معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Laure Berti-Equille *QCRI, HBKU*

lberti@qf.org.qa

Javier Borge-Holthoefer *I3, OUC*

jborgeh@uoc.edu

Scaling Up Truth Discovery

May 17, 2016

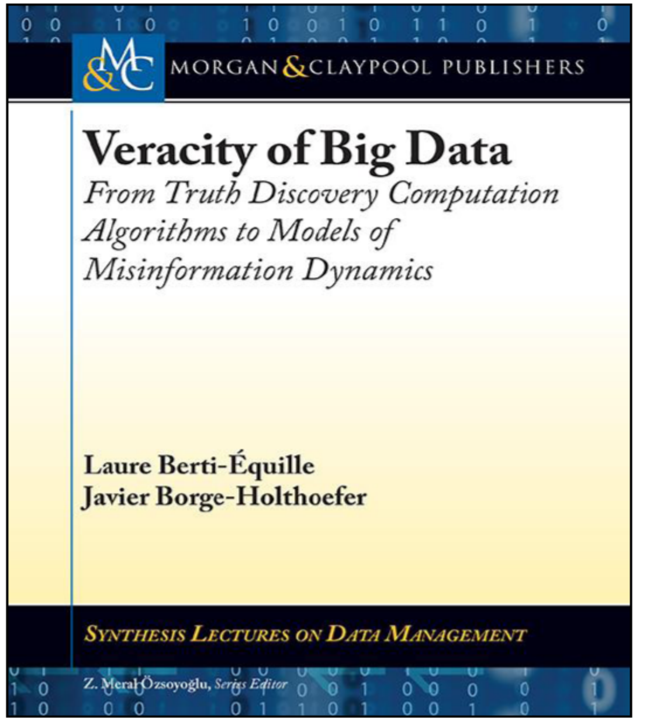
Disclaimer

Aim of the tutorial: Get the big picture

The algorithms of the main approaches will be sketched

Please don't mind if your favorite algorithm is missing

The revised version of the tutorial will be available at: http://daqcri.github.io/dafna/

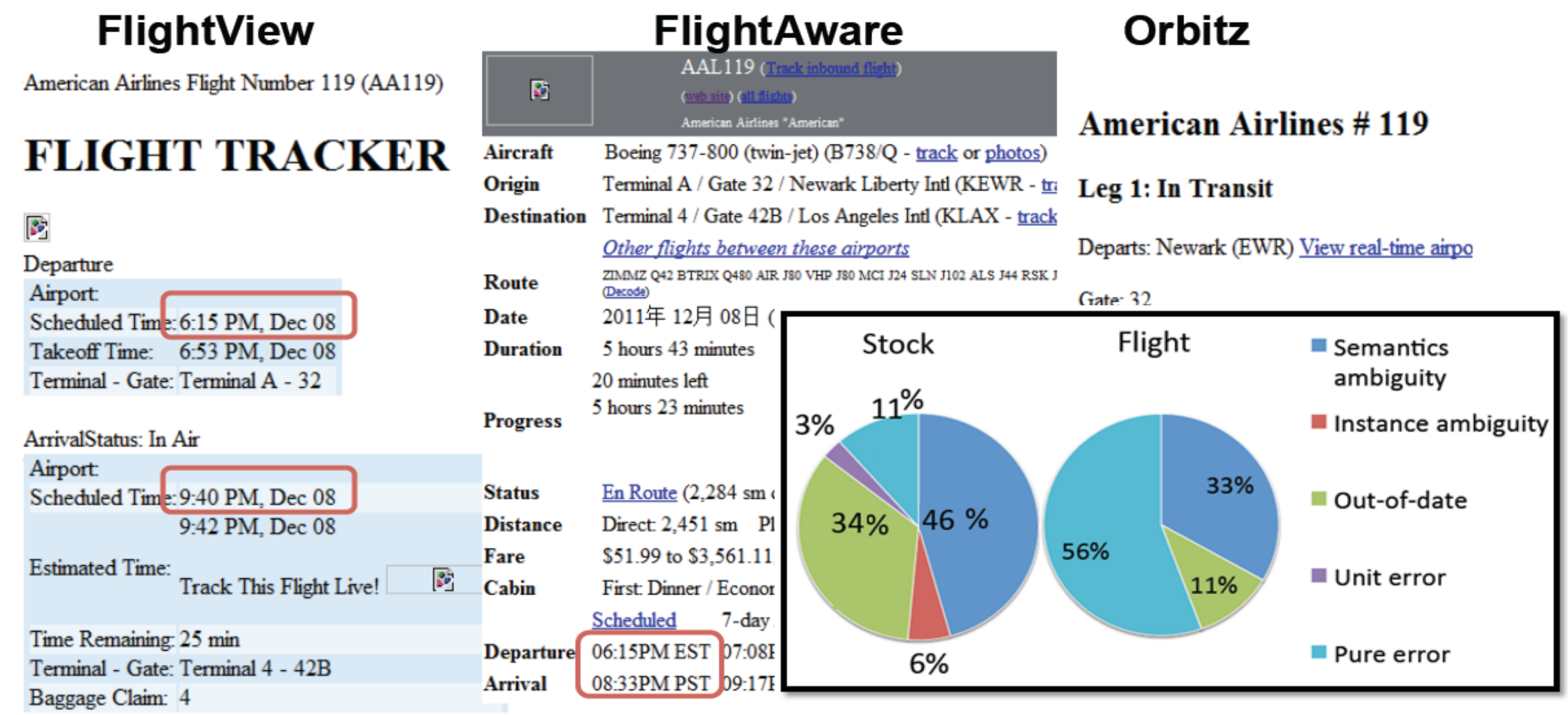


So many sources of information…

Are all these sources equally
- **accurate**
- **up-to-date**
- **and trustworthy?**

Accurate?

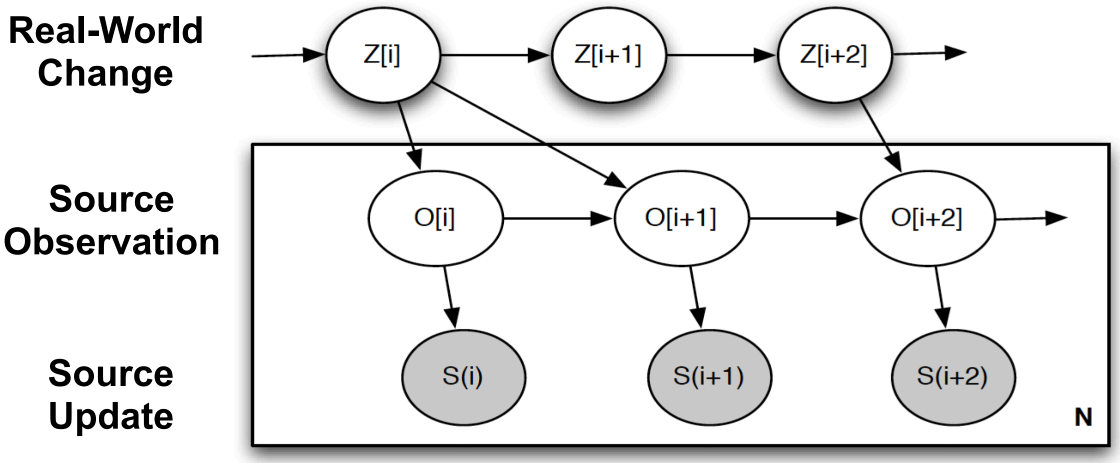
Deep Web data quality is low



X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth Finding on the Deep Web: Is the Problem Solved? PVLDB, 6(2):97–108, 2012.

Up-to-date?

Real-world entities evolve over time, but sources can delay, or even miss, reporting some of the real-world updates.



Research: 80% fund giants publish out of date fund data

15 September 2015 | By Valentina Romeo

Tweet 9 | in Share 5 | Print | Email | Comments (3)

Eight out of ten of the biggest fund groups are handing investors outdated performance information, a new survey finds.

According to fintech company Instinct Studios, 80 per cent of the largest asset managers have fund factsheets that are six weeks out of date.

A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. Proceedings of WWW '12, p. 789-798.

Trustworthy?
WikiTrust

Computed based on edit history of the page and reputation of the authors

- *B.T. Adler, L. de Alfaro, A Content-Driven Reputation System for the Wikipedia, Proceedings of the 16th International World Wide Web Conference, 2007.*
- *L. de Alfaro, B. Adler. Content-Driven Reputation for Collaborative Systems. Proceedings of Trustworthy Global Computing 2013.Lecture Notes in Computer Science, Springer, 2013.*

Information can still be trustworthy

Sources may not be "reputable", but information can still be trusted.

Authoritative sources can be wrong

YAHOO! NEWS

AFP apologises to French industrialist after death reported

AFP February 28, 2015 2:42 PM

AFP issued an apology to French industrialist Martin Bouygues, chairman and CEO of the conglomerate Bouygue...

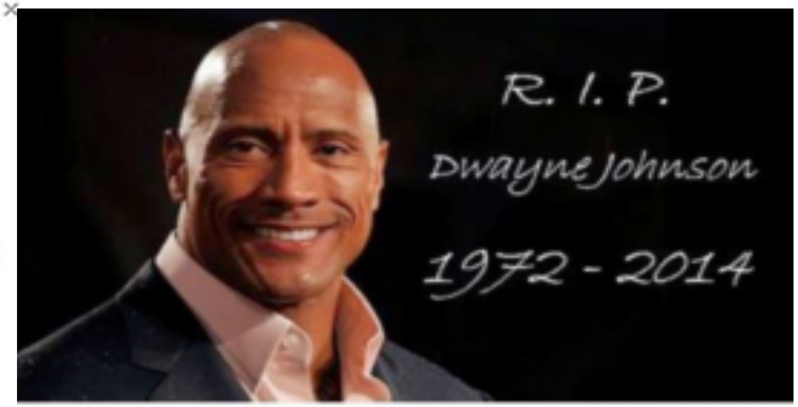
Rumors: Celebrity Death Hoaxes

成龍 Jackie Chan
June 21

Hi everybody! Yesterday, I got on a 3am flight from India to Beijing. I didn't get a chance to sleep and even had to clean my house when I got home. Today, everybody called to congratulate me on my rumored engagement. Afterward, everybody called me to see if I was alive.

If I died, I would probably tell the world! I took a photo with today's date, just in case you don't believe me! However, thank you all for your concern. Kiss kiss and love you all!

P.S. My dog is healthy, just like me! He doesn't need surgery! By the way, my dogs are golden retrievers, not Labradors.



DWAYNE JOHNSON died while filming a dangerous stunt for FAST & FURIOUS 7

REAL

RUMOR

4 tweets

retweet

Russell Crowe is **NOT** dead.

Another heinous celebrity death hoax took root online this morning with Crowe as the victim.

As was the case with previous "deaths," the actor was said to have suffered a fatal fall while filming in a remote location. Specifically, in the Hahnenkamm mountains of Austria.

New York radio station Z100 and other outlets reported the news as fact.

Fortunately, it's just another **vile, disgusting FAKE**.

The Crowe hoax comes from **FakeAWish.com**, the same disturbed "death" generator that's claimed previous victims such as **George**

(Manual) Fact Verification Web Sites (I)

PPRuNe
Professional Pilots Rumour Network

FactCheck.org

TruthOrFiction.com

Snopes.com
Rumor Has It

EL POLÍGRAFO
Chequeo de datos | La política contrastada

OpenSecrets.org
Center for Responsive Politics

CHEQUEADO•COM
La verificación del discurso público

UYCheck.com
¿nos dicen la verdad?

(Manual) Fact Verification Web Sites (II)

Global Summit of Fact-Checking in London, July 2015	**2015**	**2014**
Active fact-checking sites (tracking politicians' campaign promises)	64 (21)	44
Percentage of sites that use rating systems such as meters or labels	80	70
Sites that are affiliated with news organizations	63%	

http://reporterslab.org/snapshot-of-fact-checking-around-the-world-july-2015/

WikiLeaks

1.4 How WikiLeaks verifies its news stories

We assess all news stories and test their veracity. We send a submitted document through a very detailed examination a procedure. Is it real? What elements prove it is real? Who would have the motive to fake such a document and why? We use traditional investigative journalism techniques as well as more modern rtechnology-based methods. Typically we will do a forensic analysis of the document, determine the cost of forgery, means, motive, opportunity, the claims of the apparent authoring organisation, and answer a set of other detailed questions about the document. We may also seek external verification of the document For example, for our release of the Collateral Murder video, we sent a team of journalists to Iraq to interview the victims and observers of the helicopter attack. The team obtained copies of hospital records, death certificates, eye witness statements and other corroborating evidence supporting the truth of the story. Our verification process does not mean we will never make a mistake, but so far our method has meant that WikiLeaks has correctly identified the veracity of every document it has published.

Publishing the original source material behind each of our stories is the way in which we show the public that our story is authentic. Readers don't have to take our word for it; they can see for themselves. In this way, we also support the work of other journalism organisations, for they can view and use the original documents freely as well. Other journalists may well see an angle or detail in the document that we were not aware of in the first instance. By making the documents freely available, we hope to expand analysis and comment by all the media. Most of all, we want readers know the truth so they can make up their own minds.

Scaling Fact-Checking

Truthsquad™

Computational Journalism

ClaimBuster

Press Acknowledgement

2016 Republican Party Presidential Debate. Aug. 6, 2015, 8 p.m.

Venue: Quicken Loans Arena, Cleveland, Ohio. Broadcasted by: FOX.

Speakers: Bret Baier, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Mike Huckabee, John Kasich, Megyn Kelly, Rand Paul, Rick Perry, Marco Rubio, Donald Trump, Scott Walker, Chris Wallace

Transcript Source: http://time.com/3988276/republican-debate-primetime-transcript-full-text/

Chronological Order | Order by Score

Most Check-worthy >=1.0 <=0.9 <=0.8 <=0.7 <=0.6 <=0.5 <=0.4 <=0.3 <=0.2 <=0.1 Least Check-worthy

0.86 Secondly, we would needs test Social Security for those who are making over $200,000 dollars a year in retirement income, and have $4 to $5 million dollars in liquid assets saved.
0.80 I took the state of Ohio from an $8 billion hole and a 350,000 job loss to a $2 billion surplus and a gain of 350,000 jobs.
0.78 20 over 40 percent of small and mid-size banks that loan money to small businesses have been wiped out over the since Dodd-Frank has passed.
0.77 We came in, we balanced an $11 billion deficit on a $29 billion budget by cutting over 800 programs in the state budget.
0.77 Well, lets all be reminded, 60 million Americans are on Social Security, 60 million.
0.76 Nearly 6 million of you, 6 million, viewed the debate videos on our site, and more than 40,000 of you submitted questions: some of which you will hear us asking the candidates tonight.
0.76 We went from $1 billion of reserves to $9 billion of reserves.
0.75 Governor Kasich, You chose to expand Medicaid in your state, unlike several other governors on this stage tonight, and it is already over budget by some estimates costing taxpayers an additional $1.4 billion in just the first 18 months.
0.75 We created 1.3 million jobs.
0.73 And, finally, we went from $8 billion in the hole to $2 billion in the black.

Crowded Fact

TRUTHSQUAD ON HEALTHCARE

Orrin Hatch, U.S. Senator

"87 million Americans will be forced out of their coverage under new health care regulations from President Obama."

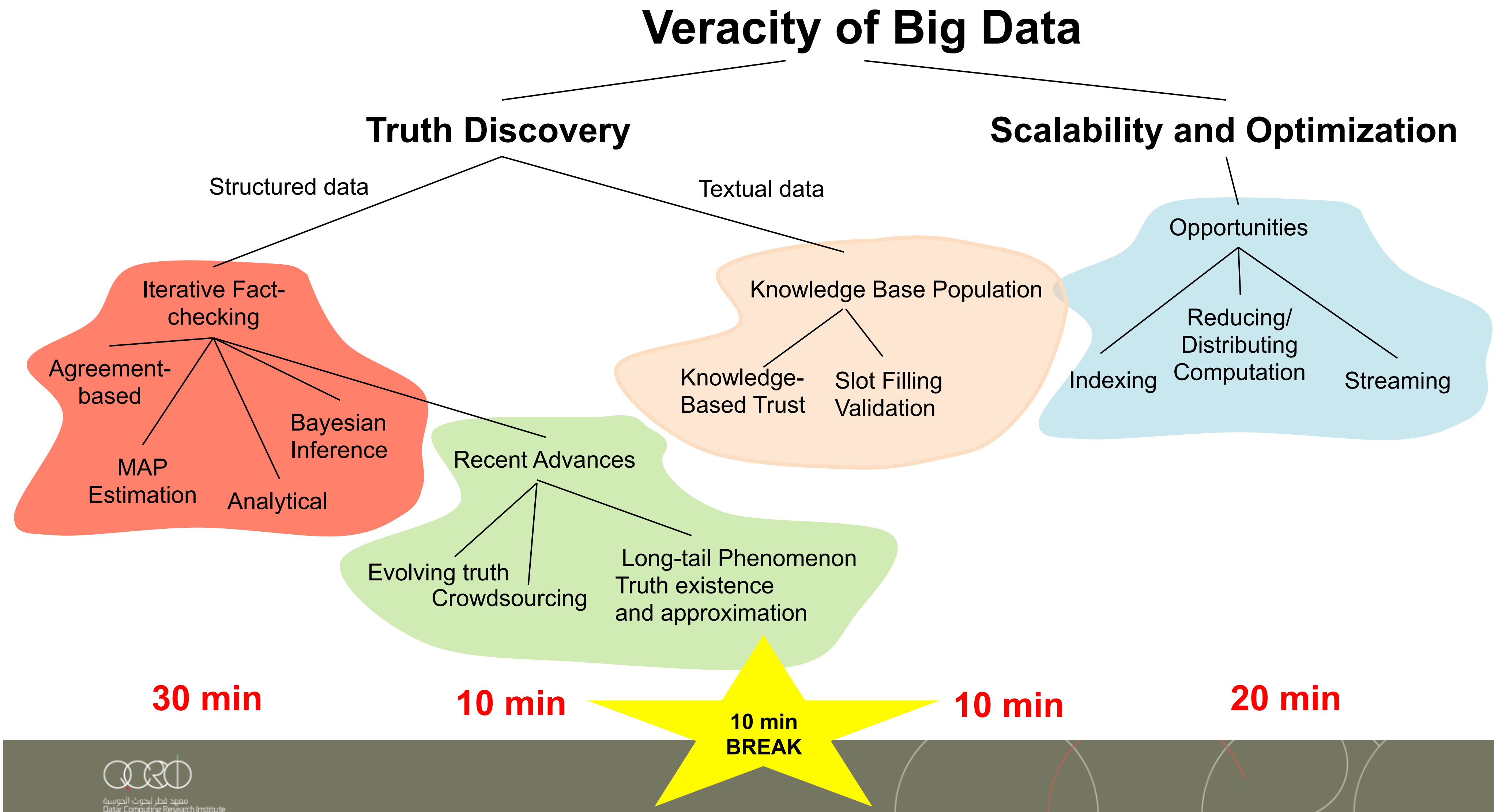
Fact-check this quote:

Is this true or false?

True | False | Not Sure

S. Cohen, J. T. Hamilton, and F. Turner. Computational journalism. CACM, 54(10):66–71, Oct. 2011.
N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In CIKM, 2015.
N.Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, C. Yu , The Quest to Automate Fact-Checking, C+J Symposium 2015
http://towknight.org/research/thinking/scaling-fact-checking/ http://blog.newstrust.net/2010/08/truthsquad-results.html

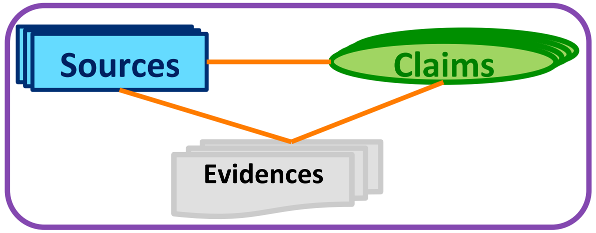
Tutorial Organization

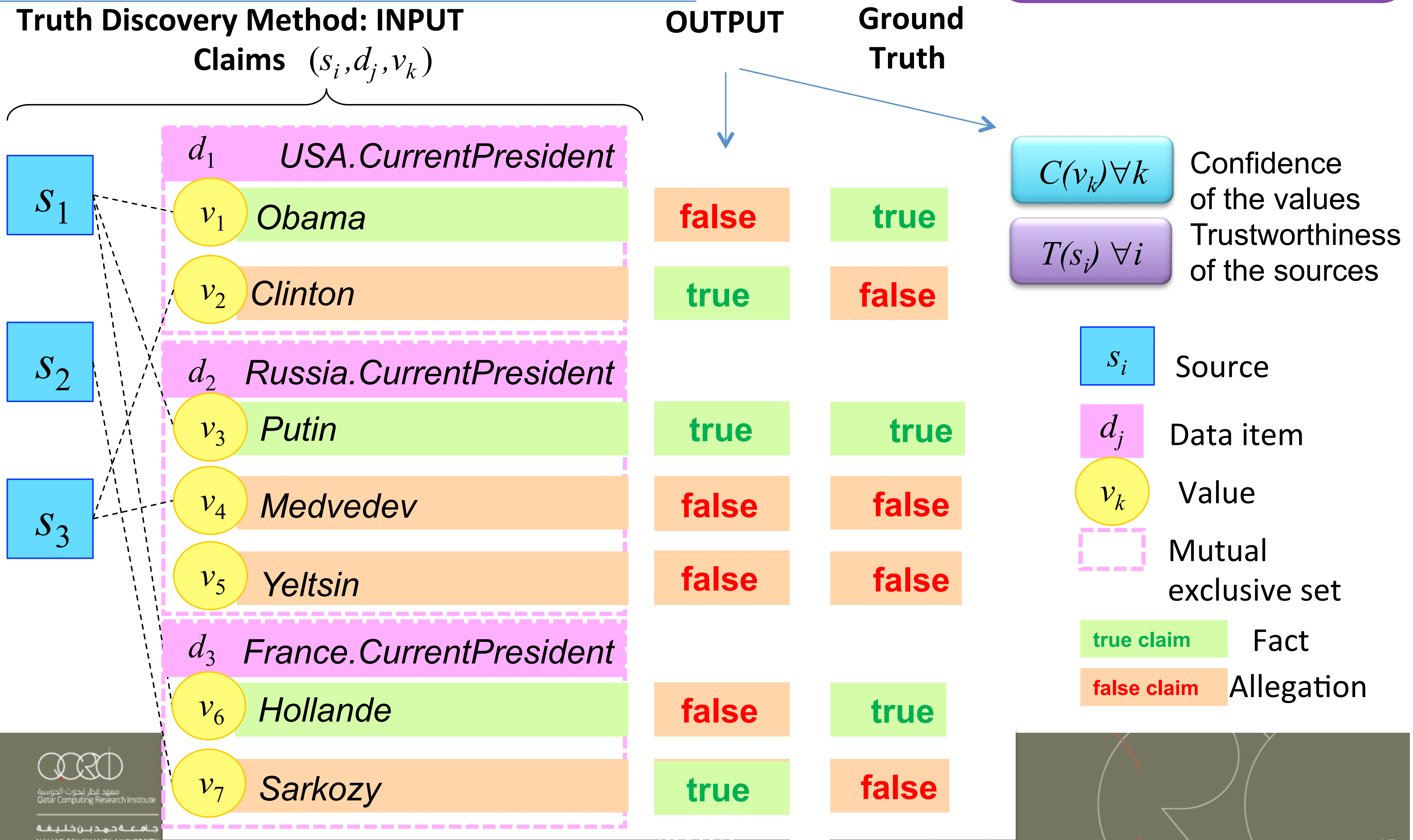


Outline

1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
4. Opportunities for scalability improvement
5. Conclusions

Terminology





Outline

1. Motivation
2. Truth Discovery from Structured Data
   - Agreement-based Methods
   - MAP Estimation-based Methods
   - Analytical Methods
   - Bayesian Methods

Agreement-Based Methods

Source Reputation Models

Source-Claim Iterative Models

Agreement-Based Methods

Source Reputation Models

Based on Web link Analysis

Compute the importance of a source in the Web graph based on the probability of landing on the source node by a random surfer

Hubs and Authorities (HITS) [Kleinberg, 1999]
PageRank [Brin and Page, 1998]
SourceRank [Balakrishnan, Kambhampati, 2009]

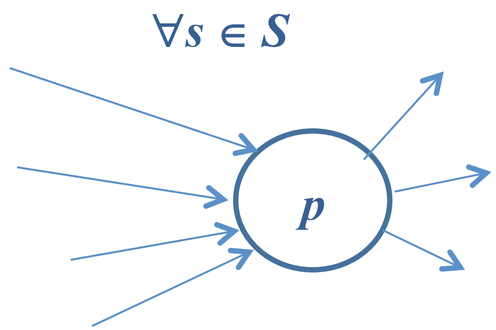
Trust Metrics: See R. Levien, Attack resistant trust metrics, PhD Thesis UC Berkeley LA, 2004

Agreement

Source Reputation

Hubs and Authorities (HITS)

- Identify Hub and Authority pages
- Each source p in S has two scores (at iteration *i*)
  – Hub score: Based on “outlinks”, links that point to other sources
  – Authority score: Based on “inlinks”, links from other sources



$$Hub^0(s) = 1$$

$$Hub^i(p) = \frac{1}{Z_h} \sum_{s \in S;\, p \to s} Auth^i(s)$$

$$Auth^i(p) = \frac{1}{Z_a} \sum_{s \in S;\, s \to p} Hub^{i-1}(s)$$

Z_a and Z_h are normalizers (L_2 norm of the score vectors)

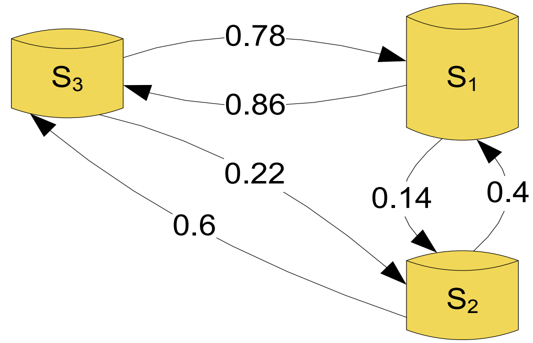
J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.

SourceRank

Agreement

Source Reputation

- Agreement graph: Markov chain with edges as the transition probabilities between the sources
- Source reputation is computed by a Markov random walk

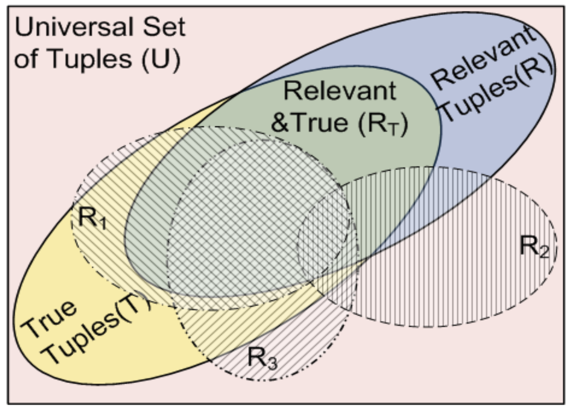


Probability of agreement of two independent false tuples

$$P_a(f_1, f_2) = \frac{1}{|U|}$$

Probability of agreement of two independent true tuples

$$P_a(r_1, r_2) = \frac{1}{|R_T|}$$



$$|U| \gg |R_T| \Rightarrow P_a(r_1, r_2) \gg P_a(f_1, f_2)$$

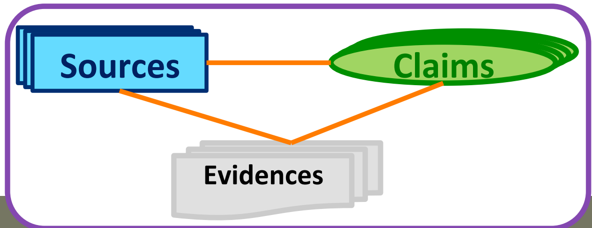
R. Balakrishnan, S. Kambhampati, SourceRank: Relevance and Trust Assessment for DeepWeb Sources Based on InterSource Agreement, In Proc. WWW 2009.

Agreement-Based Methods

Source Reputation Models

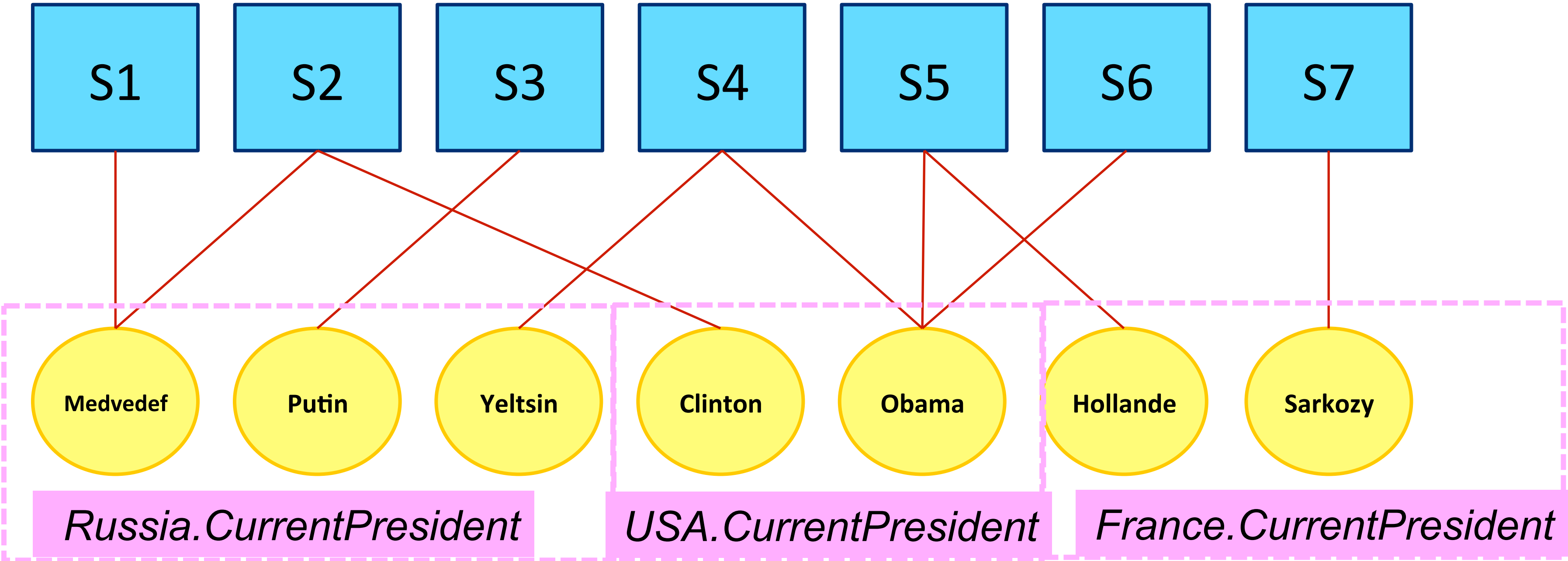
Only rely on source credibility is not enough

Source-Claim Iterative Models



Example

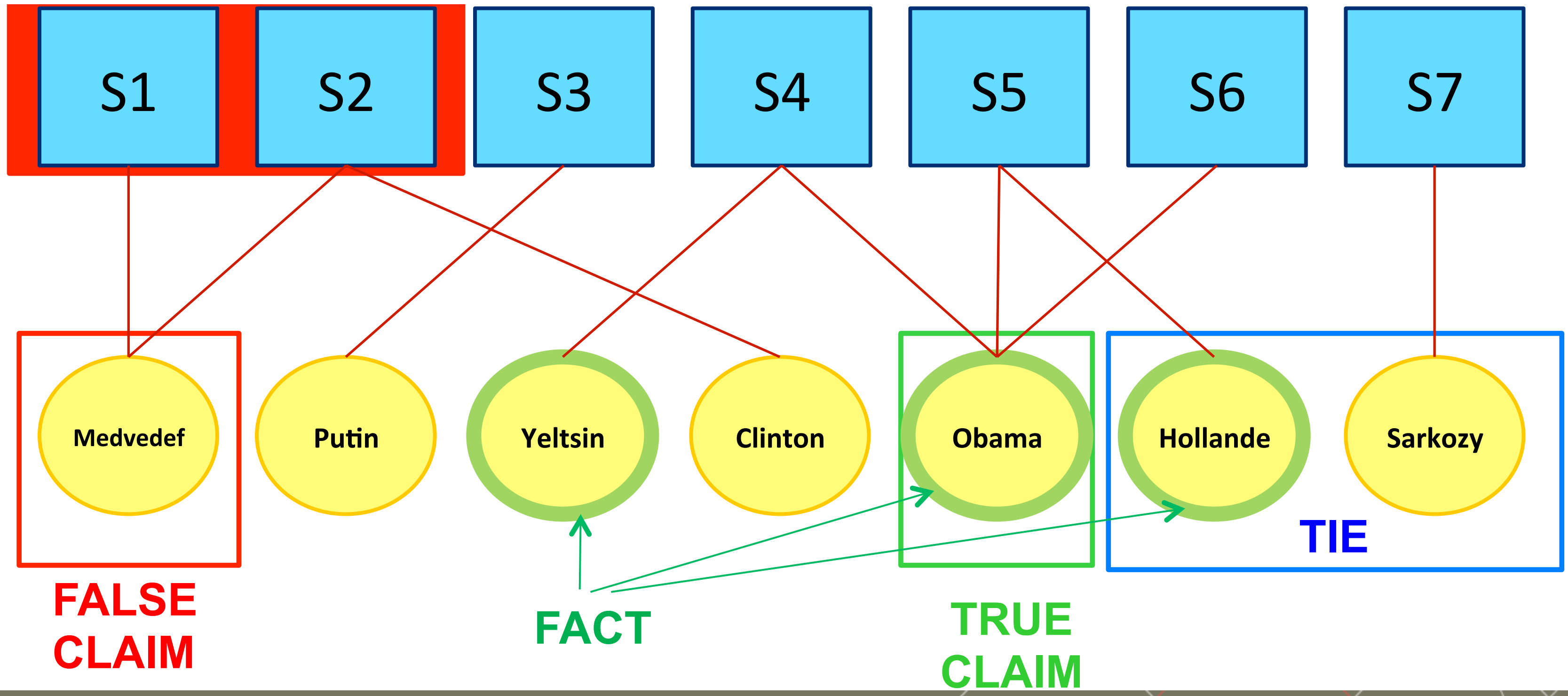
Seven sources disagree on the current president of Russia, Usa, and France
Can we discover the true values?



Solution: Majority Voting

Seven sources disagree on the current president of Russia, Usa, and France
Can we discover the true values?

Majority can be wrong!
What if these sources are not independent?



Majority Voting Accuracy : 1.5 out of 3 correct

Limit of Majority Voting Accuracy

Condorcet Jury Theorem (1785)

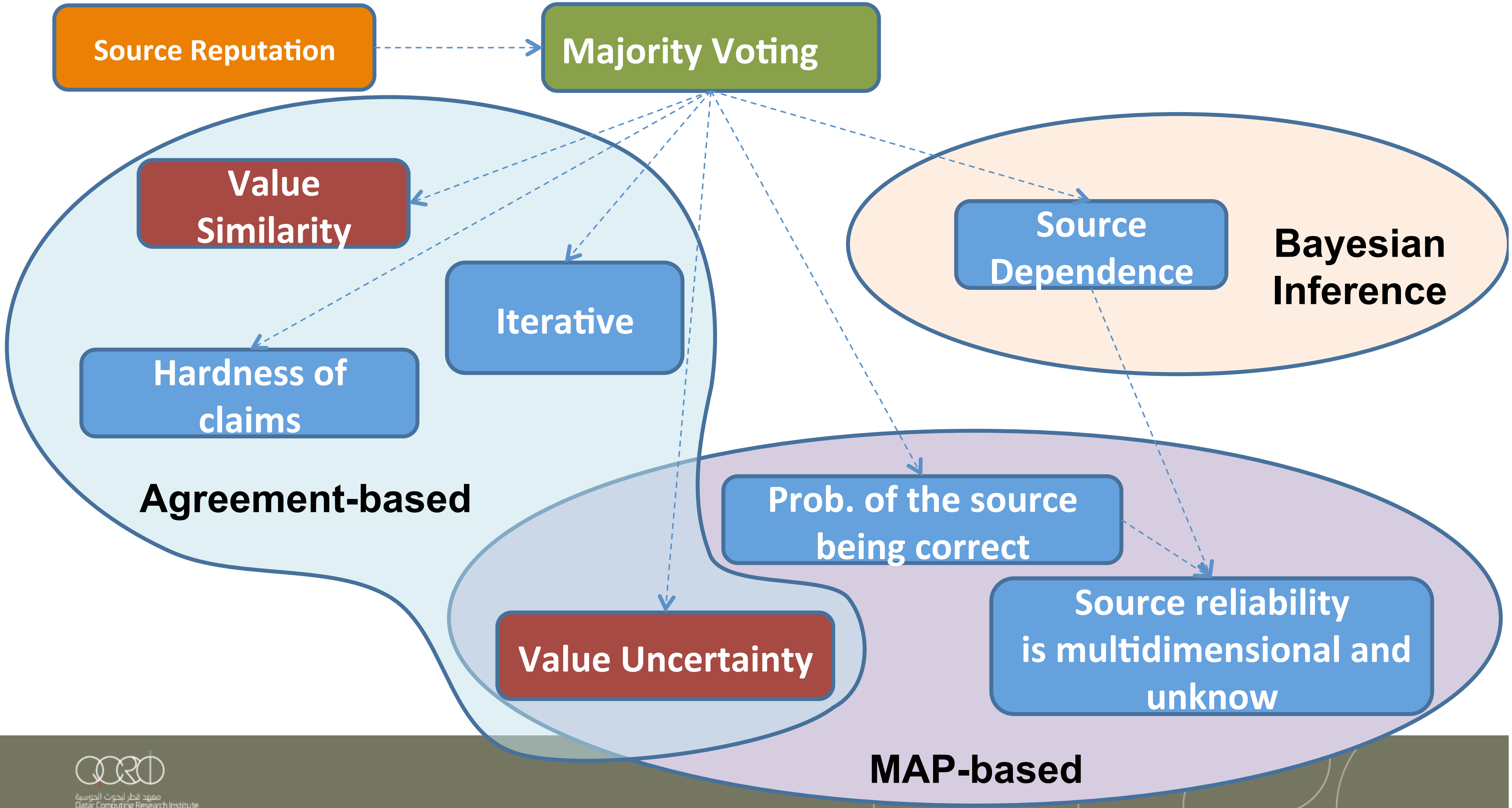
Originally written to provide theoritical basis of democracy

The majority vote will give an accurate value if at least ?$S/2+1$? independent sources give correct claims.
If each voter has a probability p of being correct, then the probability of the majority of voters being correct P_{MV} is

$$P_{MV} = \sum_{m=\lfloor S/2+1 \rfloor}^{S} \binom{S}{m} p^m (1-p)^{S-m}$$

- If $p > 0.5$, then P_{MV} is monotonically increasing, $P_{MV} \to 1$ as $S \to \infty$
- If $p < 0.5$, then P_{MV} is decreasing and $P_{MV} \to 0$ as $S \to \infty$
- If $p = 0.5$, then $P_{MV} = 0.5$ for any S

Roadmap of Modeling Assumptions

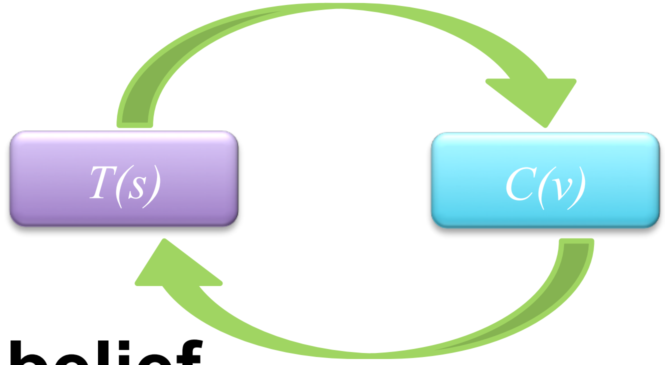


Agreement-Based Methods

Agreement

Source-Claim

Source-Claim Iterative Models



Based on iterative computation of source trustworthiness and claim belief

- Sums (adapted from HITS) *(1)*
- Average.Log, Investment, Pooled Investment *(1)*
- TruthFinder *(2)*
- Cosine, 2-Estimates, 3-Estimates *(3)*

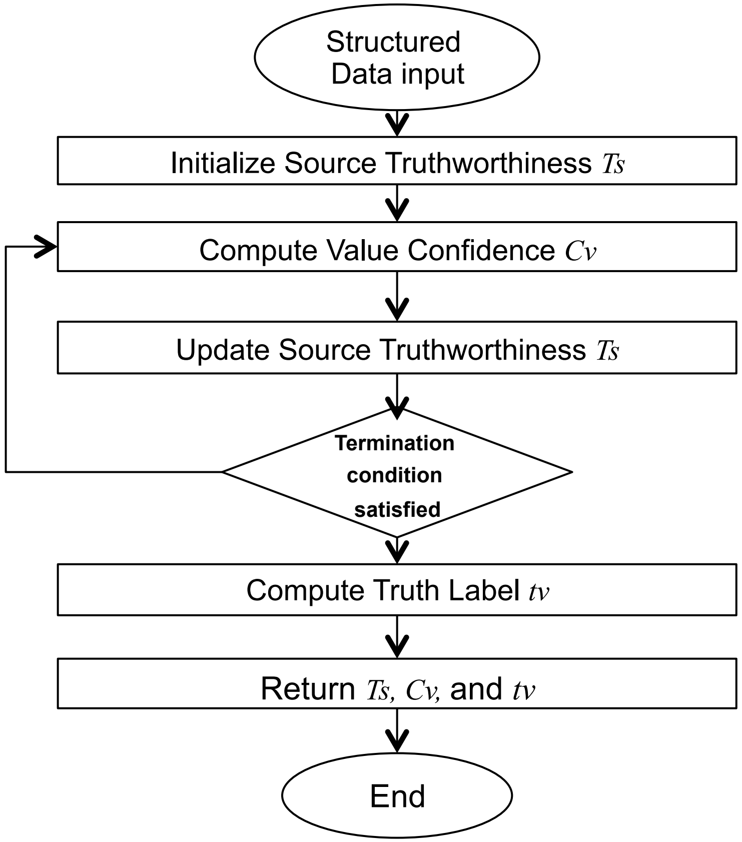
(1) J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In COLING, pages 877–885. Association for Computational Linguistics, 2010.

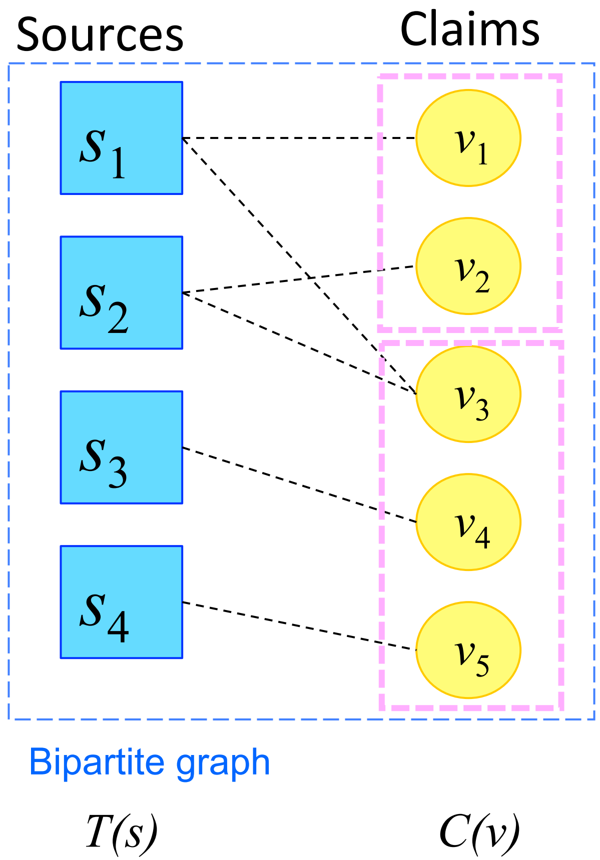
(2) X. Yin, J. Han, and P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. TKDE, 20(6):796–808, 2008.

(3) A. Galland, S . Abiteboul, A. Marian, P. Senellart. Corroborating Information from Disagreeing Views. In Proc. of the ACM International Conference on Web Search and Data Mining (WSDM), pages 131–140, 2010.

Basic Principle

Iterative and transitive voting algorithm





Agreement

Source-Claim

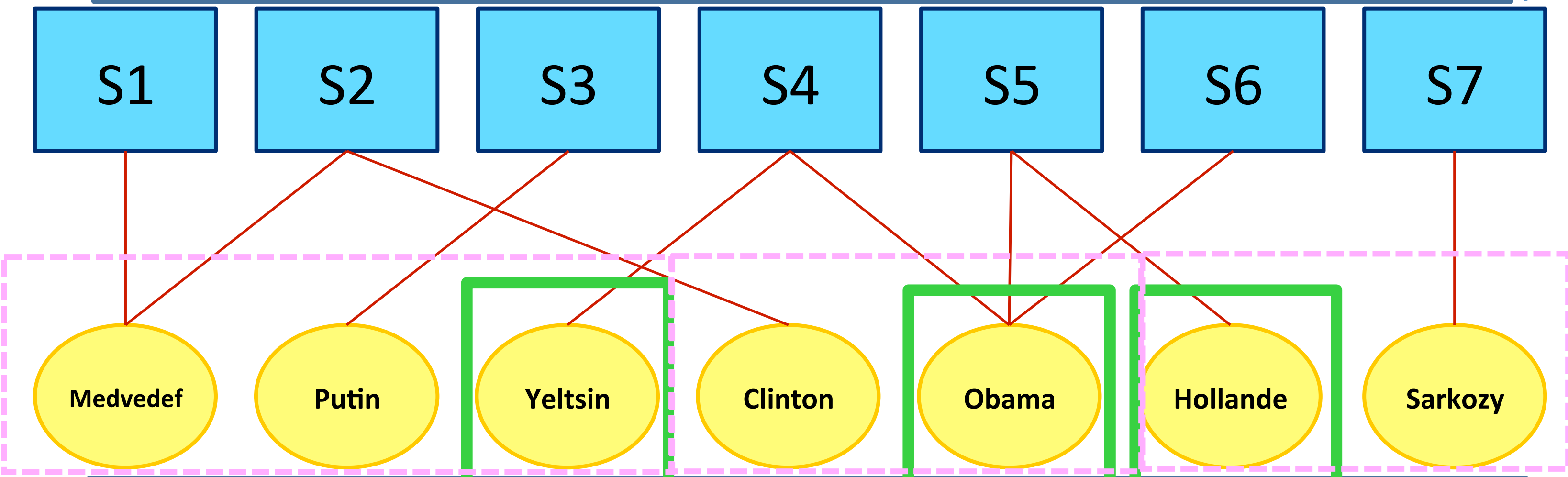
Example (cont'd)

Sums Fact-Finder: $T^i(s) = \sum_{v \in V_s} C^{i-1}(v)$ $\quad C^i(v) = \sum_{s \in S_v} T^i(s)$

Initialization: We believe in each claim equally

Source Trustworthiness T_s

	S1	S2	S3	S4	S5	S6	S7
Iteration 1:	1	2	1	2	2	1	1
Iteration 2:	3	5	1	7	7	5	1
Iteration 3:	8	13	1	26	26	19	1



Value Confidence C_v

	Medvedef	Putin	Yeltsin	Clinton	Obama	Hollande	Sarkozy
	1	1	1	1	1	1	1
Iteration 1:	3	1	2	2	5	2	1
Iteration 2:	8	1	7	5	19	7	1
Iteration 3:	21	1	26	13	71	26	1

Iterative Methods

Value Uncertainty

Agreement

Source-Claim

- Sums (adapted from HITS)

$$T^i(s) = \sum_{v \in V_s} \omega(s,v) C^{i-1}(v)$$

$$C^i(v) = \sum_{s \in S_v} \omega(s,v) T^i(s)$$

uncertainty

- Average.Log

$$T^i(s) = \log\left(\sum_{v \in V_s} \omega(s,v)\right) \cdot \frac{\sum_{v \in V_s} \omega(s,v) C^{i-1}(v)}{\sum_{v \in V_s} \omega(s,v)}$$

- Generalized Investment

$$T^i(s) = \sum_{v \in V_s} \frac{\omega(s,v) C^{i-1}(v) T^{i-1}(s)}{\sum_{v \in V_s} \omega(s,v) \cdot \sum_{r \in S_v} \frac{\omega(r,v) T^{i-1}(r)}{\sum_{b \in V_r} \omega(r,b)}}$$

$$C^i(v) = G\left(\sum_{s \in S_v} \frac{\omega(s,v) T(s)}{\sum_{v \in V_s} \omega(s,v)}\right) \quad \text{with } G(x) = x^{1.2}$$

J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In COLING, pages 877–885. Association for Computational Linguistics, 2010.

TruthFinder

Value Similarity

Agreement

Source-Claim

$$
\begin{array}{l}
\textit{Initialization.}\ \forall s \in S : T_s \leftarrow 0.8 \\
\textbf{repeat} \\
\quad \left\{ \begin{array}{l}
\textbf{for each } d \in D \\
\quad \textbf{do} \left\{ \begin{array}{l}
\textbf{for each } v \in V_d : \\
\quad \textbf{do} \left\{ \begin{array}{l}
\sigma_v \leftarrow -\sum_{s \in S_v} \ln(1 - T_s) \\
\sigma^\star_v \leftarrow \sigma_v + \rho \sum_{v' \in V_d} \sigma_{v'}.sim(v, v') \\
C_v \leftarrow \frac{1}{1+e^{-\gamma \sigma^\star_v}}
\end{array} \right.
\end{array} \right. \\
\textbf{for each } s \in S \\
\quad \textbf{do } T_s \leftarrow \frac{1}{|V_s|} \sum_{v \in V_s} C_v
\end{array} \right. \\
\textbf{until } Convergence(T_S, \delta) \\
\textbf{for each } d \in D \\
\quad \textbf{do } trueValue(d) \leftarrow \operatorname{argmax}_{v \in V_d}(C_v)
\end{array}
$$

We believe in each source equally (optimistic)

Probability to be wrong

Mutually supportive, similar values

Control parameter (ρ)

Confidence of each value

Dampening factor (γ) to compensate dependent similar values

Trustworthiness of each source

Thresholded cosine similarity of T_S between two successive iterations (δ)

X. Yin, J. Han, P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. TKDE, 20(6):796–808, 2008.

A Fine-grained Classification

1. Method Characteristics

- Initialization and parameter settings
- Repeatability
- Convergence and stopping criteria
- Complexity
- Scalability

Mono-valued: C1 (Source1,USA.CurrentPresident,Obama)
Multi-valued: C2 (Source1,Australia.PrimeMinitersList,
(Turnbull, Abott, Rudd, Gillard…))
Boolean: C3 (Source1,USA.CurrentPresident.Obama,Yes)

2. Input Data

- Type of data: categorical, string/text, continuous
- Mono- or multi-valued claims
- Similarity of claims
- Correlations between attributes or objects

3. Prior Knowledge and Assumptions

- Source Quality: Constant/evolving, non-/uniform across sources, homogeneous/ heterogeneous over data items
- Dependence of sources
- Hardness of certain claims

4. Output

- Single versus multiple true values per data item
- At least one or none true claim
- Enrichment with explanations and evidences

TruthFinder Signature

Agreement
Source-Claim

1. Method Characteristics	
❑ Initialization and parameter settings	T_s, $\boxed{?}$, $\boxed{?}$, $\boxed{?}$
❑ Repeatability	Yes
❑ Convergence and stopping criteria	$\boxed{?}$ for Cosine similarity of T_s
❑ Complexity	$O(Iter.SV)$
❑ Scalability	Yes
2. Input Data	
❑ Type of value	String, categorical, numeric
❑ Mono-/multi-valued claims	Mono- and Multi-valued claims
❑ Similarity of claims	Yes
❑ Correlations between attributes or objects	No
3. Prior Knowledge	
❑ Source Quality	Constant, uniform, homogeneous
❑ Dependence of sources	Yes (dampening factor)
❑ Hardness of certain claims	No
4. Output	
❑ Single/multiple truth per data item	Single true value per data item
❑ At least one or none true claim	At least one
❑ Enrichment (explanation/evidence)	No

Outline

1. Motivation
2. Truth Discovery from Structured Data
   - Agreement-based Methods
   - MAP-Estimation-based Methods
   - Analytical Methods
   - Bayesian Methods

Latent Credibility Analysis

MAP

EM

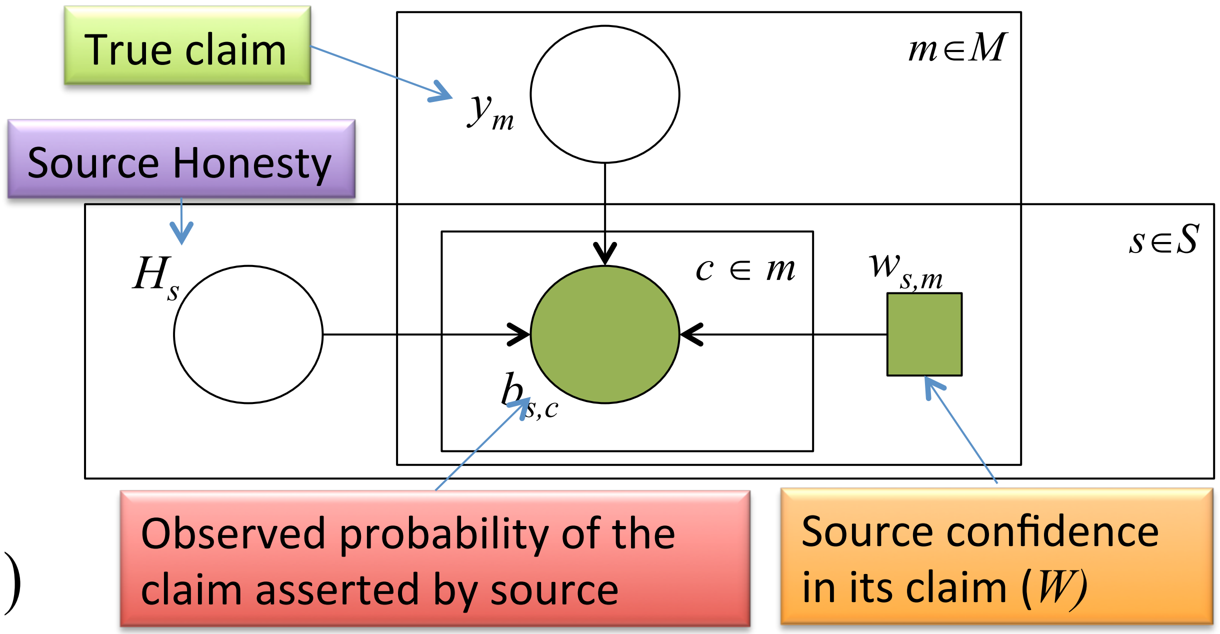
SimpleLCA, GuessLCA, MistakeLCA, LieLCA

Expectation-Maximization to find the maximum a posteriori (MAP) point estimate of the parameters

$$\theta^* = \arg\max_\theta P(X|\theta)P(\theta)$$

Then compute:

$$P(Y_U|X,Y_L,\theta^*) = \frac{P(Y_U,X,Y_L|\theta^*)}{\sum_{Y_U} P(Y_U,X,Y_L|\theta^*)}$$



Latent variables ?

- H_s: probability s makes honest, accurate claim
- D_m: probability s knows the true claims in m

J. Pasternack, D. Roth. Latent credibility analysis. In Proceedings of the 22nd International Conference on WWW 2013.

LCA Signature

MAP

EM

1. Method Characteristics

- ❑ Initialization and parameter settings
- ❑ Repeatability
- ❑ Convergence and stopping criteria
- ❑ Complexity
- ❑ Scalability

2. Input Data

- ❑ Type of value
- ❑ Mono-/multi-valued claims
- ❑ Similarity of claims
- ❑ Correlations between attributes or objects

3. Prior Knowledge

- ❑ Source Quality
- ❑ Dependence of sources
- ❑ Hardness of certain claims

4. Output

- ❑ Single/multiple truth per data item
- ❑ At least one or none true claim
- ❑ Enrichment (explanation/evidence)

W, K, $\boxed{?}_1$ *(prior truth prob./claim)*
Yes
***K* iterations**
$O(KSD)$
Yes

String, categorical
Multi-valued
Yes (as joint probability)
No

Constant, source- and entity-specific
No
Yes

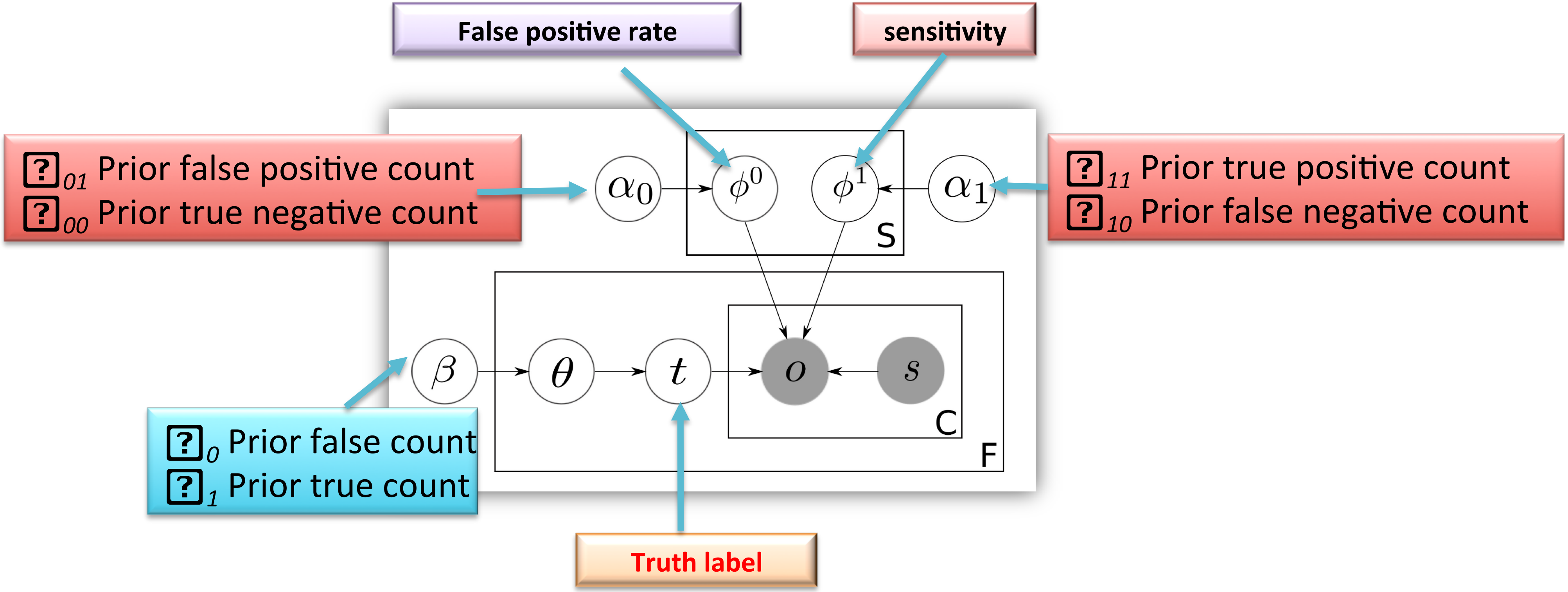
Single true value per data item
At least one
No

Latent Truth Model (LTM)

MAP

Gibbs Sampling

Collapsed Gibbs sampling to get MAP estimate for t



B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment, 5(6):550-561, 2012.

LTM Signature

MAP

Gibbs Sampling

1. Method Characteristics

- Initialization and parameter settings
- Repeatability
- Convergence and stopping criteria
- Complexity
- Scalability

2. Input Data

- Type of value
- Mono-/multi-valued claims
- Similarity of claims
- Correlations between attributes or objects

3. Prior Knowledge

- Source Quality
- Dependence of sources
- Hardness of certain claims

4. Output

- Single/multiple truth per data item
- At least one or none true claim
- Enrichment (explanation/evidence)

*(T_s, **K**, Burn-in, Thin,*
?$_{00}$,?$_{00}$, ?$_{01}$,?$_{01}$, ?$_{10}$,?$_{10}$, ?$_{11}$,?$_{11}$)
No (Gibbs sampling)
***K* iterations**
$O(KSV)$
Yes

String, categorical
Mono-valued (multiple claims/per source)
No
No

Incremental, source-specific, homog./entity
No
No

Multiple true values per data item
At least one
No

Outline

1. Motivation
2. Truth Discovery from Structured Data
    - Agreement-based Methods
    - MAP Estimation-based Methods
    - Analytical Methods
    - Bayesian Methods

Analytical Solutions

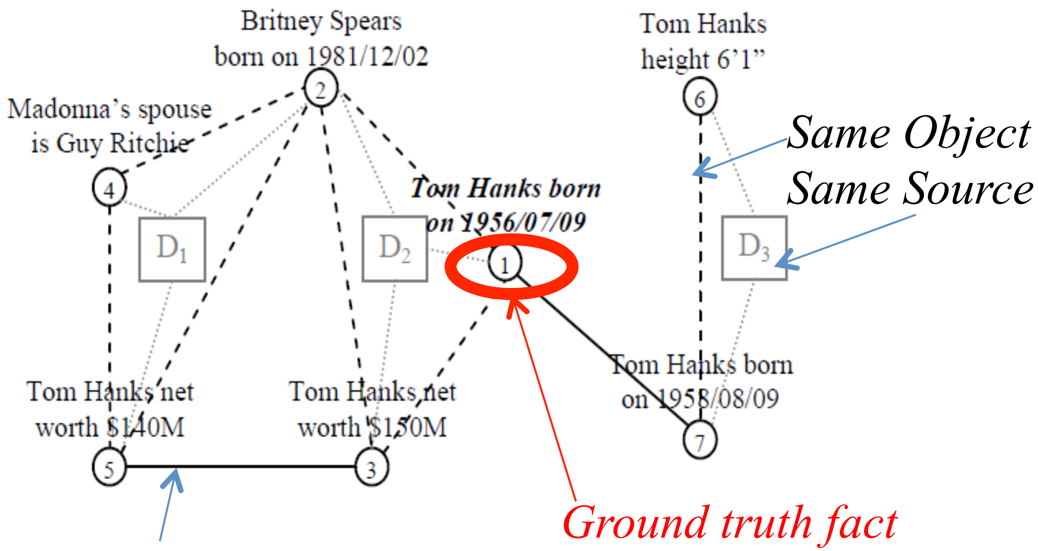
Matrix Diagonalization

Semi-Supervised Truth Discovery (SSTF)

Minimize loss funtion

$$E(\mathrm{C}) = \frac{1}{2}\sum_{i,j}\left|w_{ij}\right|(c_i - s_{ij}c_j)^2$$

$$\text{where } s_{ij} = \begin{cases} 1 \text{ if } w_{ij} \geq 0 & \text{Supportive claims} \\ -1 \text{ if } w_{ij} < 0 & \text{Claims in conflict} \end{cases}$$



w_{ij} relationship between confidence scores

$$\left.\frac{\partial E}{\partial c}\right|_{c=c^*} = 0 \Leftrightarrow (D_{uu} - W_{uu})C_u - W_{ul}C_l = 0$$

Weight Matrices

Matrix of unlabeled claim confidence scores

X. Yin, W. Tan. Semi-supervised Truth Discovery. In *Proceedings of the 20th international conference* WWW '11, 2011.

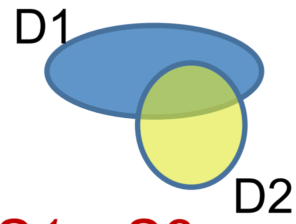
Related Work: L. Ge, J. Gao, X. Yuy, W. Fanz and A. Zhang, Estimating Local Information Trustworthiness via Multi-Source Joint Matrix Factorization, Proc. of ICDM 2012

Outline

1. Motivation
2. Truth Discovery from Structured Data
    - Agreement-based Methods
    - MAP Estimation-based Methods
    - Analytical Methods
    - Bayesian Methods

Bayesian

Source Dependence



- Sharing the same errors is unlikely if sources are independent
- Accuracy differences give the copying direction

$|Acc(D1 \cap D2)\text{-}Acc(D1\text{-}D2)| > |Acc(D1 \cap D2)\text{-}Acc(D2\text{-}D1)| \Rightarrow S1 \rightarrow S2$

Source Accuracy

$$Acc(S) = \underset{v \in V_S}{Avg}\left(P(V_s)\right)$$

Source Vote Count

$$A'(S) = \ln\left(\frac{n_f Acc(S)}{1 - Acc(S)}\right)$$

Consider dependence

$I(S)$ Prob. of independently providing value v

ValueVote Count

$$C(v) = \sum_{S \in S_v} A'(S).I(S)$$

Consider value similarity

$$C''(v) = C(v) + \rho \sum_{v' \neq v} C(v') \cdot sim(v, v')$$

Value Probability

$$\Pr(v \text{ true} \mid \Phi) = \frac{e^{C(v)}}{\sum_{v_0 \in V_d} e^{C(v_0)}}$$

X. L. Dong, L. Berti-Equille, D. Srivastava. Integrating conflicting data: the role of source dependence. In VLDB, 2009

X. L. Dong, L. Berti-Equille, Y. Hu, D. Srivastava. Global detection of complex copying relationships between sources. In VLDB, 2010

Depen Signature

Bayesian

1. Method Characteristics	
❑ Initialization and parameter settings	T_s, n_f *(nb false value)*, ε (error rate) , ? (a priori prob.), c *(copying prob.)*, δ
❑ Repeatability	Yes
❑ Convergence and stopping criteria	δ
❑ Complexity	$O(Iter.S^2V^2)$
❑ Scalability	**No**[(1)]
2. Input Data	
❑ Type of value	String, categorical, numerical
❑ Mono-/multi-valued claims	Multi-valued
❑ Similarity of claims	Yes
❑ Correlations between attributes or objects	**No**[(2)]
3. Prior Knowledge	
❑ Source Quality	Contant, uniform across sources , homogeneous across objects
❑ Dependence of sources	Yes
❑ Hardness of certain claims	No
4. Output	
❑ Single/multiple truth per data item	Single true values per data item
❑ At least one or none true claim	At least one
❑ Enrichment (explanation/evidence)	No

(1) X. Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Scaling up Copy Detection. In ICDE, 2015.

(2) R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, D. Srivastava. Fusing data with correlations. In *SIGMOD*, 2014.

Modeling Assumptions

Source

(*)***Relaxed in***

- Sources are **self-consistent**: a source does not claim conflicting claims
- The probability a source asserts a claim is independent of the truth of the claim
- Sources make their claims **independently**(1)
- A source has **uniform confidence** to all the claims it expresses(2)
- **Trust the majority**
- **Optimistic scenario : $S_{True} >> S_{False}$**

(1)[Dong et al, VLDB'09]

(2)[Pasternack Roth, WWW'13]

Claims

- Only claims with a **direct source attribution** are considered
  e.g., "S 1 claims that S2 claims A" is not condidered
- Claims are assumed to be **positive** and usually certain:
  e.g., "S claims that A is false", "S does not claim A is true" are not considered
  or "S claims that A is true with 15% uncertainty" (2)
- Claims claimed by only one source are true
- Correlations between claims/entity are not considered(3)
- One single true value exists(4)

(3)[Pochampally et al. SIGMOD'14]

(4)[Zhi et al., KDD'15]

Recap

	Truthfinder	MLE	LCA	LTM	Depen+	SSTF
Data Type	String, Categorical Numerical	**Boolean**	String, Categorical	String, Categorical	String, Categorical Numerical	String, Categorical Numerical
Mono/multi-valued claim	**Mono & Multi**	Mono	**Multi**	Mono	**Mono & Multi**	Mono
Similarity	Yes	No	Yes	No	Yes	Yes
Correlations	No	No	No	No	Yes+	Yes
Source Quality	Constant, uniform	Constant, Source-specific	Constant, **Source- and data item specific**	**Incremental, source-specific**	Constant, uniform	Constant, uniform
Source Dependence	No	No	No	No	Yes	No
Claim hardness	No	No	Yes	No	No	No
Single/multi-truth	Single	Single	Single	**Multi-truth**	Single	Single
Trainable	No	No	No	No	No	Yes

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

Outline

1. Motivation

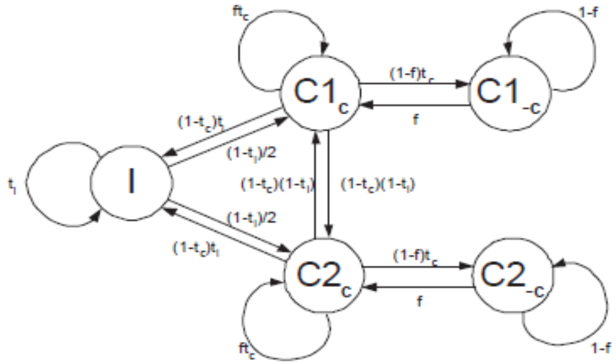
2. Truth Discovery from Structured Data

Recent Advances for Structured Data

- Evolving Truth
- Truth Finding from Crowdsourced Data
- Long-Tail Phenomenon
- Truth Existence, and Approximation

Evolving Truth

- **True values can evolve over time**
  – Lifespan of objects
  – Coverage, Exactness, Freshness of source
  – HMM model to detect lifespan and copying relationships



X. L. Dong, L. Berti-Equille, D. Srivastava. Truth discovery and copying detection in a dynamic world. In VLDB 2009.

- **Source quality changes over time**
  – MAP estimation of the source weights

Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W.Fan, J. Han. On the discovery of evolving truth. In KDD 2015.

- **New sources can be added**
  – Incremental voting over multiple trained classifiers
  – Concept drift

L. Jia, H. Wang, J. Li, H. Gao, Incremental Truth Discovery for Information from Multiple Sources. In WAIM 2013 workshop, LNCS 7901, p. 56-66, 2013

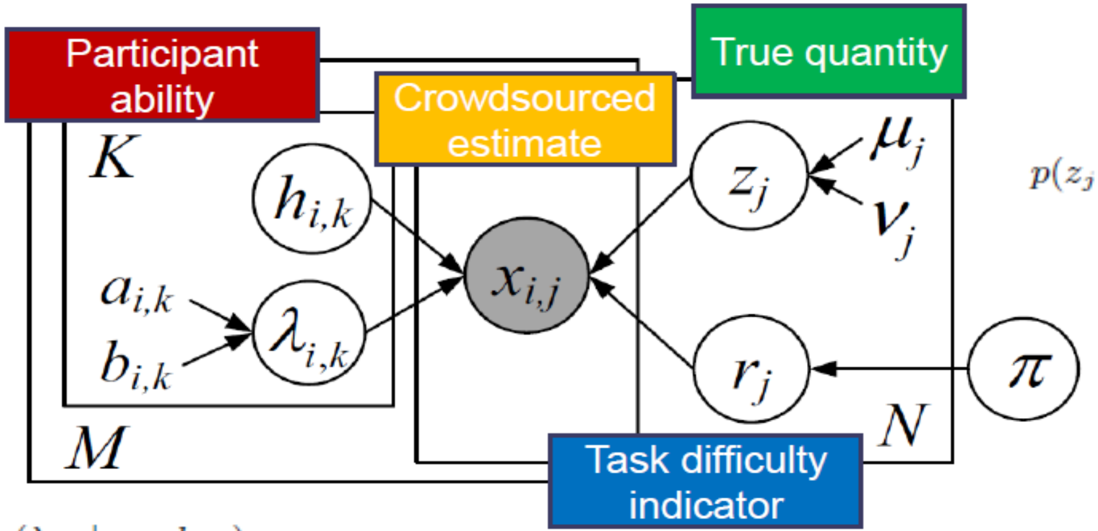
Truth discovery from crowdsourced data

Expectation Maximization

TBP (Truth Bias and Precision)

Likelihood of observing a crowdsourced estimate (given model parameters only) follows a mixture distribution

$$p(x_{i,j}|\boldsymbol{\pi}, z_j, h_{i,k}, \lambda_{i,k}) = \sum \pi_k \mathcal{N}(x_{i,j}|z_j + h_{i,k}, 1/\lambda_{i,k})$$



$$p(z_j|\mu_j, \nu_j) = \mathcal{N}(z_j|\mu_j, 1/\nu_j)$$

$$p(r_{j,k}|\pi) = \mathrm{Mul}(\mathbf{r}_j|\pi)$$

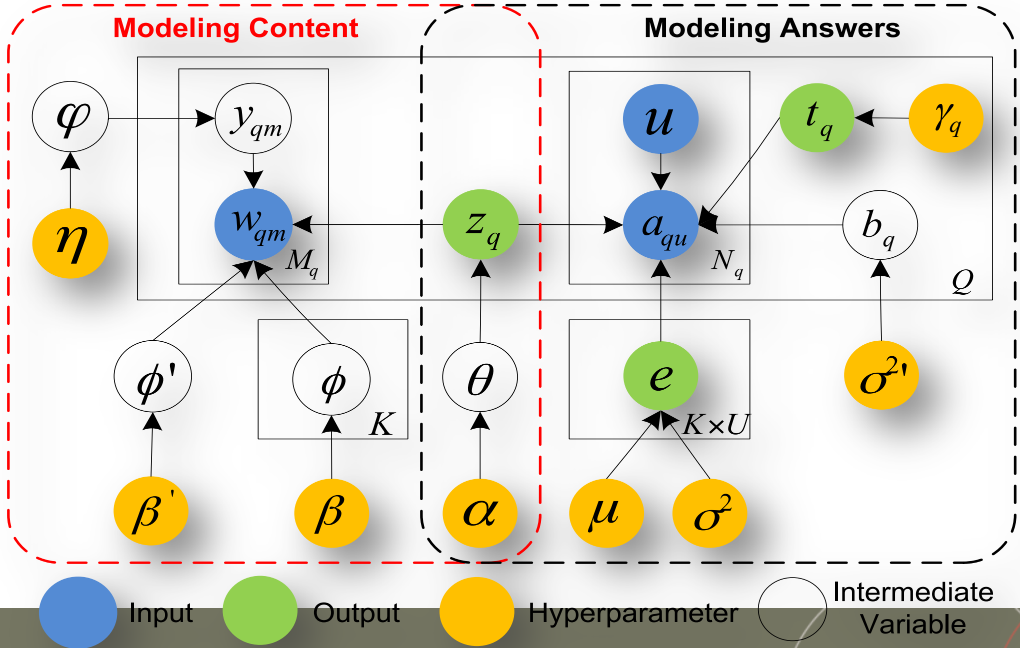
$$p(\lambda_{i,k}|a_{i,k}, b_{i,k}) = \mathrm{Gamma}(\lambda_{i,k}|a_{i,k}, b_{i,k})$$

R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, and T. J. Norman. Debiasing crowdsourced quantitative characteristics in local businesses and services. Proc. of IPSN ACM/IEEE, pp. 190-201, 2015.

Truth discovery from crowdsourced data

Faitcrowd

- **Input:** Q questions, K topics, M_q words and N_q answers per question provided by U users, hyperparameters
- **Output:** User expertise ***e***, true answers t_q, question topic labels z_q



$$t_q \sim U(\gamma_q)$$

$$b_q \sim N(0, \sigma^{2\prime})$$

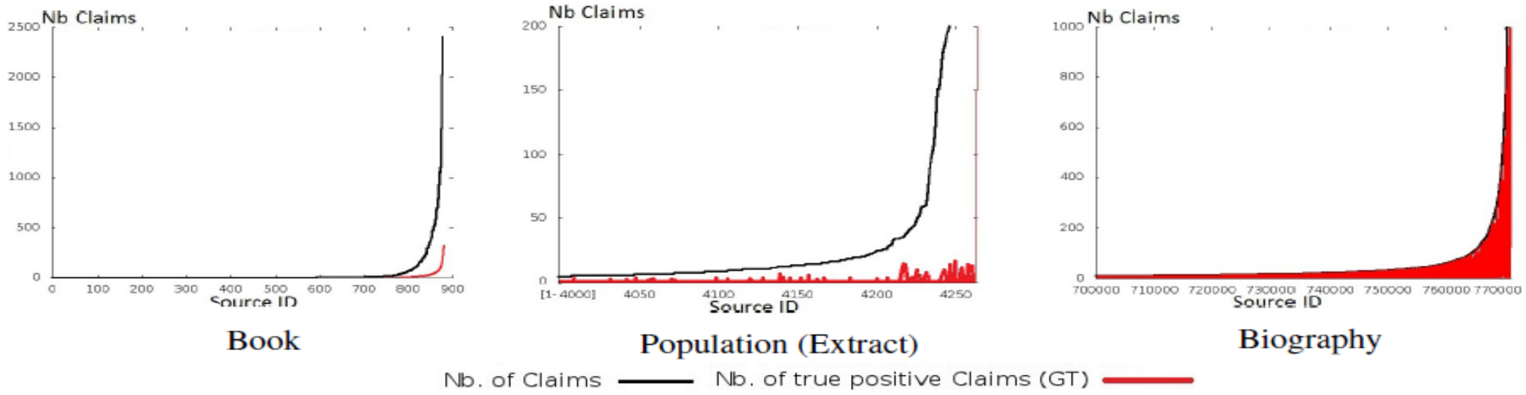
$$a_{qu}|t_q \sim logistic(e_{z_q u}, b_q)$$

$$e_{z_q u} \sim N(\mu, \sigma^2)$$

F. Ma, Y. Li, Q. Li, M. Qui, 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 KDD 2015.

Variance Minimization

Long-Tail Phenomenon



CADT Method for Independent and Benevolent Sources

Goal : Minimize the Variance of Source Reliability $\varepsilon_s \propto N(0,\sigma_s^2)$

$$\varepsilon_{combined} = \frac{\sum_{s\in S} w_s \varepsilon_s}{\sum_{s\in S} w_s}$$

$$\min_{w_s} \sum_{s\in S} w_s^2 \sigma_s^2 \quad \text{s.t.} \sum_{s\in S} w_s = 1, w_s \geq 0, \forall s \in S$$

$$w_s \propto \frac{\chi^2_{(\alpha/2, N_s)}}{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}$$

Number of claims by source s

Chi-squared probability at (1-?) confidence interval

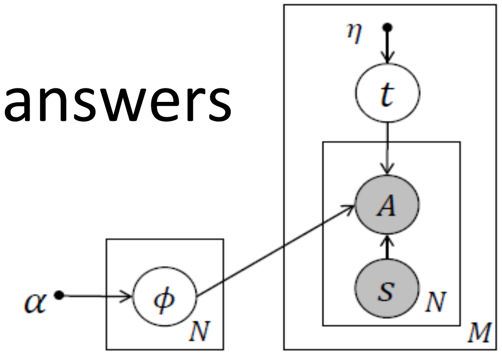
Initial value confidence for entity n

Reliability of source s

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M.Demirbas, W. Fan, and J. Han. 2014. A confidence-aware approach for truth discovery on long-tail data. Proc. VLDB Endow. 8, 4 (December 2014), 425-436.

Recent contributions

- **Modeling Truth Existence**
  - Problem of *No-truth* questions: none of the answers is true
  - EM-based algorithm similar to MLE
  - Silent rate, false and true spoken rates



S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, J. Han. Modeling Truth Existence in Truth Discovery. In Proc. of KDD

- **Multi-Truth Discovery**

X. Wang, X. Xu, X. Li. An Integrated Bayesian Approach for Effective Multi-Truth Discovery. In ICDE 2016

- **Approximate Truth Discovery**

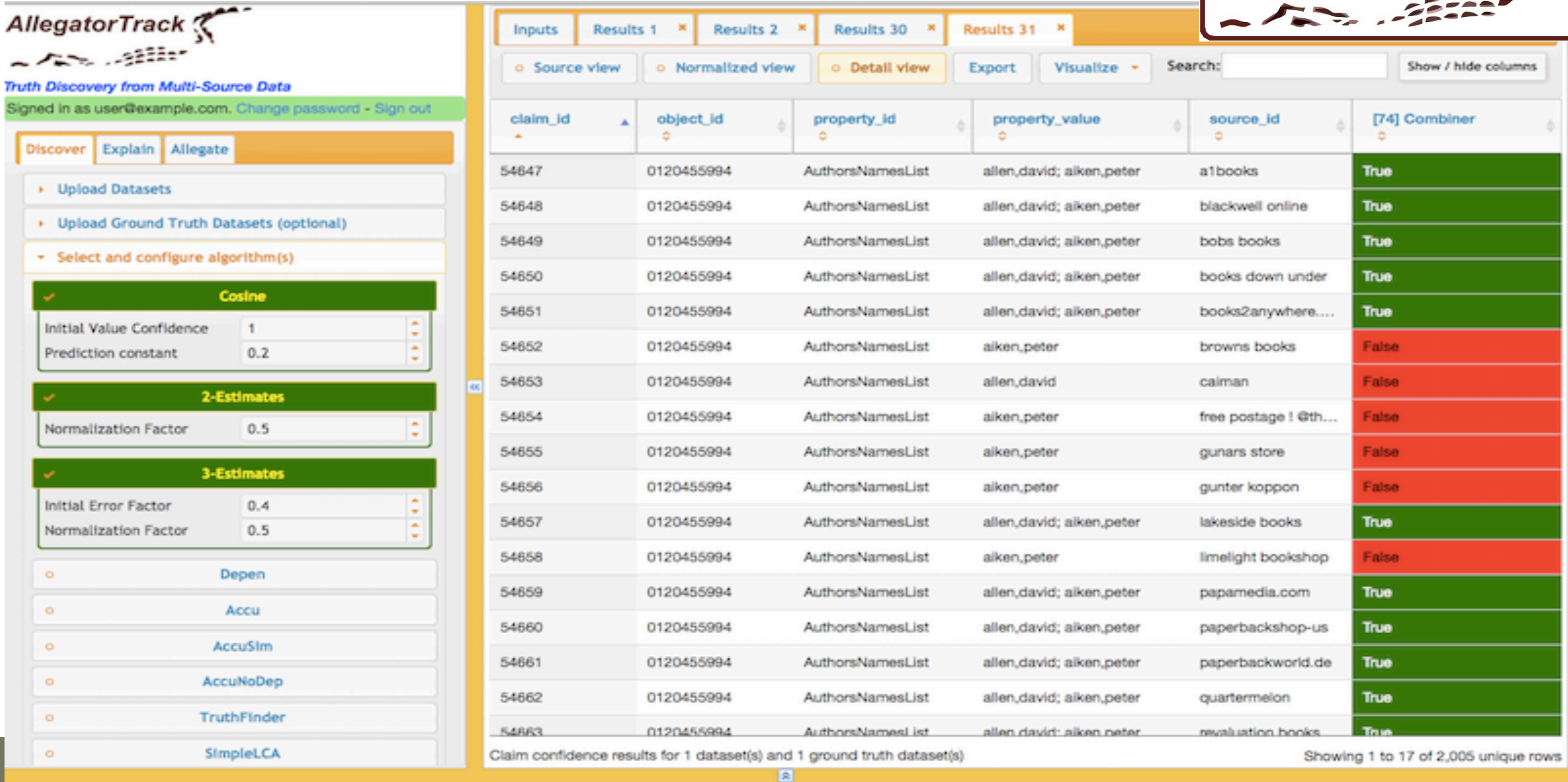
X. Wang, Q. Z. Sheng, X. S. Fang, X. Xu, X. Li, L. Yao. Approximate Truth Discovery Via Problem Scale Reduction. In ICDE 2016

Further Testing

API

AllegatorTrack

http://daqcri.github.io/dafna/



D. Attia Waguih, N. Goel, H. M. Hammady, L. Berti-Equille. AllegatorTrack: Combining and Reporting Results of Truth Discovery from Multi-source Data. *In ICDE 2015.*

Further Testing

API

AllegatorTrack

http://daqcri.github.io/dafna/

- Datasets and Synthetic Data Generator

```
java -jar DAFNA-DataSetGenerator.jar
        -src 10 -obj 10 -prop 5 -cov 1.00 -ctrlC Exp -ctrlT Exp -v 3 -ctrlV Exp -s dissSim -f "./Test"
```

Control Parameter	Value
Number of sources (S)	50; 1,000; 2,000; 5,000; 10,000
Number of data items (D)	100; 1,000; 10,000
Source Coverage (Cov)	U.25; U.75 (Uniform) L (Linear) E (Exponential)
Ground Truth (GT)	R (Random) U.25; U.75 (Uniform) FP (Fully Pessimistic) FO (Fully Optimistic) 80-P (80-Pessimistic) 80-O (80-Optimistic) E (Exponential)
Conflict Distribution (Conf)	U (Uniform) E (Exponential)
Number of Distinct Values	2...20

Outline

1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
   - Knowledge-Based Trust
   - Slot Filling Validation

Knowledge-Based Trust

Bayesian

EM

Distinguish extractor errors from source errors

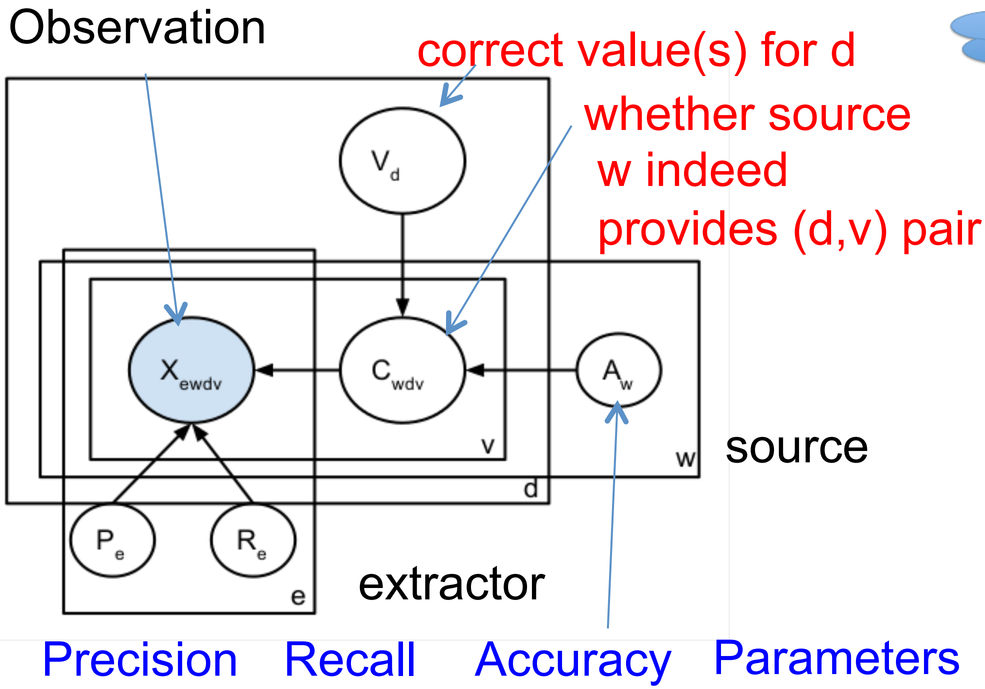
Multi-Layer Model based on EM

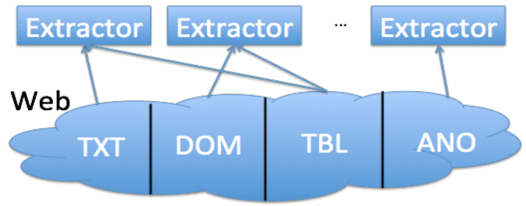
Compute P(w provide v_d | extractor quality)

Compute P(v_d | source quality)

Compute source accuracy

Compute Precision Recall of extractor





#Triples	3.0B (0.3B w. pr>=0.7)
#URLs	2.5B (28M Websites)
#Extractors	16

As of 2014

X. L. Dong, K. Murphy, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, W. Zhang. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion, In VLDB 2015

Slot Filling Validation

Method **extending Co-HITS** [Deng *et al.* 2009] over **heterogeneous networks**

Credibility Propagation

1. Initialize credibility scores c^0 for S to 1, for T with TextRank [Mihalcea 2004] and for R using linguistic indicators

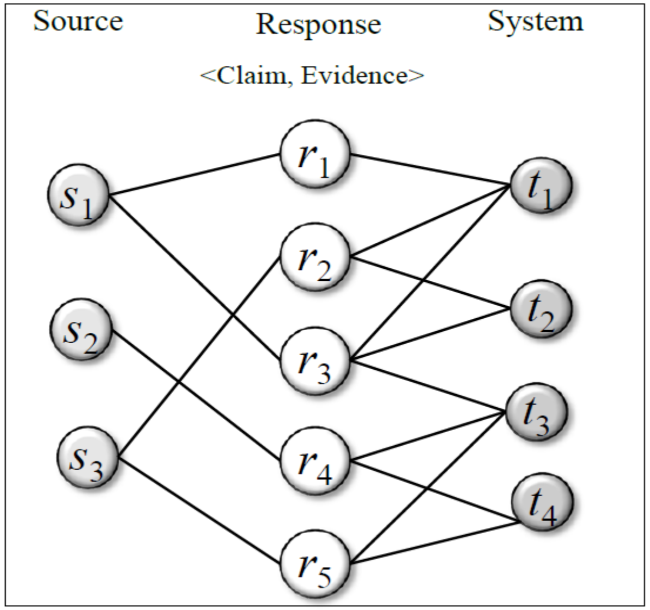
2. Construct heterogeneous networks across R, S and T with transition prob.

$$p_{ij}^{rs} = \frac{w_{ij}^{rs}}{\sum_k w_{ik}^{rs}}$$

3. Compute:

$$\begin{cases} c(s_i) = (1-\lambda_{rs})c^0(s_i) + \lambda_{rs} \sum_{r_j \in R} p_{ji}^{rs} c(r_j) \\ c(t_k) = (1-\lambda_{rt})c^0(t_k) + \lambda_{rt} \sum_{r_j \in R} p_{jk}^{rt} c(r_j) \\ c(r_j) = (1-\lambda_{sr}-\lambda_{tr})c^0(r_j) \end{cases}$$

$$+ \lambda_{sr} \sum_{s_i \in S} p_{ij}^{sr} c(s_i) + \lambda_{tr} \sum_{t_k \in T} p_{kj}^{tr} c(t_k)$$



W^{sr} W^{rt}

Weight matrices

D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. R. Voss, M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In COLING 2014, p. 1567–1578, 2014

Outline

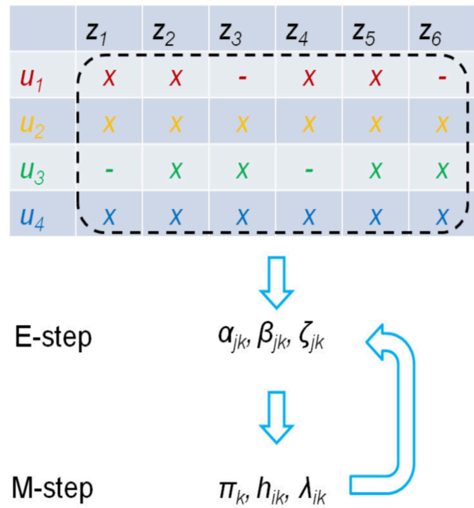
1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
4. Opportunities for scalability improvement
5. Conclusions

Scalability issues

Pairwise comparisons of sources covering the same data items

For EM-based approaches:

1. Each update needs all the data set: “out of memory” problem
2. The algorithm needs to iterate over the whole dataset several times until convergence
1. In M-step Optimal hidden variables do not have a closed-form joint optimization is required



(a) Batch truth discovery

Specialized Inverted Index

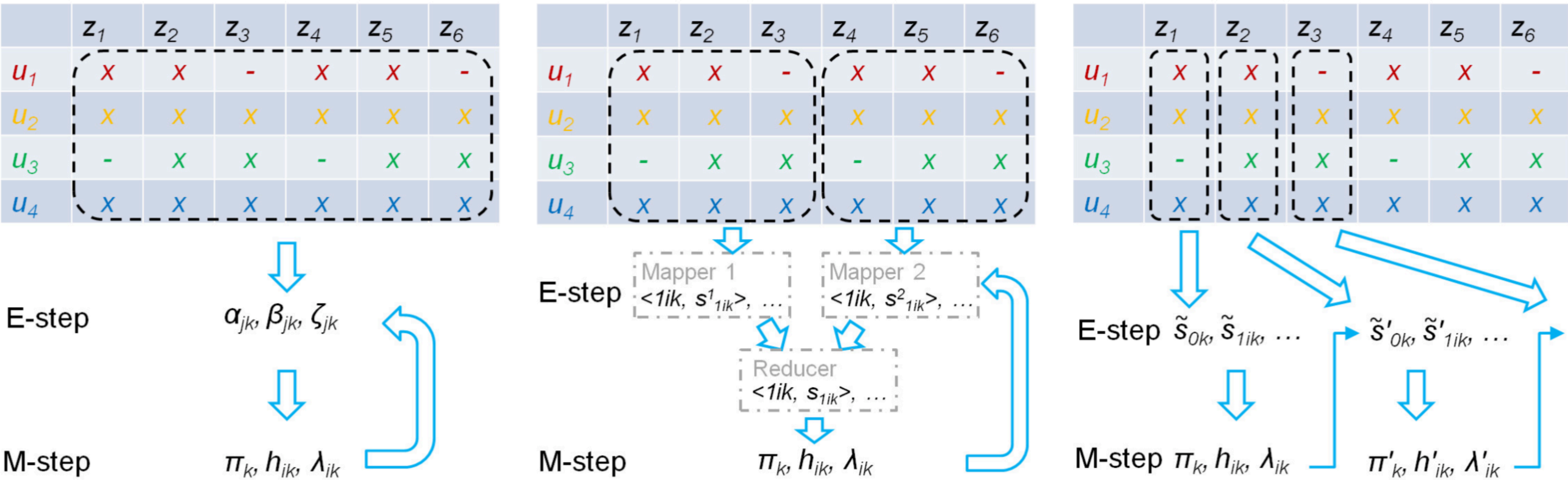
Pairwise comparisons of sources covering the same data items specialized inverted index

For EM-based approaches:

1. Each update needs all the data set out of memory problem
2. The algorithm needs to iterate over the whole dataset several times until convergence
3. Optimal hidden variables in M-step do not have a closed-form solutions and joint optimization is required

R. Wentao Ouyang, L. M. Kaplan, A.Toniolo, M. Srivastava, T. J. Norman, Parallel and Streaming Truth Discovery in Large-Scale Quantitative Crowdsourcing. *IEEE Transactions on Parallel & Distributed Systems*, doi:10.1109/TPDS. 2016.2515092

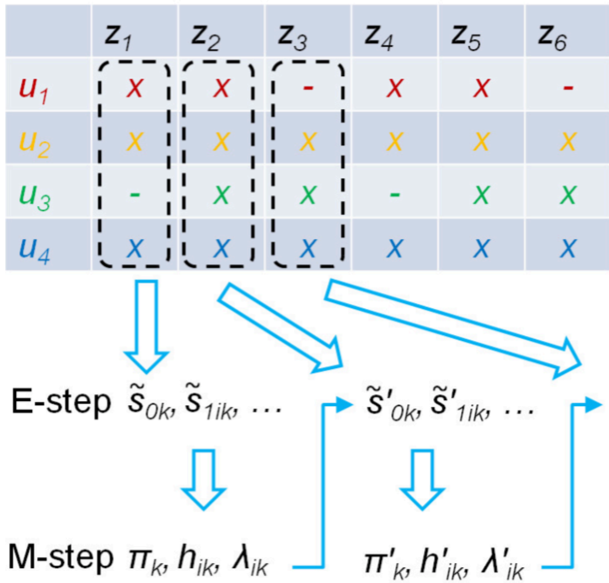
Reducing/Distributing Computation



(a) Batch truth discovery (b) Parallel truth discovery (c) Streaming truth discovery

R. Wentao Ouyang, L. M. Kaplan, A.Toniolo, M. Srivastava, T. J. Norman, Parallel and Streaming Truth Discovery in Large-Scale Quantitative Crowdsourcing. *IEEE Transactions on Parallel & Distributed Systems*, doi:10.1109/TPDS.2016.2515092

Streaming Truth Discovery



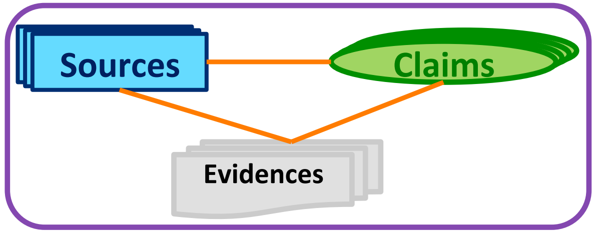
(c) Streaming truth discovery

R. Wentao Ouyang, L. M. Kaplan, A.Toniolo, M. Srivastava, T. J. Norman, Parallel and Streaming Truth Discovery in Large-Scale Quantitative Crowdsourcing. *IEEE Transactions on Parallel & Distributed Systems*, doi:10.1109/TPDS.2016.2515092

Outline

1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
4. Opportunities for scalability improvement
5. Conclusions

Truth Discovery Challenges

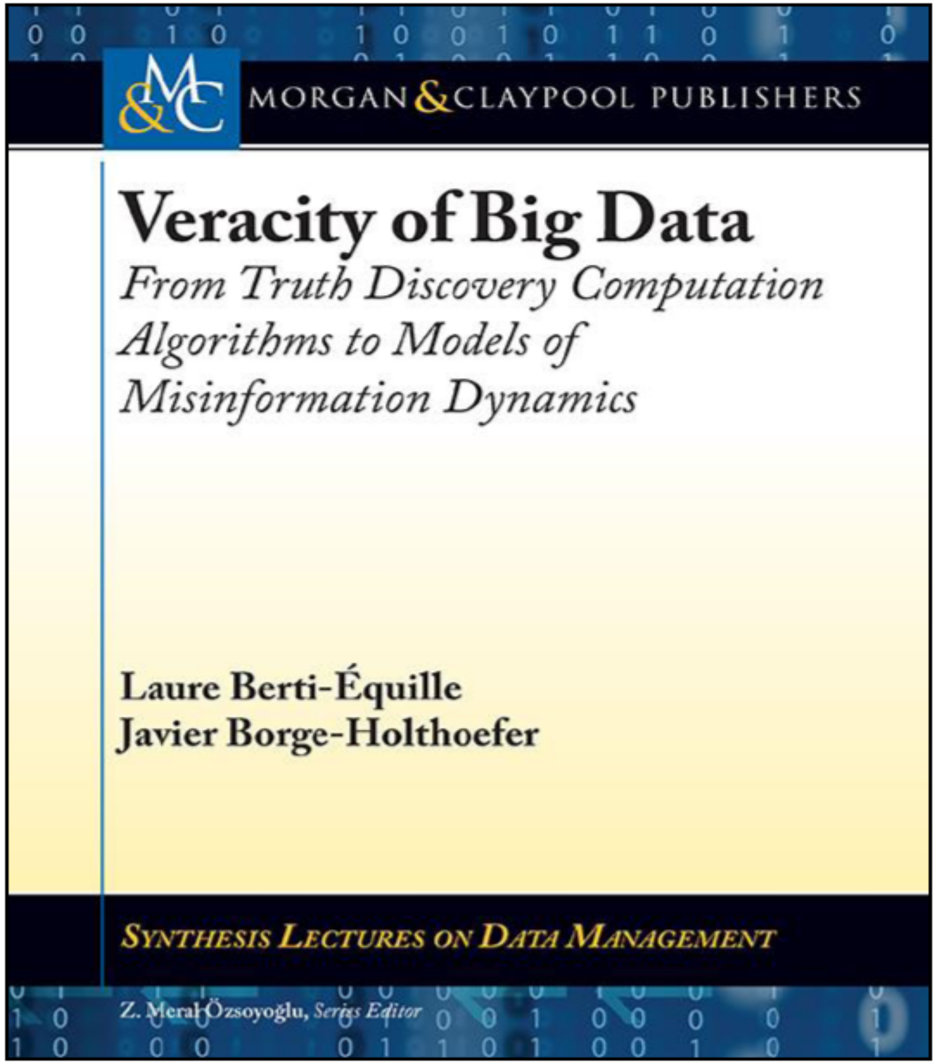


- **Data Veracity is Multidimensional**
  - Source: Coverage, Accuracy, Exacteness, Freshness, Reputation, Dependence…
  - Claims: Popularity (i.e., supported by many or few sources) (long-tail phenomena)
  - Truth: Trivial truths (hardeness), sensitive truths, uncertain, rapidly evolving
  - Data items: Information entropy (many (or few) conflicting information)
- **Truth Discovery Modeling**
  - Voting only works with benevolent sources. What about adversarial/pessimistic scenarios?
  - Need to incorporate evidences and contextual metadata (hidden agenda of sources)
  - Need to address truth discovery in the context of source/content networks
- **Algorithmic Framework**
  - Bane complex parameter setting
  - Quality performance: Ground truth data set size should be statistically significant
  - No “one-size fits all” solution
  - Need for benchmarks
- **Build a complete Truth Discovery pipeline/system**

Summary

- We presented an overview of the techniques proposed for truth discovery with opportunities for scalability and optimization improvement
- Many scientific and technical obstacles:
  - Relax modeling assumptions
  - Solve algorithmic issues related to scalability and complex parameter settings, e.g., Web-scale fact extraction/checking
  - Integrate theoretical and applied work from complex networked systems to better capture the multi-layered dynamics of misinformation
- Still a lot needs to be done for automating truth discovery for realistic and actionable scenarios
- Next step: cross-modal truth discovery

Further Reading



Veracity of Big Data (Morgan & Claypool)

Surveys

- M. Gupta and J. Han. Heterogeneous network-based trust analysis: A survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.
- K. Thirunarayan, P. Anantharam, C. A. Henson, and A. P. Sheth. Comparative trust management with applications: Bayesian approaches emphasis. Future Generation Comp. Syst., 31:182–199, 2014.

Tutorials

- Jing Gao, Qi Li, Bo Zhao, Wei Fan, Jiawei Han Truth Discovery and Crowdsourcing Aggregation: A Unified Perspective. In VLDB 2015
- Xin Luna Dong and Divesh Srivastava. Big Data Integration. In VLDB 2013
- Barna Saha and Divesh Srivastava. Data Quality: the Other Face of Big Data. In VLDB 2014
- Jeffrey Pasternack, Dan Roth, V.G. Vinod Vydiswaran. Information Trustworthiness. In AAAI 2013
- Carlos Castillo, Wei Chen, Laks V. S. Lakshmanan. Information and Influence Spread in Social Networks. In KDD 2012
- Jure Leskovec. Social Media Analytics. In KDD 2011

Experimental Study

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

Thanks!

Questions?

References

Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo, Paolo Papotti, Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources. CAiSE 2010: 83-97

Dong Wang, Lance Kaplan, and Tarek Abdelzaher, "Maximum Likelihood Analysis of Conflicting Observations in Social Sensing," ACM Transactions on Sensor Networks, Vol. 10, No. 2, Article 30, January, 2014.

Dong Wang , Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu Aggrawal, Raghu Ganti, XinLei Wang, Prasant Mohapatra, Boleslaw Szymanski, Hieu Le, "Humans as Sensors: An Estimation Theoretic Perspective," The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 14), Berlin, Germany, April, 2014.

Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Credibility Tradeoffs in Assured Social Sensing." IEEE JSAC special issue on Network Science, June, Vol. 31, No. 6, 2013.

Dong Wang, Tarek Abdelzaher, Lance Kaplan and Charu C. Aggarwal. "Recursive Fact-finding: A Streaming Approach to Truth Estimation in Crowdsourcing Applications.", 33rd International Conference on Distributed Computing Systems (ICDCS 13) Philadelphia, PA, July 2013.

Dong Wang, Tarek Abdelzaher, Lance Kaplan and Raghu Ganti. "Exploitation of Physical Constraints for Reliable Social Sensing," IEEE 34th Real-Time Systems Symposium (RTSS'13) Vancouver, Canada, December, 2013.

Dong Wang, Tarek Abdelzaher, Lance Kaplan, and Charu C. Aggarwal, "On Quantifying the Accuracy of Maximum Likelihood Estimation of Participant Reliability in Social Sensing", 8th International Workshop on Data Management for Sensor Networks (DMSN11), August 2011.

Dong Wang, Tarek Abdelzaher, Hossein Ahmadi, Jeff Pasternack, Dan Roth, Manish Gupta, Jiawei Han, Omid Fatemieh, Hieu Le, and Charu Aggarwal, "On Bayesian Interpretation of Fact-finding in Information Networks". 14th International Conference on Information Fusion (Fusion 2011).

Md Tanvir Amin, Tarek Abdelzaher, Dong Wang , Boleslaw Szymanski. "Crowd-sensing with Polarized Sources," In Proc. 10th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS) , Marina Del Rey, CA, May 2014