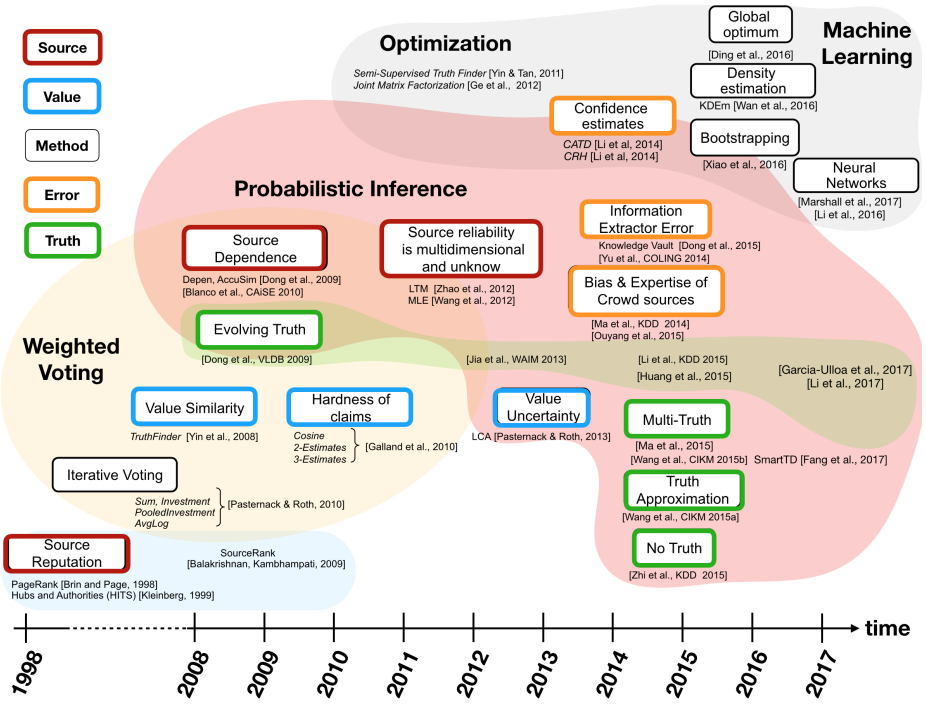

Fig. 1.3 Evolution of truth discovery research

where $d(\cdot)$ is a distance function that measures the difference between the information provided by source s and the identified truth (e.g., 0-1 loss function for categorical data or $L2$-norm for continuous data). The objective function measures the weighted distance between the value for an object-attribute pair denoted $\{v_{o_a}^s\}$ claimed by source s and the identified truth $\{v_{o_a}^*\}$. By minimizing this function, the aggregated results will be closer to the information from the sources with high weights.

As illustrated in Figure 1.3, the methods have significantly evolved in the last decade. The main trigger of this evolution is to overcome some of the limitations of previous approaches and to relax some modeling assumptions related to the sources (red boxes in the figure), their claimed values (blue boxes), the truth characteristics (green boxes), and the uncertainty or error associated (orange boxes) as we will mention in Section 1.4.2.

1.4 Key Research Findings

We will now present the main principle and core algorithm of truth discovery and review the modeling assumptions underlying the key research findings in this domain. Truth discovery methods applied to structured data take as input some conflicting quadruplets in the form of {*source*, *object*, *attribute*, *value*}, where *source* ($s \in S$) denotes the location where the data originates, *object* ($o \in O$) is an entity, *attribute* is an attribute of the object, and *value* ($v \in V^s_{o_a} \subset V$) is the value of an attribute of an object claimed by a source. For example, a quadruplet: `(imdb.com, director of Star Wars: Episode VIII - The Last Jedi, Rian Johnson)` indicates that the website "IMDb" claims that the director of the movie "Star Wars: Episode VIII - The Last Jedi" is "Rian Johnson". If a is a single-valued attribute for object o, $|V^s_{o_a}| = 1$. In case of a multi-valued attribute, e.g., "the list of actors" or "full cast & crew", $|V^s_{o_a}|$ is bigger than 1.

For each input quadruplet, the methods infer a Boolean truth label (i.e., true/false) as the output. Formally, we name the factual value of an attribute a of an object o as the ground truth, denoted by $v^*_{o_a}$. We note $(v)_m$ the label output given by a truth discovery method m for value v as the identified truth. After applying a group of truth discovery methods M one by one on the quadruplets, each method m ($m \in M$) outputs the identified truth for each object and its set of attributes. We denote the identified truth labels of all objects' attributes in O output by method m as $(V)_m$ ($(V_{o_a})_m \subset (V)_m$), and the ground truth of all objects in O, i.e., the complete ground truth of the given dataset, as V. The closer $(v_{o_a})_m$ is to $v^*_{o_a}$ for each attribute of each object (resp. $(V)_m$ to V^*), the better the method m performs. In most cases, the ground truth provided with real-world datasets is only a subset of the complete ground truth due to the prohibitive cost of collecting ground truth data.

1.4.1 Core Iterative Algorithm

As illustrated in Figure 1.4, the core algorithm of a fact-finder is an iterative, transitive voting algorithm. First, the trustworthiness of each source is initialized. For each object and each data item, the method calculates the confidence of each claim from the reliability of its sources. Then, it updates the trustworthiness of each source from the confidence of the claims it makes. The procedure is repeated until the stopping condition is satisfied. Usually, the stopping criterion is defined either as a maximum number of iterations or a threshold under which the results (*i.e.*, value confidence or source trustworthiness scores) are considered to be stable from one iteration to the next. Some methods start with initializing the confidence scores of the values instead of the source trustworthiness, then compute source trustworthiness and

update value confidence scores in a similar way. The methods differ in the way they compute and update the two scores as presented in Section 1.3.1.

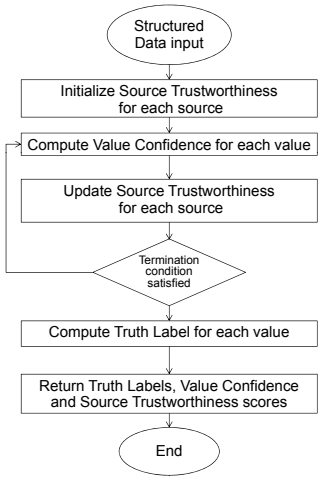

Fig. 1.4 Basic iterative algorithm for truth discovery

1.4.2 Modeling Considerations

Every truth discovery method relies on several modeling assumptions related to the sources, the input claims, the truth, and the output result.

- **Modeling the sources**. Three main modeling assumptions concern the sources: (1) Sources are assumed to be self-consistent and non-redundant. This means that a source should neither claim conflicting values for the same object-attribute pair, nor provide duplicate claims; (2) Current methods rely on trusting the majority and they are adapted only when the sources are assumed to be predominantly honest and the number of sources providing true claims is assumed to be significantly larger that the number of sources providing false claims. This assumption referred in [18] as the "optimistic scenario" is an important limitation for nowadays truth discovery scenarios.
- **Modeling the input claims**. Several considerations are important: (1) The cardinality (e.g., single or multi-valued attributes) and type of input claims (e.g., nominal vs numeric data) will determine the methods that can be applied and if data formatting is required; (2) The hardness of evaluating certain claims veracity can be considered in the model (see [10]); (3)

The extraction of structured claims is generally assumed to have no error but information extractors can generate uncertainties to be considered in truth discovery computation (see [8]); (4) Only claims with a direct source attribution are considered in current methods. This requires that each claim has its source explicitly identified with a direct correspondence; (5) Claims are usually assumed to be positive (except in [10]): e.g., "*S claims that A is false*" or "*S does not claim A is true*" are not considered by current methods; (6) Regarding claim uncertainty, few methods (e.g., LCA [16]) can handle cases such as "*S claims that A is true with 15% uncertainty*". They use a weight matrix to express the confidence each source in its assertions.

- **Modeling the relationship between source reliability and claim truthfulness**. (1) A potential limitation of current approaches is that the relationships between source reliability and claim truthfulness are often represented by simplified functions (e.g., linear, quadratic or binomial). This assumption may lead to suboptimal truth discovery results because the exact relational dependency between sources and claims is usually unknown a priori; (2) Another important related assumption is that the probability a source asserts a claim is independent of the truth of the claim. For example, some sources may choose to stay silent except for trivial or "easy" truths. This will lead to a very high trustworthiness score for these sources although they may not deserve it. Penalties in these cases can be applied to refine some existing model in order to address the "long-tail phenomenon" [12].
- **Modeling choices for the truth and output results.** (1) Each claim is assumed to be either true or false. Although value confidence score is generally returned as output by every method, truth labeling is usually Boolean with no contextual information or additional evidence (except in [19] where a posteriori explanations of the truth discovery results are provided in the form of decision trees and the robustness of a true claim is testing with generated allegations); (2) Single truth assumption: Claims related to the attribute of an object are organized into disjoint mutual exclusion sets where only one of the claims in each set is generally assumed to be true. However, some methods can handle multi-truth scenarios (e.g., LTM [23]) and situations where none of the values claimed by the sources is actually true [24].
- **Evaluation limits.** To evaluate efficiency and effectiveness performance of truth discovery methods, various metrics are traditionally used. Memory cost and running time can be used to evaluate the efficiency. Recall, precision, F1-measure, accuracy (or error rate) for categorical data, Mean of Absolute Error (MAE) and Root of Mean Square Error (RMSE) for continuous data are computed for quality performance when ground truth is available. However, the labor cost of collecting ground truth is usually prohibitive. As a consequence, ground truth is often very limited or even impossible to obtain in a reasonable amount. Methods that show the same

accuracy on sparse ground truth, may have different performance in reality. Under these circumstances, it is hard to conclude which method performs better as we cannot trust the comparison results due to its low statistical significance over sparse ground truth [4, 18].

Relaxing these modeling assumptions, enlarging the range of real-world scenarios captured by the models, evaluating and benchmarking truth discovery methods, and estimating the significance of the evaluation results based on sparse ground truth are the next challenging milestones in the truth discovery research agenda.

1.5 Novel Research Directions

Novel approaches leveraging machine learning have been recently proposed. They are promising research directions for addressing the great diversity of real-world misinformation scenarios.

ETCIBoot (Estimating Truth and Confidence Interval via Bootstrapping) [21]. Existing truth discovery methods focus on providing a point estimator for each object's truth, but in many real-world applications, confidence interval estimation of truths is more desirable because it contains richer information. ETCIBoot constructs confidence interval estimates as well as identify truths with integrating bootstrapping techniques into the truth discovery procedure. Due to the properties of bootstrapping, the estimators obtained by ETCIBoot are more accurate and robust compared with the state-of-the-art truth discovery approaches. Theoretically, the authors of [21] prove the asymptotic consistency of the confidence interval obtained by ETCIBoot.

Neural Networks. A novel neural network-based approach has been recently proposed by [14]. This method can learn complex relational dependency between source reliability and claim truthfulness. A multi-layer neural network model has been developed to solve the truth discovery problem in social sensing without any assumption on the prior knowledge of the source-claim relational dependency distribution. In particular, a neural network for truth discovery is defined by a set of input neurons which are activated by the social sensing data (i.e., sources and the claims they make). After being weighted and transformed by a learning function, the activation of these neurons are then passed on to other neurons inside the neural networks. This process is repeated until the output neuron that determines the truthfulness of a claim is activated. The complex source-claim relational dependency is learned by the neural network model through the above training process.

1.6 Conclusions

This chapter presented an overview and recent advances of truth discovery research emphasizing on the main methods and their underlying modeling assumptions.

We have observed that none of the methods constantly outperforms the others in terms of precision/recall and a "one-fits-all" approach does not seem to be achievable. Most of the current methods have been designed for excelling in *optimistic* scenarios with a reasonable number of honest and reliable sources. However, experiments reported in [18] revealed that, for *pessimistic* or adversarial scenarios when most of sources are not reliable, most of the methods have relatively low precision, some expose prohibitive runtime and may suffer from scalability or fluctuating results. Ensembling [2] and bootstrapping [21] truth discovery methods seems to be very promising research directions.

Another challenge is related to the usability of the methods. The assumptions made by current truth discovery models and their complex parameter settings make most methods still difficult to apply to the wide diversity of online information and existing scenarios on the Web. Since limited and generally sparse ground truth is available, performance evaluation and comparative studies may not be reliable and have no statistical significance; benchmarks and repeatability experiments are critically needed.

References

1. Balakrishnan, R., and Kambhampati, S. SourceRank: Relevance and trust assessment for deep Web sources based on inter-source agreement. In *Proceedings of the International Conference on World Wide Web (WWW 2011)* (2011), pp. 227–236.
2. Berti-Équille, L. Data veracity estimation with ensembling truth discovery methods. In *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015* (2015), pp. 2628–2636.
3. Berti-Équille, L. Scaling up truth discovery. In *Proceedings of the 32nd IEEE International Conference on Data Engineering (ICDE),Helsinki, Finland, May 16-20, 2016* (2016), pp. 1418–1419.
4. Berti-Équille, L., and Borge-Holthoefer, J. *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015.
5. Cohen, S., Li, C., Yang, J., and Yu, C. Computational journalism: A call to arms to database researchers. In *Proceedings of the Fifth Biennial Conference on Innovative Data Systems Research (CIDR 2011)* (2011), pp. 148–151.
6. Dong, X. L., Berti-Equille, L., Hu, Y., and Srivastava, D. Global Detection of Complex Copying Relationships Between Sources. *Proc. VLDB Endow. 3*, 1-2 (2010), 1358–1369.
7. Dong, X. L., Berti-Equille, L., and Srivastava, D. Integrating conflicting data: The role of source dependence. *PVLDB 2*, 1 (2009), 550–561.
8. Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014* (2014), pp. 601–610.
9. Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., and Zhang, W. Knowledge-based trust: Estimating the trustworthiness of web sources. *IEEE Data Eng. Bull. 39*, 2 (2016), 106–117.
10. Galland, A., Abiteboul, S., Marian, A., and Senellart, P. Corroborating Information from Disagreeing Views. In *WSDM* (2010), pp. 131–140.
11. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., and Tremayne, M. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB 10*, 12 (2017), 1945–1948.
12. Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., and Han, J. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment 8*, 4 (2014), 425–436.

13. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. A survey on truth discovery. *SIGKDD Explorations 17*, 2 (2015), 1–16.
14. Marshall, J., Argueta, A., and Wang, D. A neural network approach for truth discovery in social sensing. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* (Oct. 2018), pp. 343–347.
15. Pasternack, J., and Roth, D. Knowing what to believe (when you already know something). In *Proceedings of the Conference on Computational Linguistics (COLING'10)* (2010), pp. 877–885.
16. Pasternack, J., and Roth, D. Latent credibility analysis. In *Proceedings of the International World Wide Web Conference (WWW 2013)* (2013), pp. 1009–1020.
17. Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web* (Republic and Canton of Geneva, Switzerland, 2016), WWW '16 Companion, International World Wide Web Conferences Steering Committee, pp. 745–750.
18. Waguih, D. A., and Berti-Equille, L. Truth discovery algorithms: An experimental evaluation. *CoRR abs/1409.6428* (2014).
19. Waguih, D. A., Goel, N., Hammady, H. M., and Berti-Equille, L. AllegatorTrack: Combining and reporting results of truth discovery from multi-source data. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE 2015)* (2015), pp. 1440–1443.
20. Wang, D., Kaplan, L. M., Le, H. K., and Abdelzaher, T. F. On Truth Discovery in Social Sensing: a Maximum Likelihood Estimation Approach. In *IPSN* (2012), pp. 233–244.
21. Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., and Zhang, A. 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), KDD'16, pp. 1935–1944.
22. Yin, X., and Han, J. Truth discovery with multiple conflicting information providers on the web. In *In Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)* (2007).
23. Zhao, B., Rubinstein, B. I. P., Gemmell, J., and Han, J. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB 5*, 6 (2012), 550–561.
24. Zhi, S., Zhao, B., Tong, W., Gao, J., Yu, D., Ji, H., and Han, J. Modeling truth existence in truth discovery. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)* (2015), pp. 1543–1552.