DISCOVERING THE TRUTH ON THE WEB
One Facet of Data Forensics

Mouhamadou Lamine Ba, Laure Berti-Equille, Hossam M. Hammady

Qatar Computing Research Institute, HBKU, Doha, Qatar
{mlba, lberti, hhammady}@qf.org.qa

Truth Discovery Use Cases

Query Answering: Dealing with many possible answers

Example query: How many dead in Paris terrorist attacks?

Claim	Source	Value	Truthfulness
c_5	cnn.com	At least 128	Unknown
c_4	theguardian	120	Unknown
c_3	news.sky.com	130	Unknown
c_2	bbc.com	130	Unknown
c_1	@TBurgesWatson	35	Unknown

Multi-Source Data Fusion: Resolving conflicting values

Example: Percentage of expats in Qatar

S1: City	% expats
Qatar	75%

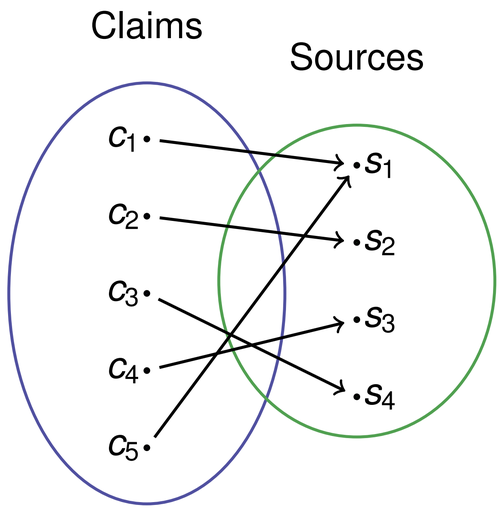
S2: City	% expats
Qatar	80%

S2: City	% expats
Qatar	60%

City	% expats
Qatar	?

Finding the most relevant results in both use cases requires the estimation of the **truthfulness level** of each data claimed by the sources.

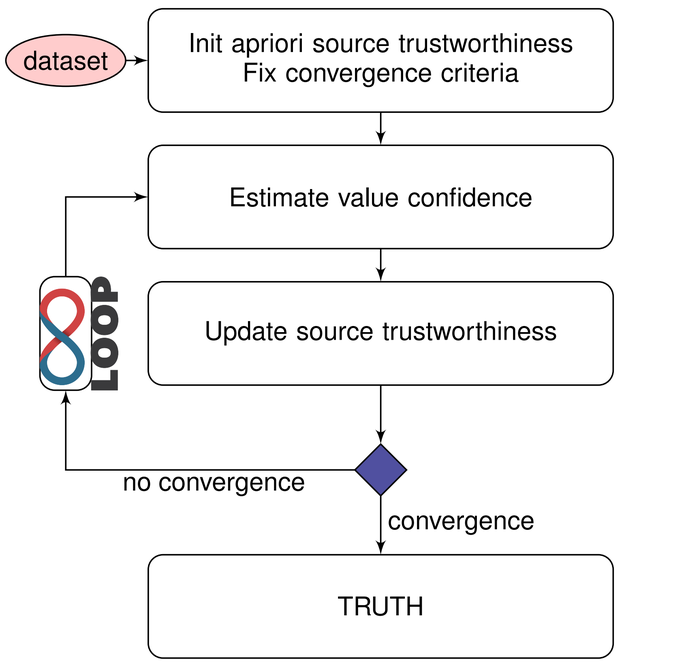
Problem statement: Given a set of claims $\{c_1, c_2, \ldots, c_n\}$ about a real-world fact f claimed by n information sources, truth discovery computes a mapping $\mathcal{F} : \{c_1, c_2, \ldots, c_n\} \mapsto \{\text{True}, \text{False}\}$



$c_1 \mapsto$ *FALSE*
$c_2 \mapsto$ *FALSE*
$c_3 \mapsto$ *TRUE*
$c_4 \mapsto$ *FALSE*
$c_5 \mapsto$ *FALSE*

- Web data integration
- Fact-checking in computational journalism
- Query answering in Web open information extraction
- Evaluation of tasks and workers in crowd-sourcing apps

Truth Discovery Pipeline



DAFNA – Data Forensics with Analytics

Data Forensics with Analytics (DAFNA) is a project initiated by the Data Analytics group at QCRI. DAFNA aims at providing a suite of tools for estimating data veracity for Data Forensics and solving some of the limitations of current truth discovery methods.

API

- Enable access to twelve state-of-the art truth discovery algorithms
- Provide explanations for truth discovery decisions

AllegatorTrack

VERA

- More scalable strategies
- Time-dependent consensus model for veracity estimation
- Self-adaptive model of truth discovery

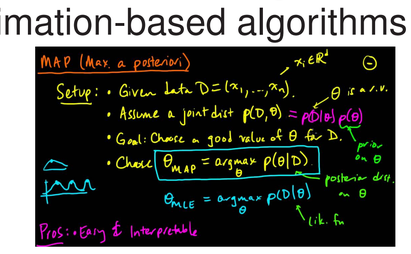
Extensive evaluation platform

Truth on the Web

Three Classes of Algorithms

One can classify existing truth discovery approaches in three classes.

- Agreement-based algorithms
- MAP Estimation-based algorithms
- MLE Estimation-based algorithms



Known Limitations

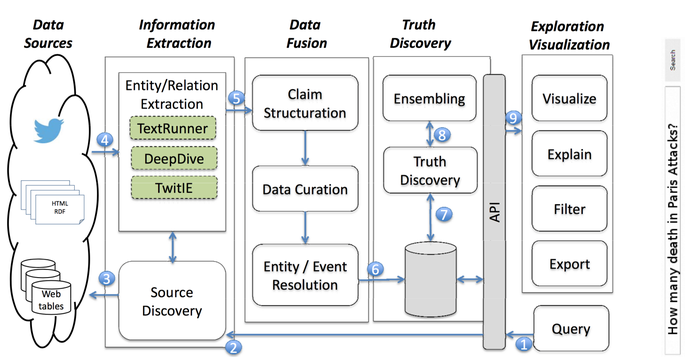
Applicability of truth discovery algorithms on the Web faces the following challenges.

- No one-fits-all solution
- No common evaluation platform
- No query-specific claim extraction
- Limited support of rapidly evolving truth

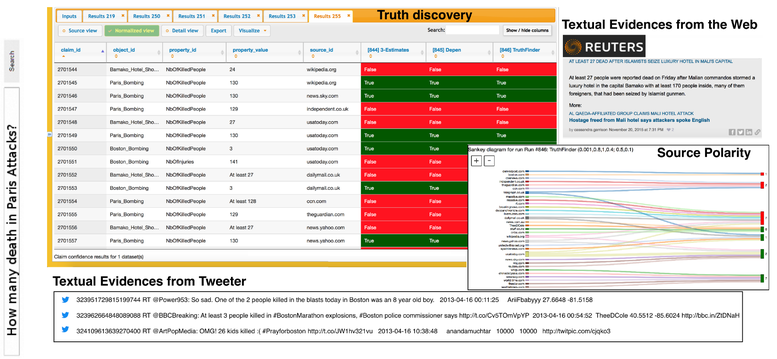
VERA – Veracity Estimation on Web Data

VERA is a Web-based platform supporting the pipeline of truth discovery from Web unstructured corpus and tweets: ranging from information extraction from raw texts and micro-texts and data fusion to truth discovery and visualization. VERA offers several advantages over previous work as it includes:

- Extraction and fusion of multi-source information to answer a factual query defined by the user;
- Use of time-dependent model in order to capture evolving truth, in particular from social media;
- Ensembling multiple truth discovery models to effectively find true values from conflicting ones;
- Visualization artifacts to better understand the information space with disagreeing vs. agreeing sources and corroborating vs. conflicting claims.



(a) System Architecture



(b) VERA Back-End and Visualization Artifacts

Publications

[1] Mouhamadou Lamine Ba, Laure Berti-Equille, Kushal Shah, Hossam M. Hammady: VERA – A Platform for Veracity Estimation over Web Data . In World Wide Web (WWW), 25th Interntional World Wide Web Conference, April 2016.

[2] Laure Berti-Equille and J. Borge-Holthoefer: Veracity of Big Data – From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics. Lectures on Data Management, Morgan & Claypool Publishers, 2015.

[3] 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 Data Engineering (ICDE), 2015 IEEE 31st International Conference on Data Engineering, pages 1440 –1443, April 2015.

[4] Dalia Attia Waguih and Laure Berti-Equille. Truth discovery algorithms: An experimental evaluation. CoRR, abs/1409.6428, 2014.