Veracity of Big Data: Challenges

LAURE BERTI-EQUILLE, MOUHAMADOU LAMINE BA, HOSSAM M. HAMMADY, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.0 [**Database Management**]: General

General Terms: Algorithms, Management, Verification

Additional Key Words and Phrases: Truth Discovery, Fact-checking, Data Quality, Data Fusion, Information Extraction

ACM Reference Format:
L. Berti-Equille, M. Lamine Ba, H. M. Hammady 2015. Veracity of Big Data: Challenges. *ACM J. Data Inform. Quality* 6, 6, Article 6 (November 2015), 3 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

Next Generation of Truth Discovery Systems

As online user-generated content grows exponentially, the reliance on Web and social media data is increasing. Truth discovery from the Web has significant practical importance as online rumor and misinformation can have tremendous impacts on our society and everyday life. One of the fundamental difficulties is that data can be biased, noisy, outdated, incorrect, misleading and thus unreliable. Conflicting data from multiple sources amplifies this problem and veracity of data has to be estimated. Beyond the emerging field of computational journalism and the success of online fact-checkers (e.g., FactCheck[1], ClaimBuster[2]), truth discovery is a long-standing and challenging problem studied by many research communities in artificial intelligence, databases, and complex systems and under various names: fact-checking, data or knowledge fusion, information trustworthiness, credibility or information corroboration (see [1] for a survey and [11] for a comparative analysis). The ultimate goal is to predict the truth label of a set of assertions claimed by multiple sources and to infer sources' reliability with no or few prior knowledge. One major line of previous work aimed at iteratively computing and updating the source's trustworthiness as a belief function in its claims, and then the belief score of each claim as a function of its sources' trustworthiness [14]. More complex probabilistic models have then incorporated various aspects beyond source trustworthiness and claim belief such as the dependence between sources [3; 2], the correlation of claims [10], the notion of evolving truth [4]. Recent contributions have further relaxed prior modeling assumptions to deal with

[1]www.factcheck.org/
[2]idir.uta.edu/claimbuster

Author's address: L. Berti-Equille, M. L. Ba, H. M. Hammady, Qatar Computing Research Institute(Current address), Hamad Bin Khalifa University, Tornado Tower 18th, P.O. Box 5825, Doha, Qatar.

DOI:http://dx.doi.org/10.1145/0000000.0000000

truth existence [15], approximate truth discovery [13; 7], truth evolution [8; 6], and applications in the context of social media and crowd sourcing [5; 9]. their evaluation depends on available samples of ground truth data. To the best of our knowledge, our work is the first one to apply an ensembling method to truth discovery and also the first to address the problem of ground truth data sample selection bias.

In this paper, we argue that the next generation of data management and data sharing systems need to manage not only volume and variety of Big Data but most importantly veracity of data. Designing truth discovery systems requires a fundamental paradigm shift in data management and goes beyond adding new layers of data fusion heuristics or developing yet another probabilistic graphical truth discovery model. Actionable and Web-scale truth discovery requires a transdisciplinary approach to incorporate the dynamic and cross-modal dimension related to multi-layered networks of contents and sources. Apart from the limitations of current truth discovery methods, we would like to highlight the following challenges.

Timely and Actionable Truth Discovery. Truth discovery from quasi real-time data could save lives in a humanitarian context for example. To be actionable, information extraction and truth discovery computation need to be streamlined, prioritized depending on the level of emergency and incompleteness of available information, and finally adjusted to the communities that will use the data (e.g., rescue team, NGOs). The long tail phenomenon problem (*i.e.* where very few sources provide the first information after a disaster) is amplified and highly time-dependent.

Cross-modal and Cross-lingual Truth Discovery. The agility of a truth discovery system is of utmost importance to efficiently extract and map information: (i) in various languages; (ii) in various data formats, structures, and semantics (e.g., texts, Web table, structured data, etc.); (iii) and conveyed by various media and technologies (e.g., tweets, instagram images, youtube videos, Web pages, etc.).

Estimation of Incompleteness, Biases and Errors in the Truth Discovery Process. Information without context can be easily distorted and misinterpreted. When a piece of information is extracted from its original content, channel or thread, it may lose its context along with important "semantic markers" that explain *when, where, how, why,* and *for which* purpose or audience it has been produced. Observation may also be incomplete and biased for various reasons, e.g., security and privacy concerns, format limitations, *observer's bias* or *disclosure bias*. Estimating the biases and errors along the entire truth discovery pipeline is crucial and challenging.

To overcome these challenges, we believe that an integrative framework is needed: (i) To define, in a principled way, a unified semantics of truth discovery; (ii) To proactively collect new evidences, contextual data, and external knowledge from multimodal data; (iii) To support continuous inference and belief revision for computing and updating data veracity estimates; (iv) And finally, to monitor and estimate errors and biases in the truth discovery process.

To address these challenges, we proposed DAFNA (*Data Forensics with Analytics*) at QCRI, an ambitious project for determining the veracity of cross-modal information from multiple Web sources. Beyond a first module demonstrated in [12], DAFNA's vision is to provide a platform for actionable and cross-modal truth discovery.

REFERENCES

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

X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. SOLOMON: Seeking the Truth Via Copying Detection. *Proc. of the VLDB Endowment*, 3(2):1617–1620, 2010.

X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global Detection of Complex Copying Relationships Between Sources. *Proc. of the VLDB Endowment*, 3(1-2):1358–1369, 2010.

X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth Discovery and Copying Detection in a Dynamic World. *Proc. of the VLDB Endowment*, 2(1):562–573, 2009.

J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han. Truth discovery and crowdsourcing aggregation: A unified perspective. *Proc. of the VLDB Endowment*, 8(12):2048–2059, 2015.

L. Jia, H. Wang, J. Li, and H. Gao. Incremental truth discovery for information from multiple data sources. In *Proc. of the Web-Age Information Management (WAIM) 2013 International Workshops*, pages 56–66, 2013.

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. *Proc. of the VLDB Endowment*, 8(4):425–436, 2014.

Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 675–684, 2015.

F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 745–754, 2015.

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 (SIGMOD'14)*, pages 433–444, 2014.

D. A. Waguih and L. Berti-Equille. Truth Discovery Algorithms: An Experimental Evaluation. QCRI Technical Report, May 2014.

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 *Proc. of ICDE*, pages 1440–1443, Seoul, Korea, 2015. IEEE.

X. Wang, Q. Z. Sheng, X. S. Fang, X. Li, X. Xu, and L. Yao. Approximate truth discovery via problem scale reduction. In *Proc. of the 24th ACM Conference on Information and Knowledge Management (CIKM)*, October 2015.

X. Yin, J. Han, and P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Trans. Knowl. Data Eng.*, 20(6):796–808, 2008.

S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1543–1552, 2015.