# Truth Discovery and Veracity Analysis

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MARCH 30, 2017

# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework
- Truth Finding: Variations and Extensions
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data
- Conclusions and Future Research

# Why Truth Finding?

- We are in the Web and "Big Data" age
  - Lots of information:  Also, lots of errors and false information!
  - Lots of information providers: Not every one is 100% reliable
- When encountering conflicting information on the same entities
  - Which piece of info is correct?
  - Which sources are trustable?
- Challenges:  We want to get trusted information!
  - Training on millions of pieces of information?
    - Too expensive, unrealistic!
  - Trust on "trusted" sources?
    - Everyone can make mistakes, even for the majority

# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework 
- Truth Finding: Variations and Extensions
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data
- Conclusions and Future Research

# Truth Validation by Info Network Analysis

- The trustworthiness problem of the web (according to a survey):
  - 54% of Internet users trust news web sites most of time
  - 26% for web sites that sell products
  - 12% for blogs
- TruthFinder: Truth discovery on the Web by link analysis
  - Among multiple conflict results, can we automatically identify which one is likely the true fact?
- Veracity (conformity to truth):
  - Given a large amount of conflicting information about many objects, provided by multiple web sites (or other information providers), how to discover the true fact about each object?
- Our first work: Xiaoxin Yin, Jiawei Han, Philip S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web", TKDE'08

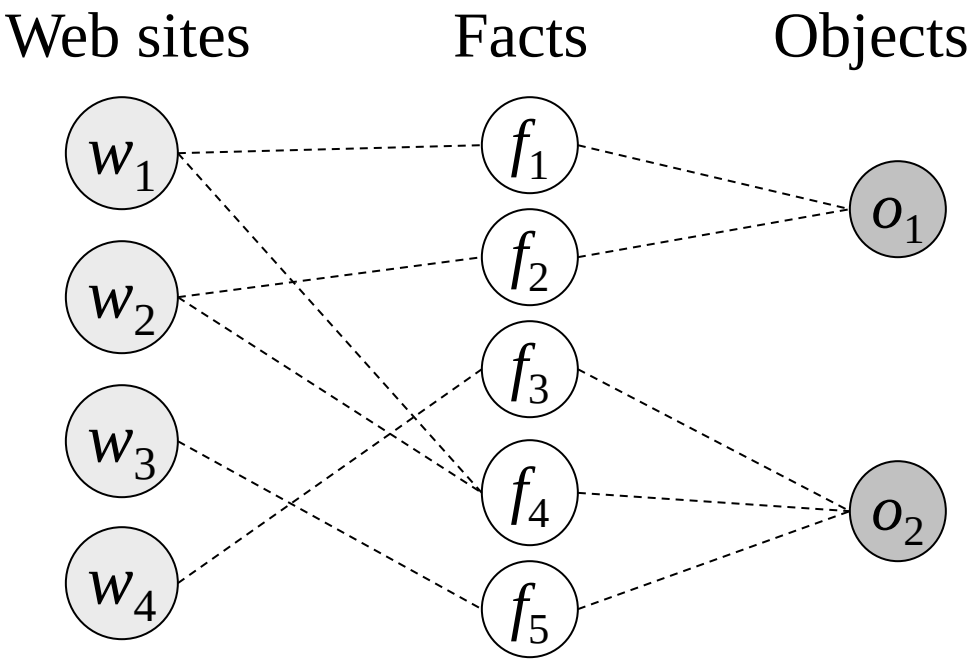
# Conflicting Information on the Web

- Different websites often provide conflicting info. on a subject, e.g., Authors of *"Rapid Contextual Design"*

| ***Online Store*** | ***Authors*** |
|---|---|
| Powell's books | Holtzblatt, Karen |
| Barnes & Noble | Karen Holtzblatt, Jessamyn Wendell, Shelley Wood |
| A1 Books | Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood |
| Cornwall books | Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood |
| Mellon's books | Wendell, Jessamyn |
| Lakeside books | WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY |
| Blackwell online | Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley |

# Our Setting: Info. Network Analysis

- Each object has a set of ***conflictive*** facts
  - E.g., different author names for a book
- And each web site provides some facts
- How to find the true fact for each object?



# Basic Heuristics for Problem Solving

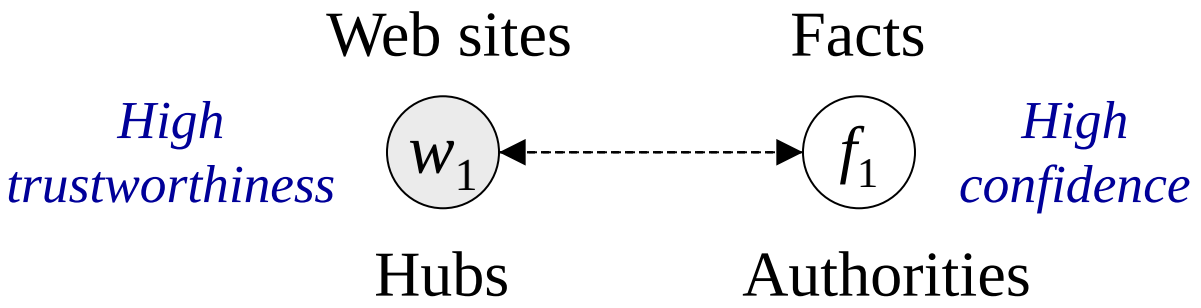
1. There is usually only one true fact for a property of an object
2. This true fact appears to be the same or similar on different web sites
   - E.g., “Jennifer Widom” vs. “J. Widom”
3. **The false facts on different web sites are less likely to be the same or similar**
   - False facts are often introduced by random factors
4. **A web site that provides mostly true facts for many objects will likely provide true facts for other objects**

# Overview of the TruthFinder Method

- *Confidence of facts* ↔ *Trustworthiness of web sites*
  - A fact has *high confidence* if it is provided by (many) trustworthy web sites
  - A web site is *trustworthy* if it provides many facts with high confidence
- The TruthFinder mechanism, an overview:
  - Initially, each web site is equally trustworthy
  - Based on the above four heuristics, infer fact confidence from web site trustworthiness, and then backwards
  - Repeat until achieving stable state

# Analogy to Authority-Hub Analysis

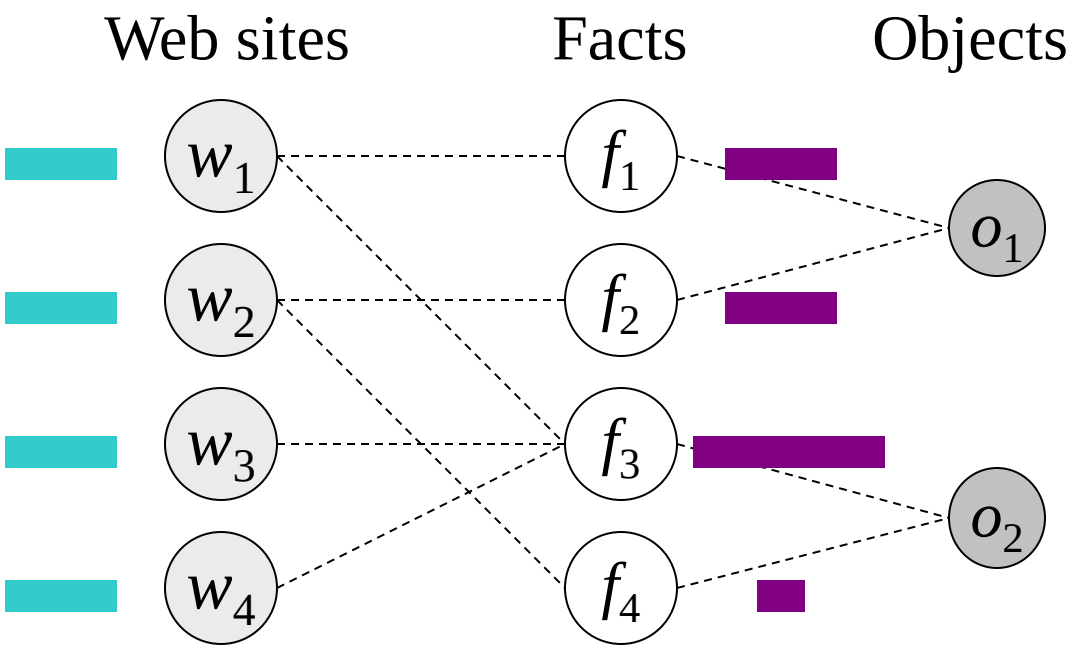
- Facts ↔ Authorities, Web sites ↔ Hubs




- Difference from authority-hub analysis
  - Linear summation cannot be used
    - A web site is trustable if it provides accurate facts, instead of many facts
    - Confidence is the probability of being true
  - Different facts of the same object influence each other

# Inference on Trustworthness

- Inference of web site trustworthiness & fact confidence




- True facts and trustable web sites will become apparent after some iterations

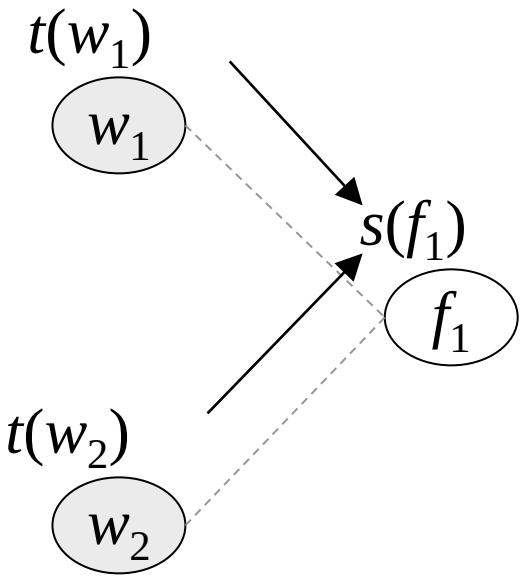
# Computation Model: t(w) and s(f)

- **The trustworthiness of a web site *w*: *t*(*w*)**
  - Average confidence of facts it provides

$$t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|}$$

*Sum of fact confidence*

*Set of facts provided by w*




- **The confidence of a fact *f*: *s*(*f*)**
  - One minus the probability that all web sites providing *f* are wrong

$$s(f) = 1 - \prod_{w \in W(f)} (1 - t(w))$$

*Probability that w is wrong*

*Set of websites providing f*

# Experiments: Finding Truth of Facts

- Determining authors of books
  - Dataset contains 1265 books listed on abebooks.com
  - We analyze 100 random books (using book images)

| Case | *Voting* | *TruthFinder* | *Barnes & Noble* |
|---|---|---|---|
| Correct | 71 | 85 | 64 |
| Miss author(s) | 12 | 2 | 4 |
| Incomplete names | 18 | 5 | 6 |
| Wrong first/middle names | 1 | 1 | 3 |
| Has redundant names | 0 | 2 | 23 |
| Add incorrect names | 1 | 5 | 5 |
| No information | 0 | 0 | 2 |

# Experiments: Trustable Info Providers

- Finding trustworthy information sources
  - Most trustworthy bookstores found by TruthFinder vs. top-ranked bookstores by Google (query “bookstore”)

**TruthFinder**

| Bookstore | *trustworthiness* | *#book* | *Accuracy* |
|---|---|---|---|
| TheSaintBookstore | 0.971 | 28 | 0.959 |
| MildredsBooks | 0.969 | 10 | 1.0 |
| Alphacraze.com | 0.968 | 13 | 0.947 |

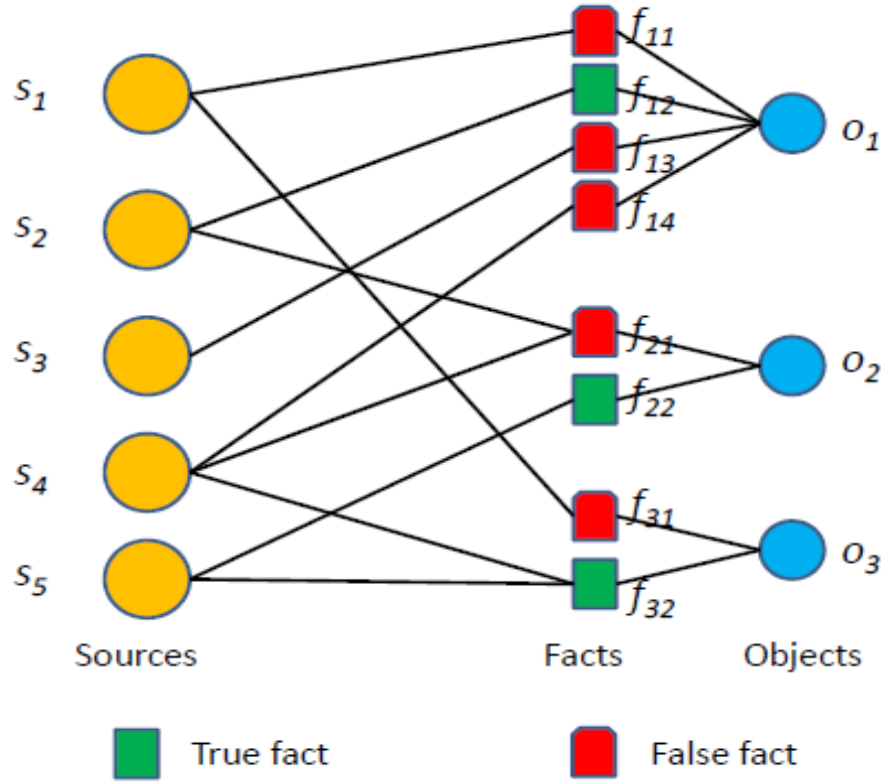
**Google**

| Bookstore | *Google rank* | *#book* | *Accuracy* |
|---|---|---|---|
| Barnes & Noble | 1 | 97 | 0.865 |
| Powell’s books | 3 | 42 | 0.654 |

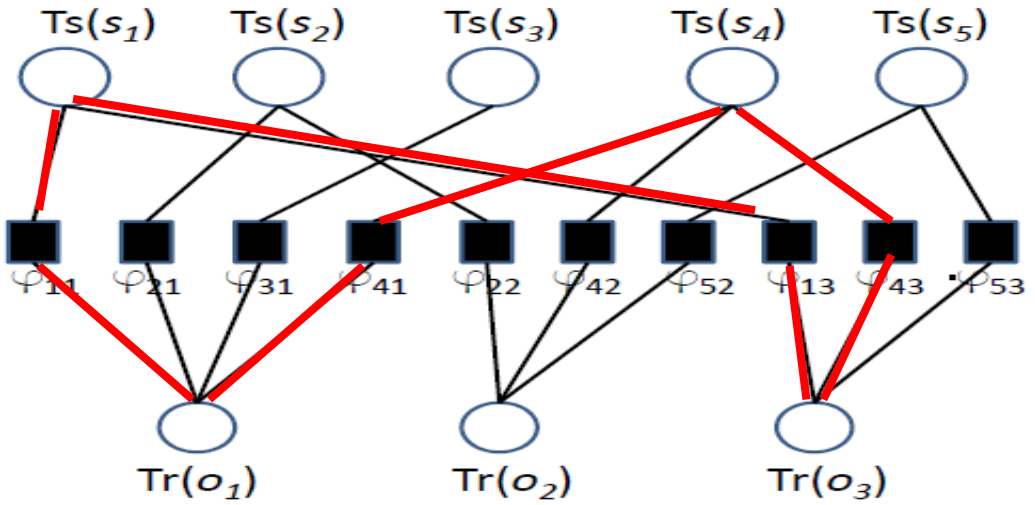
# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework
- Truth Finding: Variations and Extensions 
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data
- Conclusions and Future Research

# Mining Collective Intelligence in Groups: Latent Dirichlet Truth Discovery



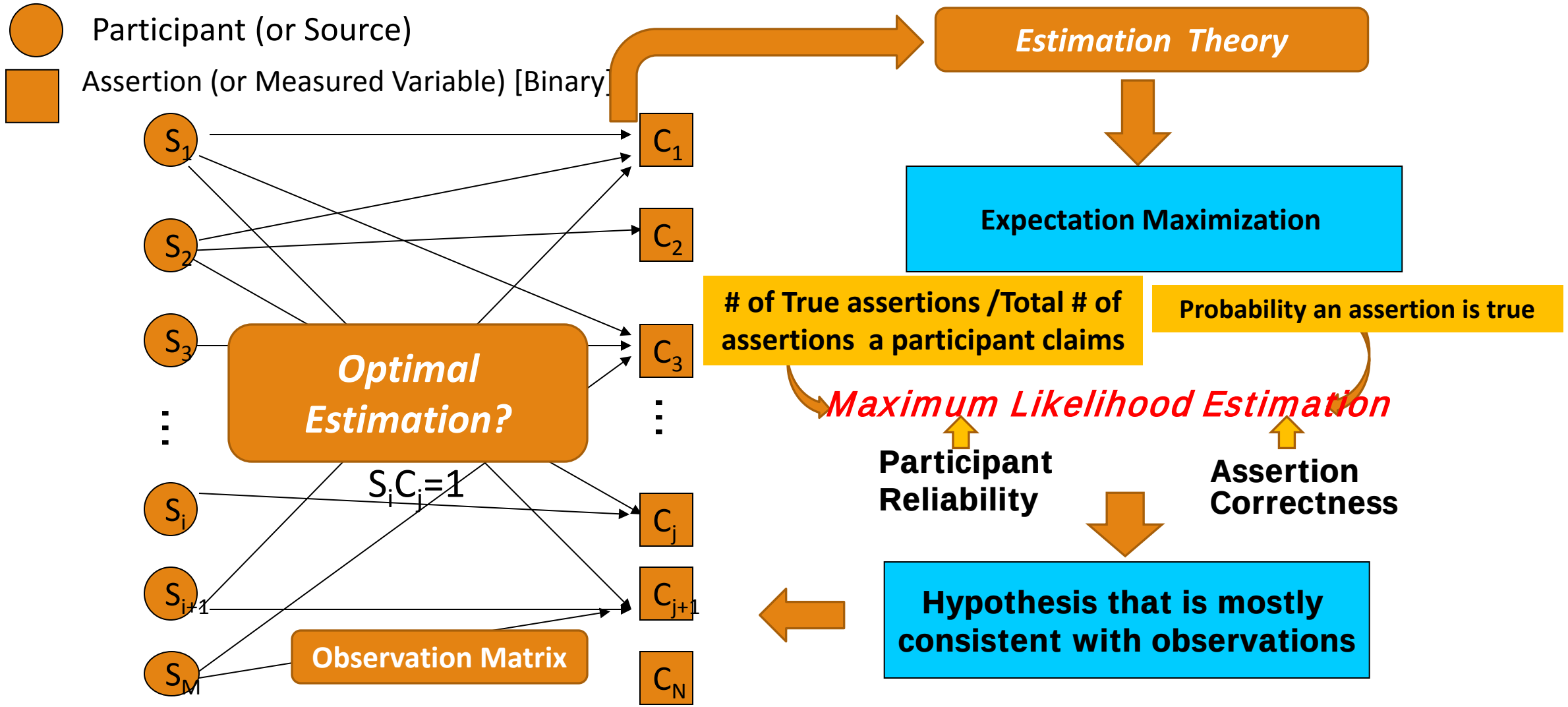

**A Source-Claim-Object Framework**




- Explicitly exploit the fact-object hierarchy
- Considers all the facts as a whole for each object
- Model the behavior of both trustworthy and untrustworthy sources explicitly

Guo-Jun Qi, Charu C. Aggarwal, Jiawei Han, and Thomas Huang, "Mining Collective Intelligence in Groups", WWW'13, May 2013

# An Expectation Maximization Framework




D. Wang, L. Kaplan and T. Abdelzaher, "On Truth Discovery in Social Sensing with Conflicting Observations: A Maximum Likelihood Estimation Approach", ACM Transaction on Sensor Networks (TOSN), 2014

# Generalized Fact-Finding: Considering Additional Information

- Ex: *"Obama was born in Hawaii" vs. "Obama was born in Kenya"*
  - **Uncertainty in the claims:** "I'm 90% sure Obama was born in Hawaii"
  - **Attributes of sources:** The source making the first claim is Authority
  - **Similarity between claims:** A source claiming "Kenya" implicitly prefers neighboring "Uganda" over "Hawaii"
- Key Features: J. Pasternack and D. Roth, "Making Better Informed Trust Decisions with Generalized Fact-Finding", IJCAI'11
  - Consider **Additional/Background Knowledge** in Fact-finding (e.g., source uncertainty, claim similarity, group information)
  - Model Additional Knowledge **as Link Weights** to Generalize Fact-finding algorithms
  - Generalize bi-partite graph to **k-partite graph** to consider source groups or attributes

# Truth Discovery and Copying Detection in a Dynamic World

- Luna Dong's series work on truth discovery and data integration
- Find true values and determine the copying relationship between sources (VLDB'09)
- Quality of sources over time: coverage, exactness and freshness
- Use hidden Markov model (HMM) to decide whether a source is a copier of another and identifies the specific moments it copies
- Use a Bayesian model that aggregates info from sources to decide the true value for a data items and the evolution of the true value over time
- Further study on truth finding on the Web (e.g., VLDB'13)
  - Lot of inconsistencies on even deep web, in some highly "trusted" domains, e.g., stock and flight
  - For 70% of data items, > 1 value is provided: Widely open!

# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework
- Truth Finding: Variations and Extensions
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data
- Conclusions and Future Research

# From TruthFinder to Latent Truth Model (LTM)

- HITS-like Random Walk methods (e.g., TruthFinder(KDD'08), 3-Estimate(WSDM'10), Invest(COLING'10), …)
  - Higher Quality Sources ↔ More Probable Facts
- Limitations
  - Quality as a single value: Precision or Accuracy
  - In practice, some sources tend to ignore true attributes (False Negatives), while some others tend to produce false attributes (False Positives).
  - FN ǂ FP when there are multiple truths per entity!
- LTM (Latent Truth Model): Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han, "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration", VLDB'12
- LTM: A Principled Probabilistic Model
- Model negative claims and two-sided source quality with Bayesian regularization

# What Is Latent Truth Model?

- Different but real situations (new assumptions)
  - Multiple facts can be true for each entity (object)
    - One book may have 2+ authors
  - A source can make multiple claims per entity, where more than one of them can be true
    - A source may claim a book w. 3 authors
  - Sources and objects are independent respectively
    - Assume book websites and books are independent
  - The majority of data coming from many sources are not erroneous
    - Trust the majority of the claims

**Table 1: An example raw database of movies.**

| *Entity (Movie)* | *Attribute (Cast)* | *Source* |
|---|---|---|
| Harry Potter | Daniel Radcliffe | IMDB |
| Harry Potter | Emma Waston | IMDB |
| Harry Potter | Rupert Grint | IMDB |
| Harry Potter | Daniel Radcliffe | Netflix |
| Harry Potter | Daniel Radcliffe | BadSource.com |
| Harry Potter | Emma Waston | BadSource.com |
| Harry Potter | Johnny Depp | BadSource.com |
| Pirates 4 | Johnny Depp | Hulu.com |
| ... | ... | ... |

# Challenge: Why Voting Does Not Work?

- Input:
  - Facts = (Entity, Attribute)
  - Claims = (Relation, Source, Observation)
- Output: Truth(Relation) → {1, 0}

- Many sources are not of high quality: Cannot trust all sources in voting
- Movie: Harry Potter and the Deathly Hallows
  - IMDB: Daniel Radcliffe, Emma Watson
  - Nexflix: Daniel Radcliffe
  - BadSource.Com: Daniel Radcliffe, Johnny Depp

| RID | Entity | Attribute |
|---|---|---|
| 1 | HP7 | Daniel Radcliffe |
| 2 | HP7 | Emma Watson |
| 3 | HP7 | Johnny Depp |

Input: Fact Table

| RID | Source | Obs. |
|---|---|---|
| 1 | IMDB | 1 |
| 2 | IMDB | 1 |
| 3 | IMDB | 0 |
| 1 | Netflix | 1 |
| 2 | Netflix | 0 |
| 3 | Netflix | 0 |
| 1 | BadSource | 1 |
| 2 | BadSource | 0 |

Input: Claim Table

False Negative (threshold = 0.5)

False Positive (threshold = 0.3)

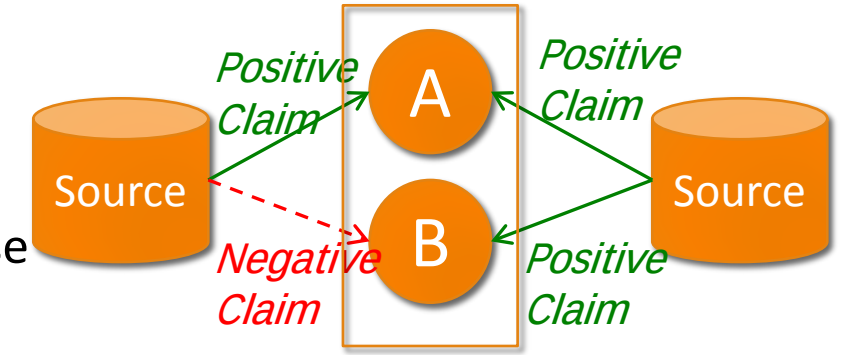
Optimal Threshold?

| RID | Truth |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 0 |

Output: Truth Table

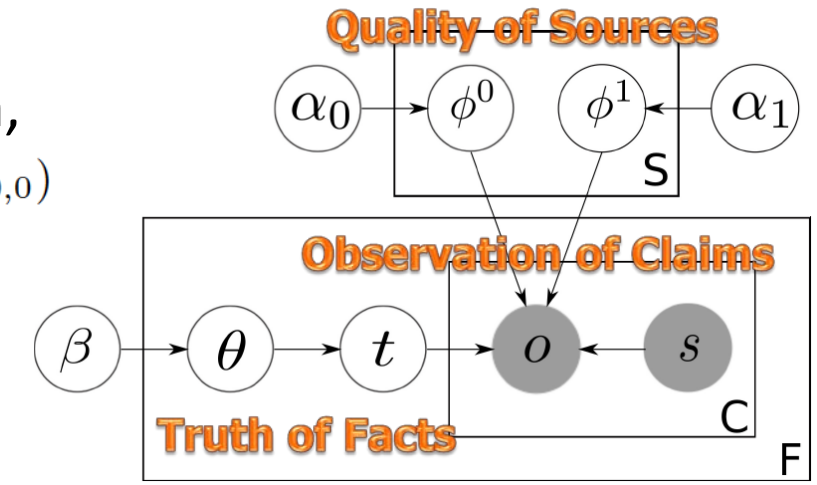
# Multiple Truths for Same Entities

- Implicit negative claims by source s:
  - For those facts not claimed true by source s, but by some other sources
- Modeling negative claims and two-sided errors (false positives, false negatives) is essential for supporting multiple truths:
  - Negative claims can help detect false attributes
  - Negative claims by high recall sources are usually very accurate, e.g., IMDB
  - Negative claims by low recall sources should not count much, e.g., Netflix
- Thus, LTM Naturally supports multiple true attribute values
- LTM can naturally incorporate prior domain knowledge through Bayesian priors
- An efficient and scalable linear complexity inference algorithm
  - LTM can run in either batch or online streaming modes for incremental truth finding



# The Latent Truth Model

- For each source **k**
  - Generate false positive rate (with **strong** regularization, believing most sources have low FPR): $\phi_k^0 \sim Beta(\alpha_{0,1}, \alpha_{0,0})$
  - Generate its sensitivity (1-FNR) with uniform prior, indicating low FNR is more likely: $\phi_k^1 \sim Beta(\alpha_{1,1}, \alpha_{1,0})$
- For each fact **f**
  - Generate its prior truth prob, uniform prior: $\theta_f \sim Beta(\beta_1, \beta_0)$
  - Generate its truth label: $t_f \sim Bernoulli(\theta_f)$
- For each claim **c** of fact **f,** generate observation of **c**.
  - If **f** is false, use false positive rate of source: $o_c \sim Bernoulli(\phi_{s_c}^0)$
  - If **f** is true, use sensitivity of source: $o_c \sim Bernoulli(\phi_{s_c}^1)$



Graphical Representation

# Inferring Truth

- MAP inference: find truth assignment that maximizes posterior probabilities

$$\hat{\boldsymbol{t}}_{MAP} = \arg\max_{\boldsymbol{t}} \int\int\int p(\boldsymbol{o}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{\phi}^{\mathbf{0}}, \boldsymbol{\phi}^{\mathbf{1}}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\phi}^{\mathbf{0}} \mathrm{d}\boldsymbol{\phi}^{\mathbf{1}}$$

- Collapsed Gibbs sampling (more efficient, only sample truth, other parameters integrated out)

$$p(t_f = i | \boldsymbol{t}_{-f}, \boldsymbol{o}, \boldsymbol{s}) \propto \beta_i \prod_{c \in \mathbf{C}_f} \frac{n^{-f}_{s_c,i,o_c} + \alpha_{i,o_c}}{n^{-f}_{s_c,i,1} + n^{-f}_{s_c,i,0} + \alpha_{i,1} + \alpha_{i,0}}$$

- Prediction of the truth from samples of the latent truth variable.
  - Burn-in: throw away first b samples
  - Thinning: Take every k sample from all the samples
  - Calculate expectation of the taken samples

# Incremental Truth Finding

- Read-off Source Quality Parameters

$$sensitivity(s) = \phi_s^1 = \frac{E[n_{s,1,1}] + \alpha_{1,1}}{E[n_{s,1,0}] + E[n_{s,1,1}] + \alpha_{1,0} + \alpha_{1,1}}$$

$$specificity(s) = 1 - \phi_s^0 = \frac{E[n_{s,0,0}] + \alpha_{0,0}}{E[n_{s,0,0}] + E[n_{s,0,1}] + \alpha_{0,0} + \alpha_{0,1}}$$

- Incremental Prediction (LTMinc)
  - Assuming source quality unchanged, directly utilize source quality to make prediction (very efficient)

$$p(t_f = 1 | \boldsymbol{o}, \boldsymbol{s}) = \frac{\beta_1 \prod_{c \in \mathbf{C}_f} (\phi_s^1)^{o_c} (1 - \phi_s^1)^{1 - o_c}}{\sum_{i=0,1} \beta_i \prod_{c \in \mathbf{C}_f} (\phi_s^i)^{o_c} (1 - \phi_s^i)^{1 - o_c}}$$

  - Can also rerun inference for incremental update by using previous quality counts as Bayesian priors

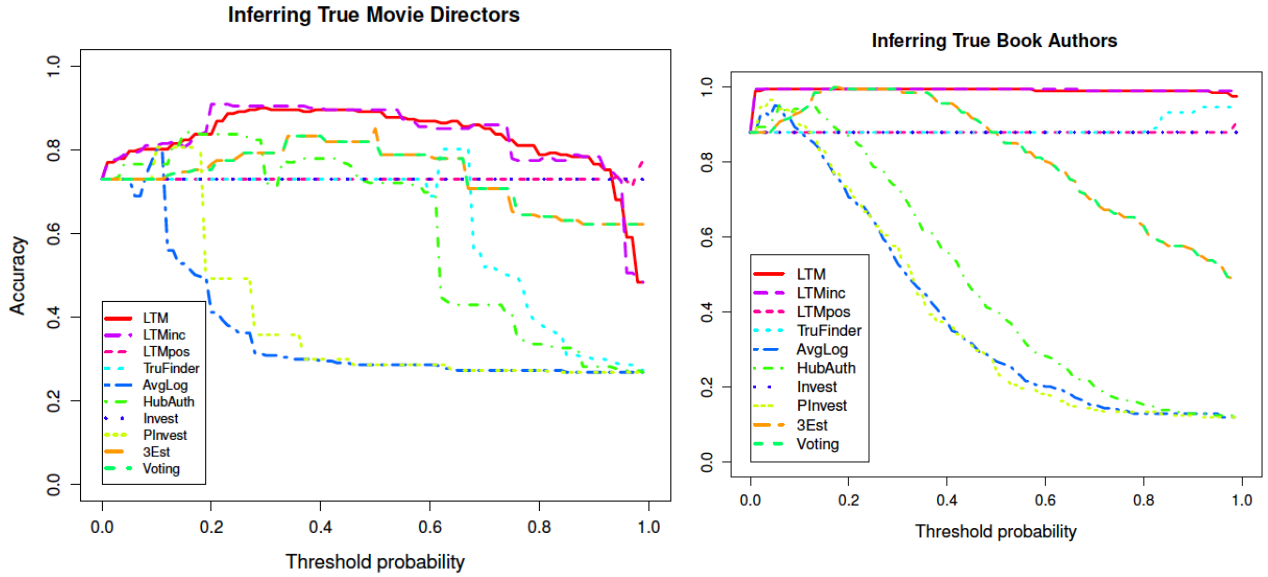
# Experiments (Effectiveness)

- Datasets:
  - Book Authors from abebooks.com (1263 books, 879 sources, 48153 claims, 2420 book-author, 100 labeled)
  - Movie Directors from Bing (15073 movies, 12 sources, 108873 claims, 33526 movie-director, 100 labeled)

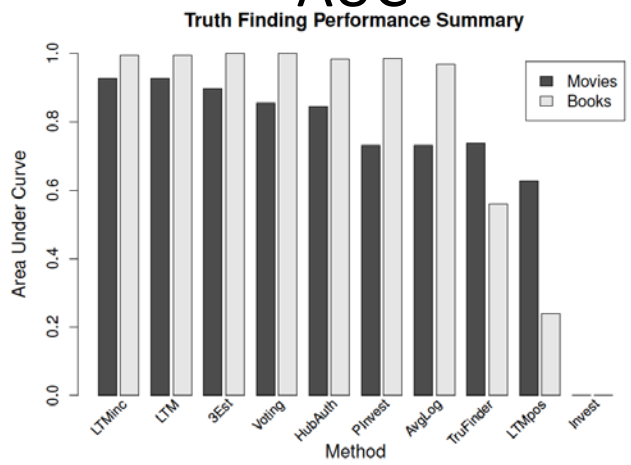
| | *Results on book data* | | | | | *Results on movie data* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *One-sided error* | | | *Two-sided error* | | *One-sided error* | | | *Two-sided error* | |
| | *Precision* | *Recall* | *FPR* | *Accuracy* | *F1* | *Precision* | *Recall* | *FPR* | *Accuracy* | *F1* |
| **LTMinc** | **1.000** | 0.995 | **0.000** | **0.995** | **0.997** | 0.943 | 0.914 | 0.150 | **0.897** | **0.928** |
| **LTM** | **1.000** | 0.995 | **0.000** | **0.995** | **0.997** | 0.943 | 0.908 | 0.150 | 0.892 | 0.925 |
| 3-Estimates | **1.000** | 0.863 | **0.000** | 0.880 | 0.927 | 0.945 | 0.847 | 0.133 | 0.852 | 0.893 |
| Voting | **1.000** | 0.863 | **0.000** | 0.880 | 0.927 | 0.855 | 0.908 | 0.417 | 0.821 | 0.881 |
| TruthFinder | 0.880 | **1.000** | 1.000 | 0.880 | 0.936 | 0.731 | **1.000** | 1.000 | 0.731 | 0.845 |
| Investment | 0.880 | **1.000** | 1.000 | 0.880 | 0.936 | 0.731 | **1.000** | 1.000 | 0.731 | 0.845 |
| HubAuthority | **1.000** | 0.322 | **0.000** | 0.404 | 0.488 | **1.000** | 0.620 | **0.000** | 0.722 | 0.765 |
| AvgLog | **1.000** | 0.169 | **0.000** | 0.270 | 0.290 | **1.000** | 0.025 | **0.000** | 0.287 | 0.048 |
| LTMpos | 0.880 | **1.000** | 1.000 | 0.880 | 0.936 | 0.731 | **1.000** | 1.000 | 0.731 | 0.845 |
| PooledInvestment | **1.000** | 0.142 | **0.000** | 0.245 | 0.249 | **1.000** | 0.025 | **0.000** | 0.287 | 0.048 |

# Experiments (Effectiveness) cont.

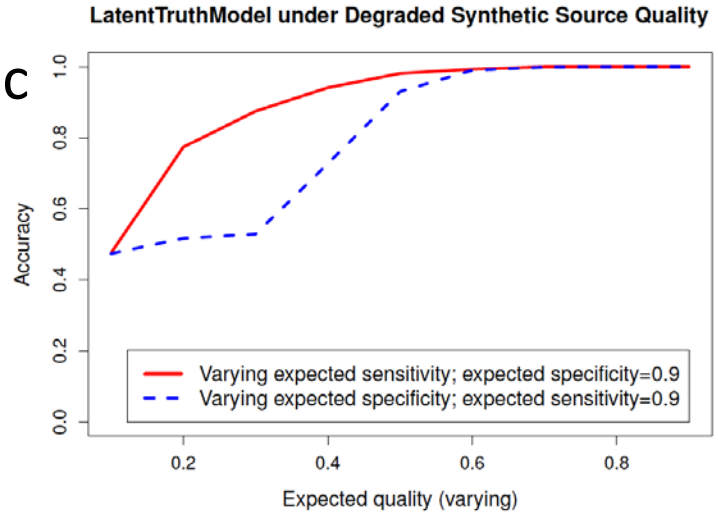
## Varying cutoff threshold (consistently better)




## AUC




## Varying synthetic quality (more tolerant of low sensitivity)

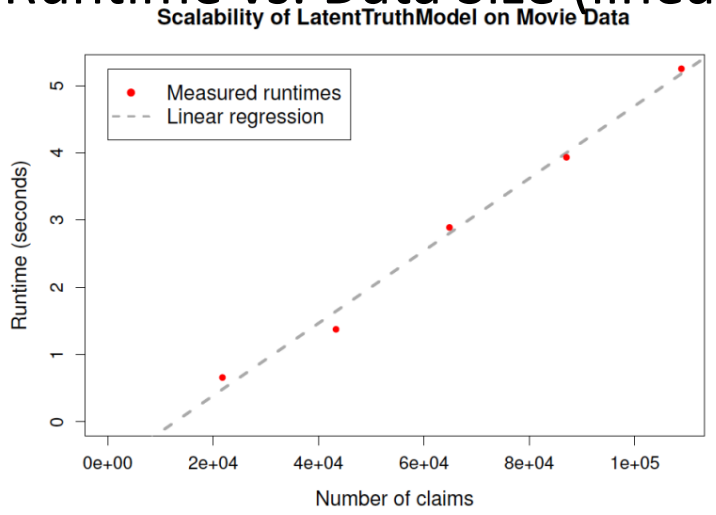



## Case Study of Movie Sources

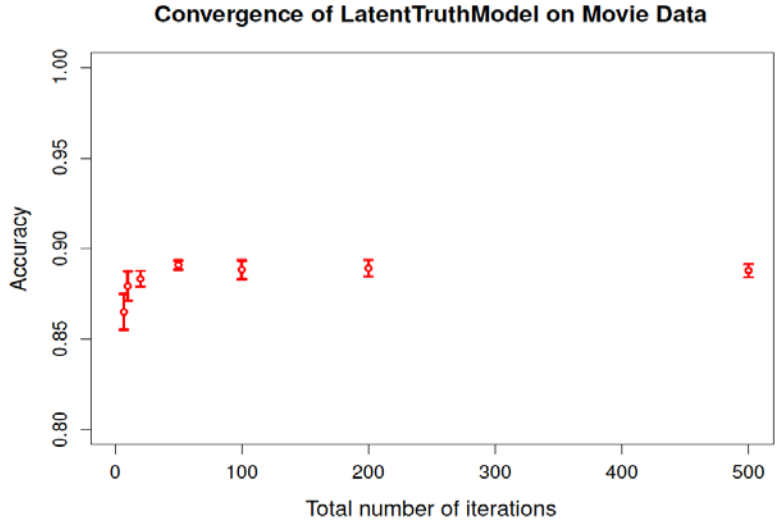
| *Source* | *Sensitivity* | *Specificity* |
|---|---|---|
| imdb | **0.911622836** | 0.898838631 |
| netflix | 0.894019034 | 0.934833904 |
| movietickets | 0.862889367 | 0.978844687 |
| commonsense | 0.809752315 | 0.982347827 |
| cinemasource | 0.794184357 | 0.985847745 |
| amg | 0.776583683 | 0.690600694 |
| yahoomovie | 0.760589896 | 0.897654374 |
| msnmovie | 0.749192861 | 0.987870636 |
| zune | 0.744272491 | 0.973922421 |
| metacritic | 0.678661638 | 0.987957893 |
| flixster | 0.584223615 | 0.911078627 |
| fandango | 0.499623726 | **0.989836274** |

# Experiments (Efficiency)

## Runtime vs. Data Size (linear)




## Convergence Rate (stable at 50)




## Running Time

| | *Runtimes (secs.) vs. #Entities* | | | | |
|---|---|---|---|---|---|
| *#Entities* | 3k | 6k | 9k | 12k | 15k |
| *1 iteration* | | | | | |
| **Voting** | **0.004** | **0.008** | **0.012** | **0.027** | **0.030** |
| LTMinc | **0.004** | **0.008** | **0.012** | 0.037 | 0.048 |
| *100 iterations* | | | | | |
| **AvgLog** | 0.150 | **0.297** | 0.446 | **0.605** | **0.742** |
| **HubAuthority** | **0.149** | **0.297** | **0.445** | 0.606 | 0.743 |
| PooledInvestment | 0.175 | 0.348 | 0.514 | 0.732 | 0.856 |
| TruthFinder | 0.195 | 0.393 | 0.587 | 0.785 | 0.971 |
| Investment | 0.231 | 0.464 | 0.690 | 0.929 | 1.143 |
| 3-Estimates | 0.421 | 0.796 | 1.170 | 1.579 | 1.958 |
| LTM | 0.660 | 1.377 | 2.891 | 3.934 | 5.251 |

*Standard LTM is slightly slower than state-of-the-art due to random sampling*

*LTMinc is as fast as voting!*

# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework
- Truth Finding: Variations and Extensions
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data 
- Conclusions and Future Research

# Estimating Real-Valued Truth from Conflicting Sources

- Truth finding from categorical data to numerical data
  - Real valued data could be critical in many applications
- B. Zhao and J. Han, "A Probabilistic Model for Estimating Real-Valued Truth from Conflicting Sources", Proc. of 2012 Int. Workshop on Quality in Databases (QDB'12), Aug. 2012.
- **GTM** (Gaussian Truth Model): A principled probabilistic model
  - Leverage two Gaussian generative processes to simulate the generation of numerical truth and claims
  - Source quality adapts to numerical data
  - Prior on truth and source quality can be easily incorporated as Bayesian priors
  - Efficient inference

# Automatic Finding of True Real Values

- Problem Formulation
  - Input: Claim Table in form of (entity, value, source).
  - Output: Truth Table (entity, true value)

| Entity (city) | Value (pop.) | Source |
|---|---|---|
| NYC | 8,346,794 | Freebase |
| NYC | 8,244,910 | Wikipedia |
| NYC | 8,175,133 | US Census |
| NYC | 7,864,215 | BadSource |
| Urbana | 36,395 | US Census |
| Urbana | 36,395 | Wikipedia |
| Urbana | 34,774 | Freebase |
| Urbana | 1,215 | BadSource |
| … | … | … |

**Input: Claim Table**

| Entity | Value |
|---|---|
| NYC | 8,175,133 |
| Urbana | 36,395 |
| … | … |

**Output: Truth Table**

# Intuition: Normalization vs. Outliers

- Quality varies for different sources
  - US Census > Wikipedia > Freebase > BadSource
- Normalization
  - NYC is more difficult to reach consensus, sources should be punished less for making same amount of error, comparing with Urbana
- Outlier
  - (Urbana, 1215) is an outlier, needs to be recognized

| Entity | Value | Source |
|---|---|---|
| NYC | 8,346,794 | Freebase |
| NYC | 8,244,910 | Wikipedia |
| NYC | 8,175,133 (truth) | US Census |
| NYC | 7,864,215 | BadSource |
| Urbana | 36,395 (truth) | US Census |
| Urbana | 36,395 (truth) | Wikipedia |
| Urbana | 34,774 | Freebase |
| Urbana | 1,215 | BadSource |
| … | … | … |

# Normalization and Outlier Detection

- Leverage truth priors, e.g., median, the most frequent value, or output of any truth-finding algorithms
- Remove outliers by absolute error, relative error, and Gaussian confidence interval (with prior truth as mean, iteratively computed variance)
  - Outliers can significantly shift empirical variance, so update variance after outliers are detected and try to detect outliers based on updated variance
- Normalize all claims to standard Gaussian N(0,1)
  - Prevent biased estimation of source quality

{Outlier Detection}
**for all** $e \in \mathcal{E}$ **do**
  {based on relative error and absolute error}
  **for all** $c \in \mathcal{C}_e$ **do**
    **if** $|v_c - \hat{t}_e|/\hat{t}_e > \delta_0$ or $|v_c - \hat{t}_e| > \delta_1$ **then**
      $outlier[c] \leftarrow True$
  {based on Gaussian confidence intervals}
  $\hat{\sigma}_e \leftarrow calcuate\_standard\_deviation(\mathcal{C}_e)$
  **repeat**
    $new\_outlier \leftarrow False$
    **for all** $c \in \mathcal{C}_e$ **do**
      **if** $|v_c - \hat{t}_e|/\hat{\sigma}_e > \delta_2$ **then**
        $outlier[c] \leftarrow True$
        $new\_outlier \leftarrow True$
    $\hat{\sigma}_e \leftarrow calcuate\_standard\_deviation(\mathcal{C}_e)$
  **until** $new\_outlier = False$
{Normalization}
**for all** $e \in \mathcal{E}$ **do**
  **for all** $c \in \mathcal{C}_e$ **do**
    $o_c \leftarrow (v_c - \hat{t}_e)/\hat{\sigma}_e$

**Preprocessing Routine**

# GTM: Truth Generation Mechanism

- For each source **k**
  - Generate source quality:

$$\sigma_s^2 \sim Inv\text{–}Gamma(\alpha, \beta)$$

$$\sim (\sigma_s^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma_s^2}\right)$$

- For each entity **e**
  - Generate its true value:

(set $\mu_0 = 0$ and $\sigma_0^2 = 1$)

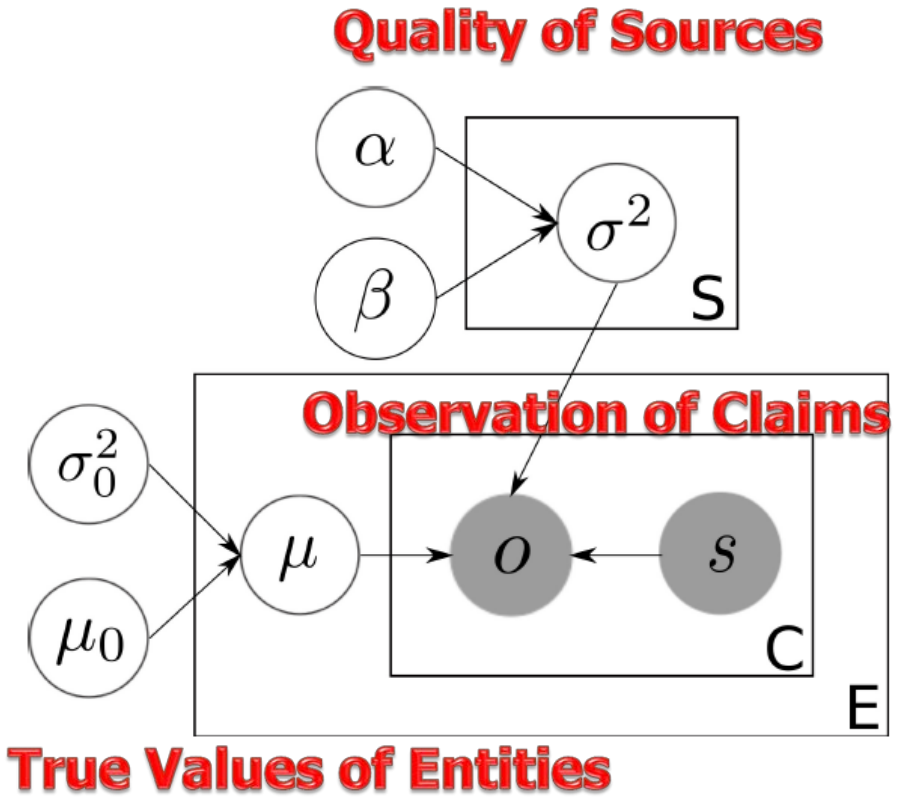
$$\mu_e \sim Gaussian(\mu_0, \sigma_0^2)$$

$$\sim \exp\left(-\frac{(\mu_e - \mu_0)^2}{2\sigma_0^2}\right)$$

- For each claim **c** of entity **e**
  - Generate observation of **c**:

$$o_c \sim Gaussian(\mu_e, \sigma_{s_c}^2)$$

$$\sim \sigma_{s_c}^{-1} \exp\left(-\frac{(o_c - \mu_e)^2}{2\sigma_{s_c}^2}\right)$$



# Inference

- Likelihood:

$$p(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mu_0, \sigma_0^2, \alpha, \beta) =$$

$$\prod_{s \in \mathcal{S}} p(\sigma_s^2 | \alpha, \beta) \times \prod_{e \in \mathcal{E}} \left( p(\mu_e | \mu_0, \sigma_0^2) \prod_{c \in \mathcal{C}_e} p(o_c | \mu_e, \sigma_e^2) \right)$$

- MAP Inference of truth:

$$\hat{\boldsymbol{\mu}}_{MAP} = \arg\max_{\boldsymbol{\mu}} \int p(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \mu_0, \sigma_0^2, \alpha, \beta) \mathrm{d}\boldsymbol{\sigma}^2$$

- Many inference algorithms can be applied, e.g. Gibbs sampling, EM, etc.

- To get actual truth: $\hat{t} + \hat{\mu}_e \hat{\sigma}_e$ , or the closest claimed value

# EM Algorithm

- Given source quality, optimal truth is:

$$\hat{\mu}_e = \frac{\frac{\mu_0}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{1}{\sigma_{s_c}^2} \frac{v_c - \hat{t}_e}{\hat{\sigma}_e}}{\frac{1}{\sigma_0^2} + \sum_{c \in \mathcal{C}_e} \frac{1}{\sigma_{s_c}^2}}$$

Regularization

Weighted by source quality

- Given truth, optimal source quality is:

$$\hat{\sigma}_s^2 = \frac{2\beta + \sum_{c \in \mathcal{C}_s} \frac{(v_c - \hat{t}_{e_c} - \hat{\sigma}_{e_c} \mu_{e_c})^2}{\hat{\sigma}_{e_c}^2}}{2(\alpha + 1) + |C_s|}$$

Regularization / smoothing

How close claims are to the truth

# Experiments on Edit History of Wikipedia

- Datasets: Edit history of Wikipedia
  - Population: 2,415 sources; 1,148 city-year; 4,119 claims.
  - Biography: 607,819 sources; 9,924 dates of birth or death; 1,372,066 claims.
- Evaluation: Mean Absolute Error, Root Mean Square Error

| | *Results on the population data* | | *Results on the bio data* | |
|---|---|---|---|---|
| | *MAE* | *RMSE* | *MAE* | *RMSE* |
| **GTM** | **1498.59** | **8339.99** | **228.19** | **4831.53** |
| 3-Estimates | 1640.83 (+9.49%) | 8822.50 (+5.78%) | 237.35 (+4.01%) | 4847.80 (+0.33%) |
| TruthFinder | 1633.60 (+9.00%) | 8824.09 (+5.80%) | 660.66 (+189.51%) | 6959.72 (+44.04%) |
| Investment | 1787.65 (+19.28%) | 9358.80 (+12.21%) | 3858.90 (+1591.04%) | 26237.65 (+443.05%) |
| LTM | 3040.90 (+102.91%) | 12865.52 (+54.26%) | 396.78 (+73.87%) | 5837.66 (+20.82%) |
| Voting | 10327.20 (+589.12%) | 126217.98 (+1413.40%) | 237.35 (+4.01%) | 4847.80 (+0.33%) |
| Median | 10241.81 (+583.42%) | 126198.86 (+1413.17%) | 244.04 (+6.94%) | 4854.90(+0.48%) |
| Average | 10368.54 (+591.88%) | 126199.76 (+1413.18%) | 253.41 (+11.05%) | 4860.28 (+0.59%) |

# Outline

- Motivation: Why Truth Finding?
- TruthFinder:  A Source-Claim-Object Network Framework
- Truth Finding: Variations and Extensions
- LTM: Latent Truth Model (Modeling Multi-Valued Truth and Two-sided Errors)
- GTM: A Gaussian Truth Model for Finding Truth among Numerical Data
- Conclusions and Future Research

# Conclusions

- Truth finding: A critical issue in data cleaning, information integration, and quality of information → Information Trust
- A fundament framework in Truth Finding:
  - A Source-Claim-Object network + iterative enhancement
- LTM (Latent Truth Model): Modeling multi-valued truth and two-sided errors
  - Cares subtlety and demonstrates its power on truth modeling
- GTM: A Gaussian truth model for numerical data
- Integration of methodologies of truth-finding and crowdsourcing
- Still a widely open area: Lots more to be studied!!

# Selected References on Truth Analysis


- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, “Integrating conflicting data: The role of source dependence”, PVLDB, 2(1):550–561, 2009.
- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, “Truth discovery and copying detection in a dynamic world”, PVLDB, 2(1):562–573, 2009.
- Manish Gupta, Yizhou Sun, and Jiawei Han, “Trust Analysis with Clustering", WWW'11
- Guo-Jun Qi, Charu C. Aggarwal, Jiawei Han, and Thomas Huang, "Mining Collective Intelligence in Groups", Proc. of 2013 Int. Conf. on Word Wide Web (WWW'13), Rio de Janeiro, Brazil, May 2013.
- Dong Wang, Lance Kaplan, and Tarek Abdelzaher, “On Truth Discovery in Social Sensing with Conflicting Observations: A Maximum Likelihood Estimation Approach”, ACM Transaction on Sensor Networks (TOSN)
- Xiaoxin Yin, Jiawei Han, and Philip S. Yu, “Truth Discovery with Multiple Conflicting Information Providers on the Web”, IEEE TKDE, 20(6):796-808, 2008
- Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han, "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration", PVLDB, 2012
- Bo Zhao and Jiawei Han, "A Probabilistic Model for Estimating Real-Valued Truth from Conflicting Sources", Proc. of 10th Int. Workshop on Quality in Databases (QDB'12), Istanbul, Turkey, Aug. 2012.