Scaling Up Truth Discovery

Laure Berti-Équille
Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
Email: lberti@qf.org.qa

***Abstract*—The evolution of the Web from a technology platform to a social ecosystem has resulted in unprecedented data volumes being continuously generated, exchanged, and consumed. User-generated content on the Web is massive, highly dynamic, and characterized by a combination of factual data and opinion data. False information, rumors, and fake contents can be easily spread across multiple sources, making it hard to distinguish between what is true and what is not. Truth discovery (also known as fact-checking) has recently gained lot of interest from Data Science communities. This tutorial will attempt to cover recent work on truth-finding and how it can scale Big Data. We will provide a broad overview with new insights, highlighting the progress made on truth discovery from information extraction, data and knowledge fusion, as well as modeling of misinformation dynamics in social networks. We will review in details current models, algorithms, and techniques proposed by various research communities whose contributions converge towards the same goal of estimating the veracity of data in a dynamic world. Our aim is to bridge theory and practice and introduce recent work from diverse disciplines to database people to be better equipped for addressing the challenges of truth discovery in Big Data.**

I. Introduction

The importance of veracity and dynamics of information on the Web has led to a substantial amount of research over the past few years raising lot of interest not only from the academia and the Web industry, but also from national Intelligence and News agencies for its direct application to homeland security and computational journalism. Veracity of Big Data is a timely topic of particular interest to many database people and to the data science community at large. It attracts a large audience of data scientists, researchers, and practitioners interested in the challenges when ascertaining the veracity of online information, evolving and spreading at a very fast pace. The goal of the tutorial is to provide a comprehensive and cohesive overview of the contributions related to every stage of the truth discovery pipeline, ranging from information extraction, data and knowledge fusion, probabilistic inference, and truth discovery computation to the modeling of the propagation of falsified and distorted information (*i.e.*, rumour) in social networks. A particular emphasis will be on scalability of current approaches.

In the introduction, the tutorial will first start with a variety of real-world examples to motivate the importance of checking the veracity of online information and illustrate how it may affect our society, economy, and privacy.

II. From Information Extraction to Truth Discovery Computation

The problem of truth discovery is intellectually and technically interesting enough to have attracted a lot of prior studies from diverse research communities in information retrieval, artificial intelligence, and data management, sometimes investigated under the names of "fact-checking", "information trustworthiness", "information credibility", "trust management", "information corroboration", "data fusion" or "knowledge fusion".

In the first part of the tutorial, we will introduce the main approaches of truth discovery as illustrated in Figure 1 and we will describe the truth discovery pipeline, starting with the crucial stage of information extraction. We will provide a comprehensive survey of the advances in text analysis and natural language processing, more specifically on contradiction detection and knowledge base population (KBP) with presenting entity linking and slot filling techniques. Each stage of content analysis (from preprocessing to entity matching) may produce errors and we will present techniques to mitigate information extraction problems in the truth discovery computation. We will then review the work on data fusion that is particularly relevant to truth discovery, with a particular focus on recent scalable approaches.

Many truth discovery methods have been proposed to deal with data veracity estimation (see [1] for a survey). With majority voting as their underlying principle, previous methods are mostly applied to structured data and iteratively compute and update the trustworthiness of a source as a function of the belief in its claims, and then the belief score of each claim as a function of the trustworthiness of the sources asserting it. Since the early work of Yin et al. in 2007, various models have been proposed for truth discovery to incorporate various aspects beyond source trustworthiness and claim belief for structured data, such as: *(i)* a prior knowledge either about the claimed assertions or about the source reputation via trust assessment; *(ii)* the dependence between sources (e.g., [5], [4]); *(iii)* the temporal dimension of evolving truth; *(iv)* the difficulty of estimating the veracity of certain claims; and *(v)* the management of complex data structures such as collections of entities in a claim, and correlation of claims. As illustrated in the figure, recent contributions relaxing prior modeling assumptions have been proposed to deal with truth existence and approximate truth discovery. Other new developments are related to incremental truth discovery, truth discovery from data streams, and dynamic truth discovery in social media and crowd sourcing applications. To tackle the problem of truth discovery in Web data, recent approaches have been developed

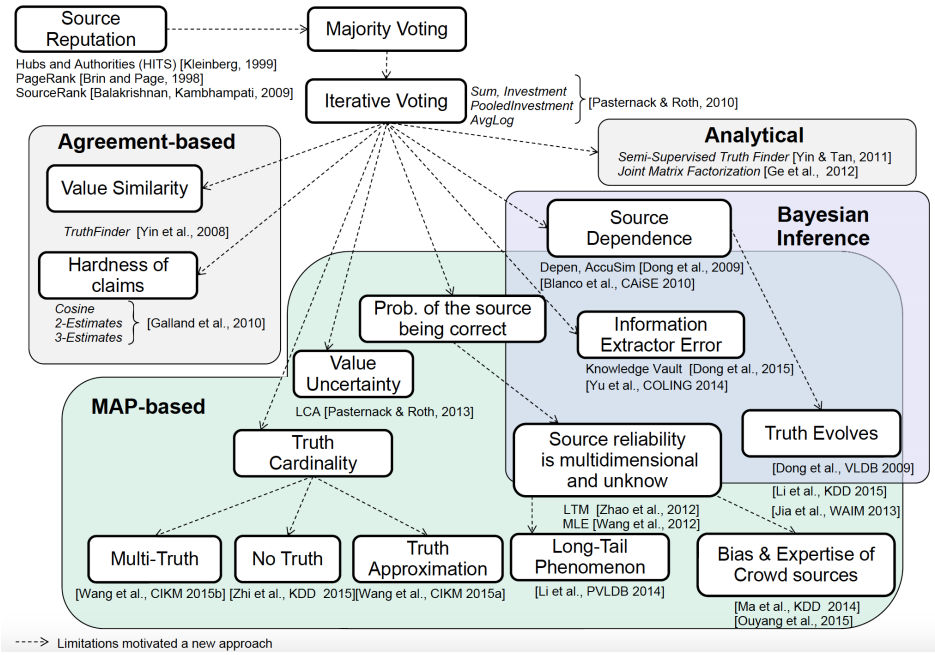

Fig. 1. Truth Discovery Roadmap

to discover true values extracted from textual content in a large corpus of Web sources using various information extractors (e.g., [3]). These solutions extend previous probabilistic models based on iterative vote counting and integrate the extraction systems' error in truth discovery computation at Web scale. We will review all the approaches and present in detail their modeling assumptions, algorithms, applications and limitations (as shown in the study of [2]). Since "one-fits-all" solution does not seem to be achievable for a wide range of truth discovery scenarios, we will discuss several opportunities for new research, such as ensembling truth discovery methods [7] in order to improve the quality performance of current results. We expect research along this line can attract the audience's interest.

III. Scaling Up Truth Discovery

Real-world evolves, data changes from one provider to another, and information mutates more quickly than we can ever acknowledge. Conflicting data can be easily spread across multiple sources. True contents can mutate and become viral rumors and lies. Therefore, many questions can be legitimately asked: How do we figure out that a lie has been told often enough that it is now considered to be true? How many lying sources are required to introduce confusion in what you knew before to be the truth? What is the origin of a rumor or a fake news? How can we detect and mitigate allegations or falsified information?

In the second part of our tutorial, we will address the truth-finding problem as amplified and made even more complex by data volume and data velocity. The audience will understand how these two dimensions of Big Data can affect truth discovery algorithms and how distributed and scalable versions of current algorithms can be achieved. We also discuss the need for designing realistic truth discovery scenarios with benchmark data sets for testing the methods at scale with various perturbations (e.g., attacks by injecting allegations).

Finally, we will end the tutorial by discussing cutting-edge open problems for discovering truth in settings where information from multiple sources is rapidly evolving, distorted, and propagated by mostly non reliable sources (i.e., cases where majority voting is not adequate). This tutorial will also review the challenges of truth discovery in Big Data with awareness of misinformation dynamics. We will discuss how close we are to meeting these challenges and we will identify various open problems for future research in data management and truth finding. Truth discovery and misinformation dynamics give clear opportunities to the ICDE community for cutting-edge research in data and knowledge management to ultimately design systems that effectively support the fourth "V" (Veracity) of Big Data.

References

[1] L. Berti-Equille and J. Borge-Holthoefer, *Veracity of Big Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*, Morgan & Claypool,2015.

[2] D. Attia Waguih and L. Berti-Equille, *Truth Discovery Algorithms: An Experimental Evaluation*, CoRR 1409.6428, 2014.

[3] X.L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun and W. Zhang, *Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion*, KDD'14, p. 601–610, 2014.

[4] X.L. Dong, L. Berti-Equille and D. Srivastava, *Integrating Conflicting Data: The Role of Source Dependence*, PVLDB (2):1, p. 550–561, 2009.

[5] X.L. Dong, L. Berti-Equille, Y. Hu a and D. Srivastava, *SOLOMON: Seeking the Truth Via Copying Detection*, PVLDB (3):2, p. 1617–1620, 2010.

[6] D. Attia Waguih, N. Goel, H.M. Hammady, L. Berti-Equille, *AllegatorTrack: Combining and Reporting Results of Truth Discovery from Multi-Source Data*, ICDE'15, p. 1440–1443, 2015.

[7] L. Berti-Equille, *Data Veracity Estimation with Ensembling Truth Discovery Methods*, IEEE Big Data, 2015.