Towards Veracity Challenge in Big Data

Jing Gao[1], Qi Li[1], Bo Zhao[2], Wei Fan[3], and Jiawei Han[4]

[1]SUNY Buffalo; [2]LinkedIn;

[3]Baidu Research Big Data Lab; [4]University of Illinois

Big data challenge

- **Volume**
  - The quantity of generated and stored data

Big data challenge

- **Velocity**
  - The speed at which the data is generated and processed

Big data challenge

- **Variety**
  - The type and nature of the data

Big data challenge

- **Veracity**
  - The quality of captured data

Causes of Veracity Issue

- Rumors
- Spammers
- Collection errors
- Entry errors
- System errors
- …

Aspects of Solving Veracity Problems

- **Sources and claims**
  - We know who claims what
  - Truth discovery
- **Features of sources and claims**
  - Features of sources, eg. history, graphs of sources
  - Features of claims, eg. hashtags, lexical patterns
  - Rumor detection
  - Source trustworthiness analysis

Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References

Overview

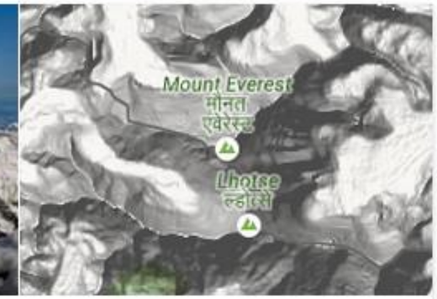
1. Introduction
2. **Truth Discovery: Veracity Analysis from Sources and Claims**
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References

Truth Discovery

- **Problem**
  - Input: Multiple conflicting information about the same set of objects provided by various information sources
  - Goal: Discover trustworthy information (i.e., the **truths**) from conflicting data on the same object

Example 1: Knowledge Base Construction

- **Knowledge base**
  - Construct knowledge base based on huge amount of information on Internet
- **Problem**
  - Find true facts from multiple conflicting sources

Mount Everest

Mountain in Asia

Mount Everest, also known in Nepal as Sagarmāthā and in Tibet as Chomolungma, is Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. Its peak is 8,848 metres above sea level. Wikipedia

Elevation: 29,029'

First ascent: May 29, 1953

Prominence: 29,029'

First ascenders: Tenzing Norgay, Edmund Hillary

Mountain range: Himalayas, Mahalangur Himal


What Is The Height Of Mount Everest?

Google

what is the height of mount everest

Mount Everest - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Mount_Everest**
By the same measure of base to summit, **Mount** McKinley, in Alaska, is also taller than **Everest**. Despite its **height** above sea level of only 6,193.6 m (20,320 ft), **...**
List of deaths on eight ... - Edmund Hillary - Timeline of climbing Mount - 1996

Mt. Everest Height Mystery May Be Answered : Discovery News
news.discovery.com/.../**everest**-official-**height**-120301.htm
Mar 1, 2012 – The plunge from 71581 feet was a success. Next up: 120000 feet.

Facts About **Mt. Everest**
teacher.scholastic.com/activities/hillary/archive/evefacts.htm
Number of people to successfully climb **Mt. Everest**: 660. Number of people who have died trying to climb **Mt. Everest**: 142. **Height**: 29,028 feet, or 5 and a half **...**

Mount Everest by the Numbers: Deaths, Cost to Climb, and More **...**

www.thedailybeast.com/.../**mount-everest**-by-th...
May 22, 2012
8,000: **Height** in meters (approximately 26,000 feet) at **Mount Everest's** "death zone," the low-oxygen area above **...**

More videos for **what is the height of mount everest »**

What is the height of Mount Everest
wiki.answers.com › ... › Geography › Landforms › Mountains
Mt. Everest is 29,002 feet hight. And 348,024 inches high. What is the real **height of Mount Everest**? 12,000 ft!!! Everest is, to begin with, 18,000 ft above sea level **...**

Height of Mount Everest (Everest, Mount) -- Britannica Online **...**
www.britannica.com/EBchecked/.../**Height-of-Mount-Everest**
The **height of Mount Everest**, according to the most recent and reliable data, is 29035 feet (8850 metres). In 1999 an American survey, sponsored by the (U.S.) **...**

Mount Everest - Overview of **Mount Everest**
geography.about.com › ... › Specific Places of Interest
With a peak **elevation** of 29,035 feet (8850 meters), the top of **Mount Everest** is the world's highest point above sea level. As the world's highest mountain, **...**

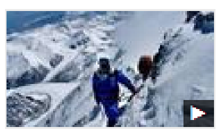
Google what is the height of mount everest

[Mount Everest - Wikipedia, the free encyclopedia]
en.wikipedia.org/wiki/**Mount_Everest**
By the same measure of base to summit,
Everest. Despite its **height** above sea lev
List of deaths on eight ... - Edmund Hillary

[**Mt. Everest Height** Mystery May B]
news.discovery.com/.../**everest**-official-**he**
Mar 1, 2012 – The plunge from 71581 feet

[Facts About **Mt. Everest**]
teacher.scholastic.com/activities/hillary/ar
Number of people to successfully climb **M**
died trying to climb **Mt. Everest**: 142. **Hei**

[**Mount Everest** by the Numbers: D]



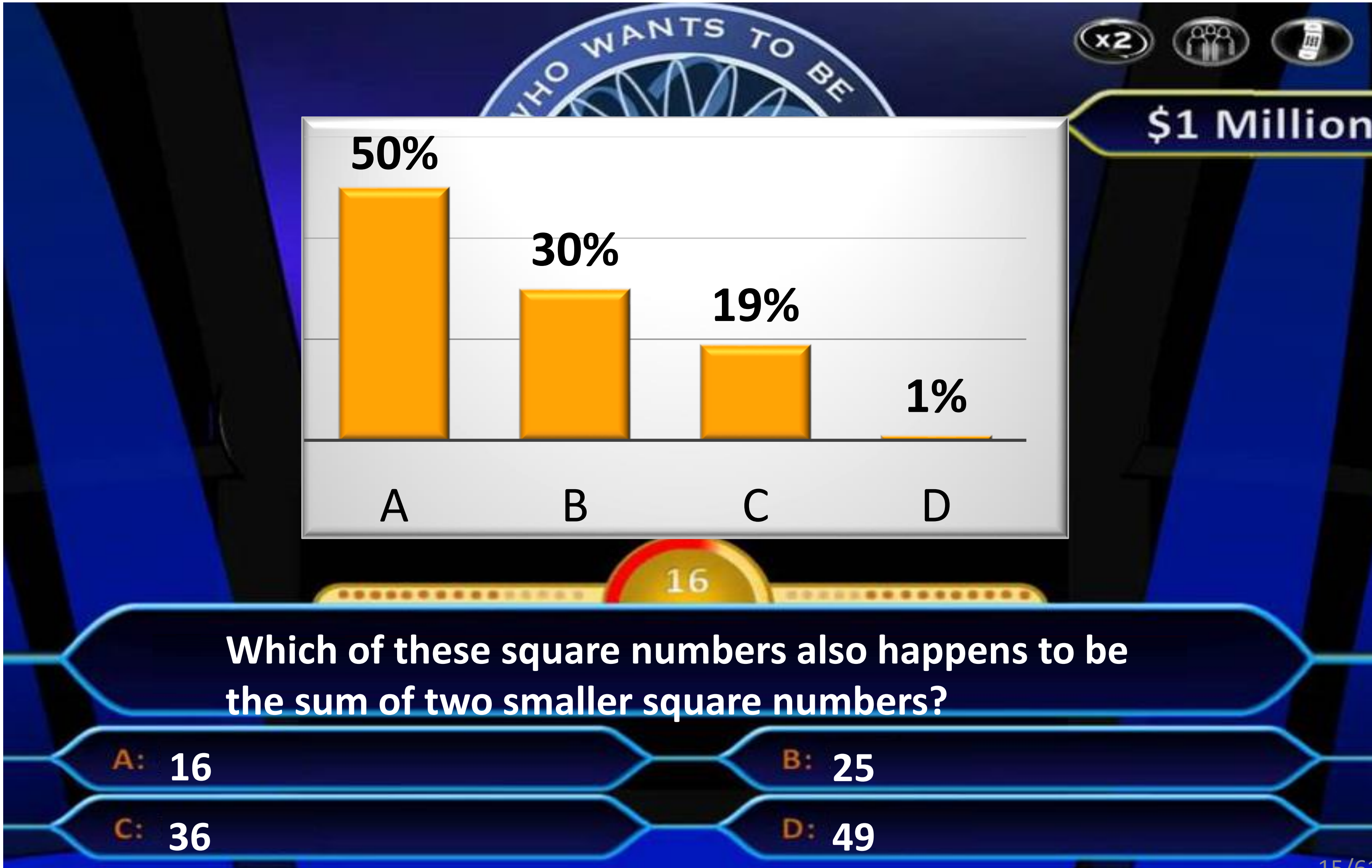
www.thedailybeast.c
May 22, 2012
8,000: **Height** in me
Everest's "death zo
More videos for **what is the height of mo**

[**What is the height of Mount Eve**]
wiki.answers.com › ... › Geography › Land
Mt. Everest is 29,002 feet hight. And 348,
Mount Everest? 12,000 ft!!! Everest is, to

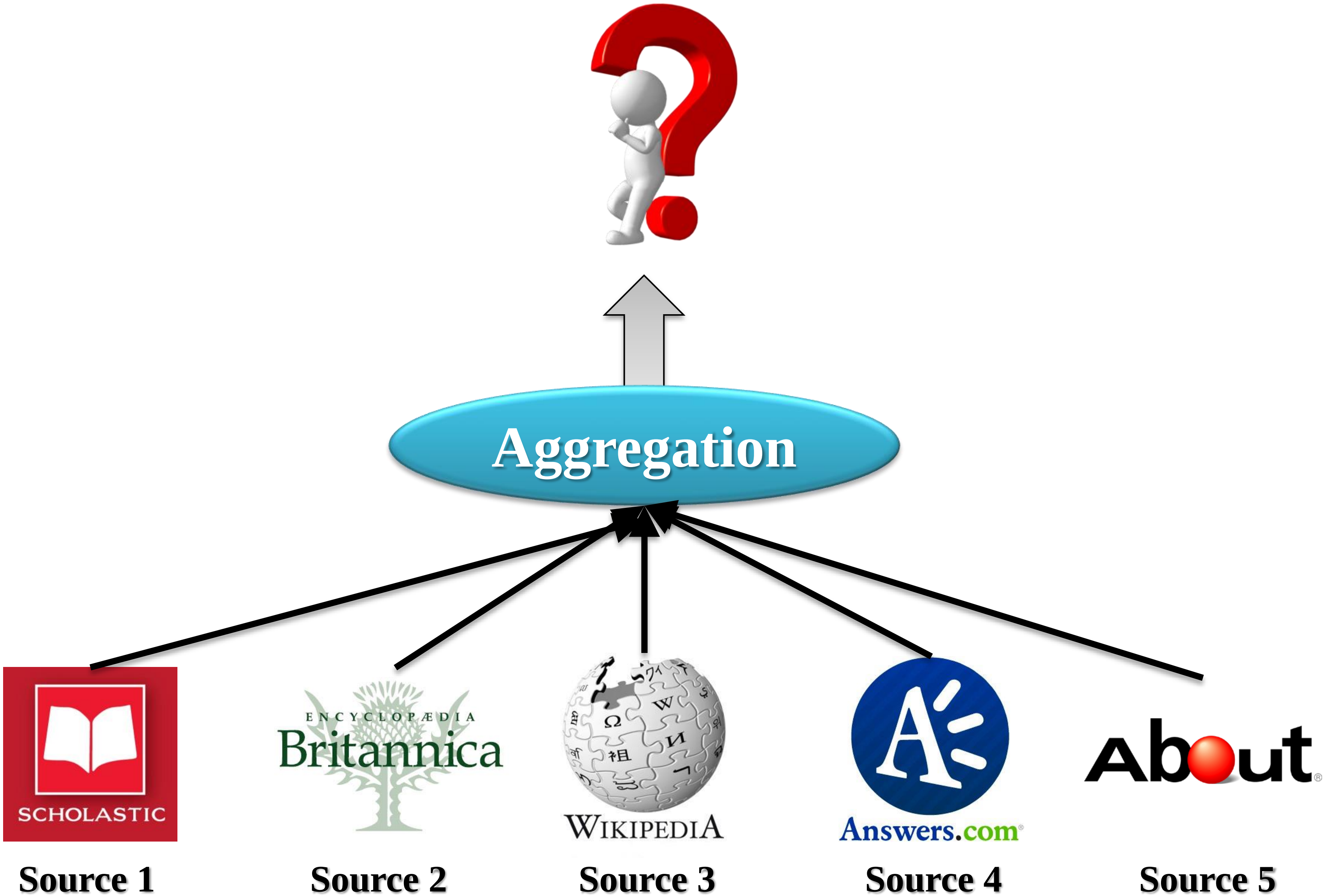
[**Height of Mount Everest** (Everes]
www.britannica.com/EBchecked/.../**Heigh**
The **height of Mount Everest**, according
feet (8850 metres). In 1999 an American s

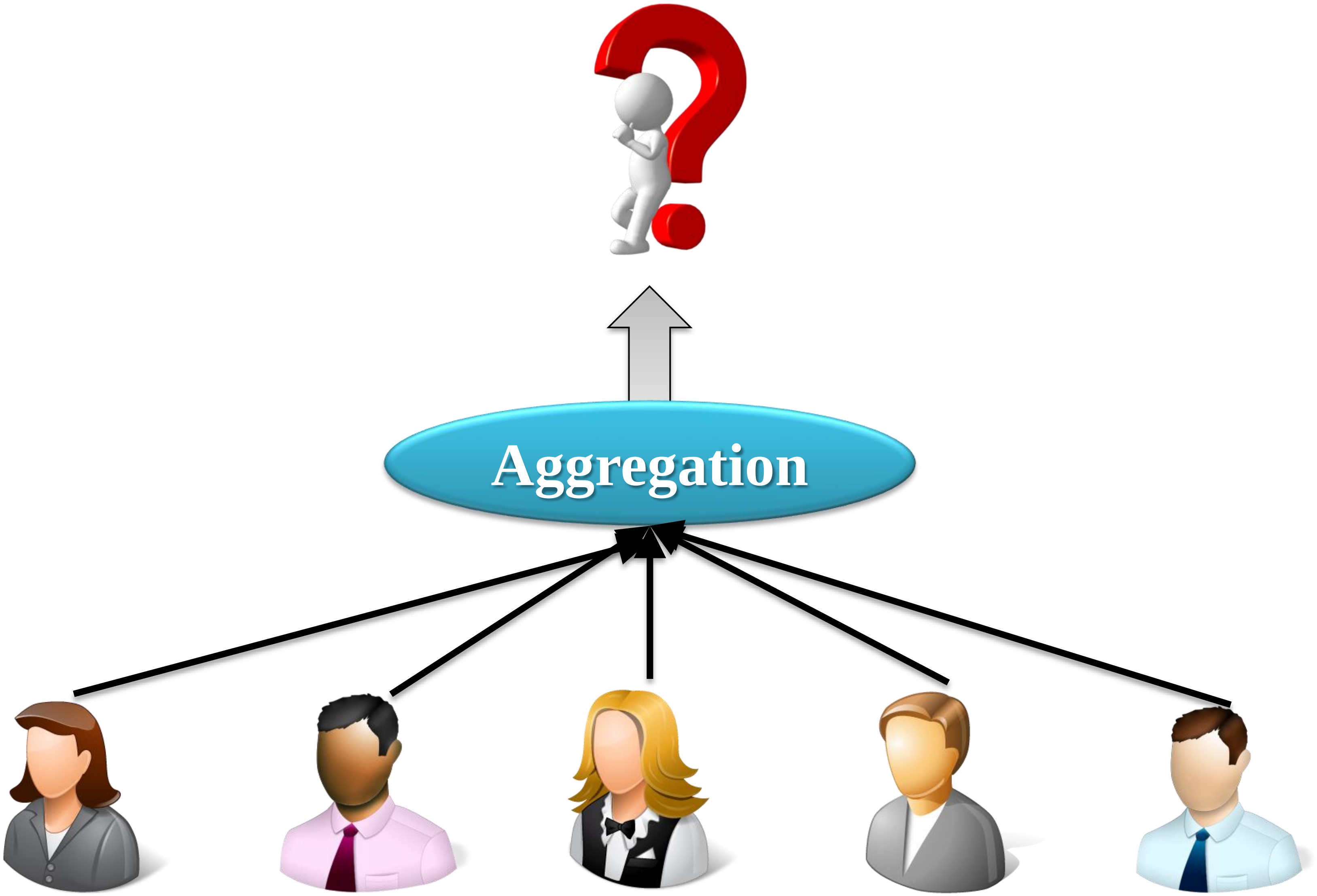
[**Mount Everest** - Overview of **Mou**]
geography.about.com › ... › Specific Place
With a peak **elevation** of 29,035 feet (885
world's highest point above sea level. As the world's highest mountain, ...

Example 2: Crowdsourced Question Answering



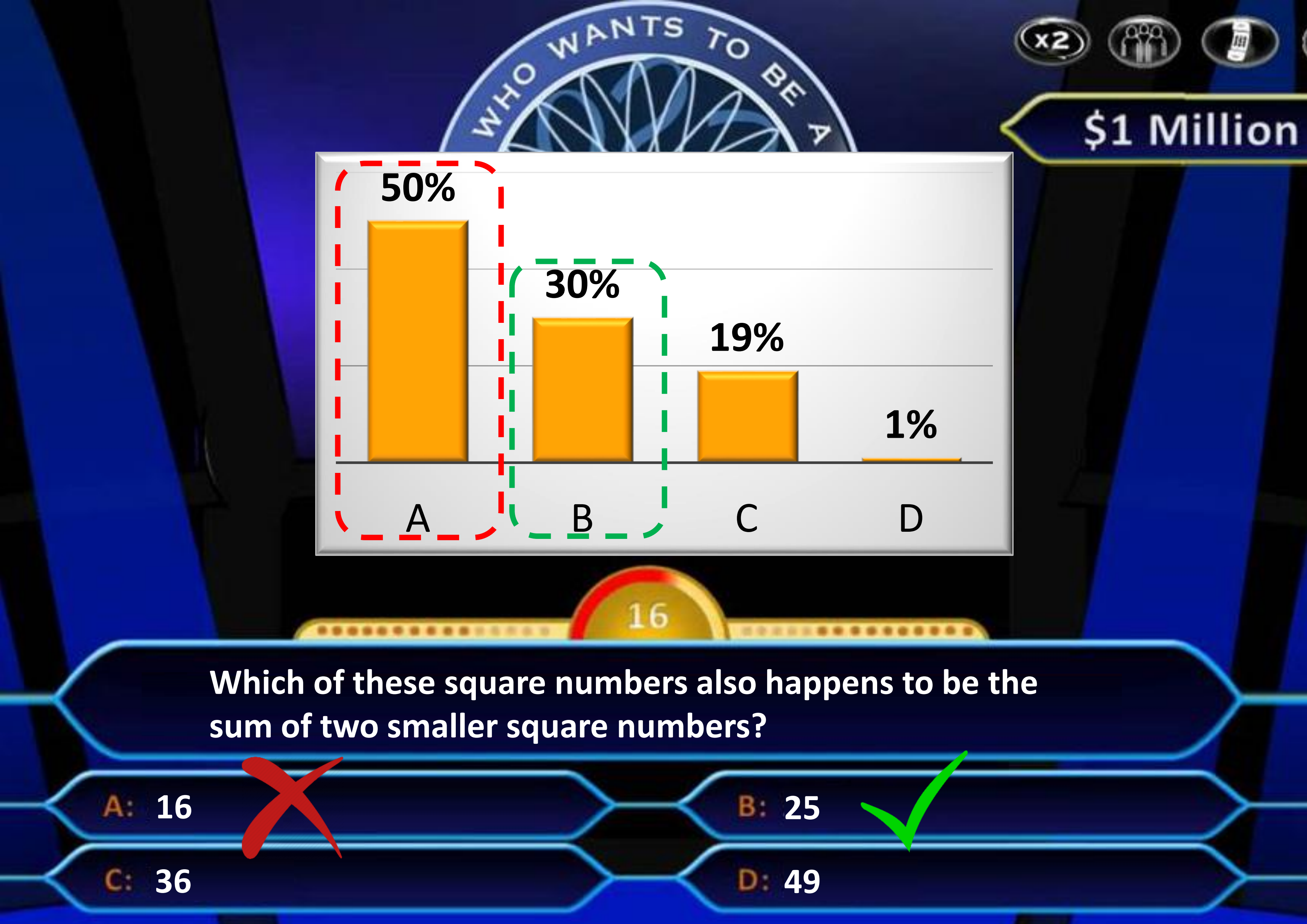
https://www.youtube.com/watch?v=BbX44YSsQ2I


Aggregation
SCHOLASTIC
ENCYCLOPÆDIA Britannica
WIKIPEDIA
Answers.com
About
Source 1
Source 2
Source 3
Source 4
Source 5


Aggregation

A Straightforward Aggregation Solution

- **Voting/Averaging**
  - Take the value that is claimed by majority of the sources
  - Or compute the mean of all the claims
- **Limitation**
  - Ignore source reliability
- **Source reliability**
  - Is crucial for finding the true fact but unknown



https://www.youtube.com/watch?v=BbX44YSsQ2I

A Straightforward Aggregation Solution

- **Voting/Averaging**
  - Take the value that is claimed by majority of the sources
  - Or compute the mean of all the claims
- **Limitation**
  - Ignore source reliability
- **Source reliability**
  - Is crucial for finding the true fact but unknown

Truth Discovery

- **Principle**
  - Infer both truth and source reliability from the data
    - A source is reliable if it provides many pieces of true information
    - A piece of information is likely to be true if it is provided by many reliable sources

Model Categories

- Optimization model (OPT)
- Statistical model (STA)
- Probabilistic graphical model (PGM)

Optimization Model (OPT)

- General model

$$\arg\min_{\{w_s\},\{v_o^*\}} \sum_{o\in O} \sum_{s\in S} g(w_s, v_o^*)$$
$$s.t.\ \delta_1(w_s) = 1, \delta_2(v_o^*) = 1$$

- What does the model mean?
  - Find the optimal solution that minimize the objective function
  - Jointly estimate true claims v_o^* and source reliability w_s under some constraints $\delta_1, \delta_2, \ldots$.
  - Function $g(\cdot,\cdot)$ can be distance, entropy, etc.

Optimization Model (OPT)

- General model

$$\arg\min_{\{w_s\},\{v_o^*\}} \sum_{o\in O} \sum_{s\in S} g(w_s, v_o^*)$$
$$s.t.\ \delta_1(w_s) = 1, \delta_2(v_o^*) = 1$$

- How to solve the problem?
  - Use the method of Lagrange multipliers
  - Block coordinate descent to update parameters
  - If each sub-problem is convex and smooth, then convergence is guaranteed

OPT - CRH Framework

$$\min_{\mathcal{X}^{(*)},\, \mathcal{W}} \quad f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m\left(v_{im}^{(*)}, v_{im}^{(k)}\right)$$

$$\text{s.t.} \;\; \delta(\mathcal{W}) = 1, \qquad \mathcal{W} \geq 0.$$

CRH is a framework that deals with the heterogeneity of data. Different data types are considered, and the estimation of source reliability is jointly performed across all the data types together.

[Li et al., SIGMOD'14]

OPT - CRH Framework

$$\min_{\mathcal{X}^{(*)},\, \mathcal{W}} \quad f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m \left(v_{im}^{(*)}, v_{im}^{(k)} \right)$$

$$\text{s.t.} \quad \delta(\mathcal{W}) = 1, \qquad \mathcal{W} \geq 0.$$

Basic idea

- Truths should be close to the claims from reliable sources
- Minimize the overall weighted distance to the truths in which reliable sources have high weights

OPT - CRH Framework

- **Loss function**
  - d_m: loss on the data type of the m-th property
  - Output a high score when the claim deviates from the truth
  - Output a low score when the claim is close to the truth
- **Constraint function**
  - The objective function may go to $-\infty$ without constraints
  - Regularize the weight distribution

OPT - CRH Framework

- **Run the following until convergence**
  - Truth computation
    - Minimize the weighted distance between the truth and the sources' claims

$$v_{im}^{(*)} \leftarrow \arg\min_{v} \sum_{k=1}^{K} w_k \cdot d_m\left(v, v_{im}^{(k)}\right)$$

  - Source reliability estimation
    - Assign a weight to each source based on the difference between the truths and the claims made by the source

$$\mathcal{W} \leftarrow \arg\min_{\mathcal{W}} f(\mathcal{X}^{(*)}, \mathcal{W})$$

Statistical Model (STA)

- General goal:
  - **To find the (conditional) probability of a claim being true**
- Source reliability:
  - **Probability(ies) of a source/worker making a true claim**

Statistical Model (STA)

- Models
  - Apollo-MLE [Wang et al., ToSN'14]
  - TruthFinder [Yin et al., TKDE'08]
  - Investment, Pool Investment [Pasternack&Roth, COLING'10]
  - Cosine, 2-estimate, 3-estimate [Galland et al., WSDM'10]

STA - TruthFinder

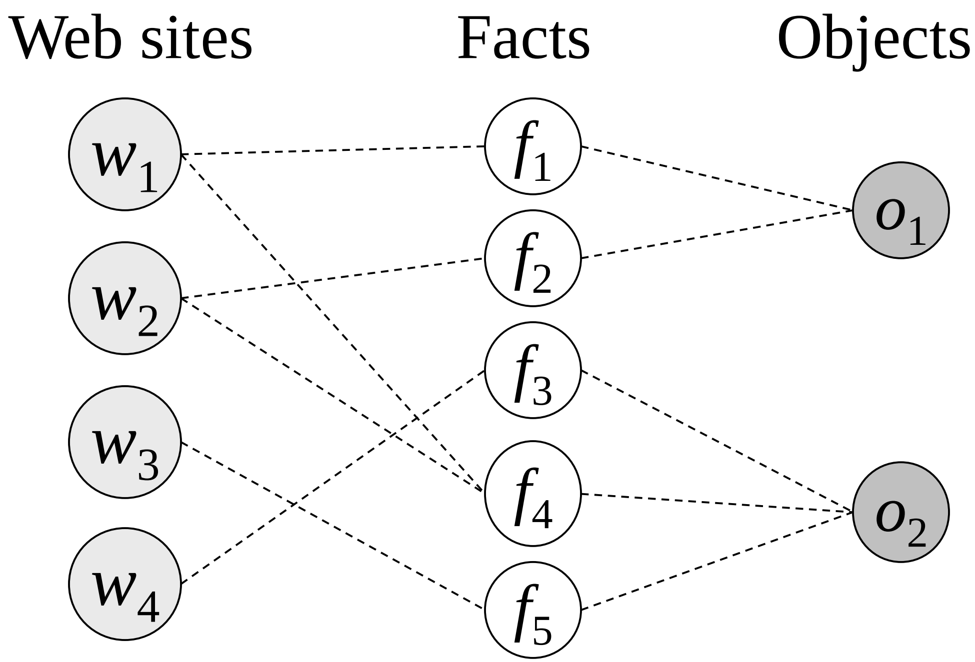
Different websites often provide conflicting information on a subject, e.g., Authors of *"Rapid Contextual Design"*

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

[Yin et al., TKDE'08]

STA - TruthFinder

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



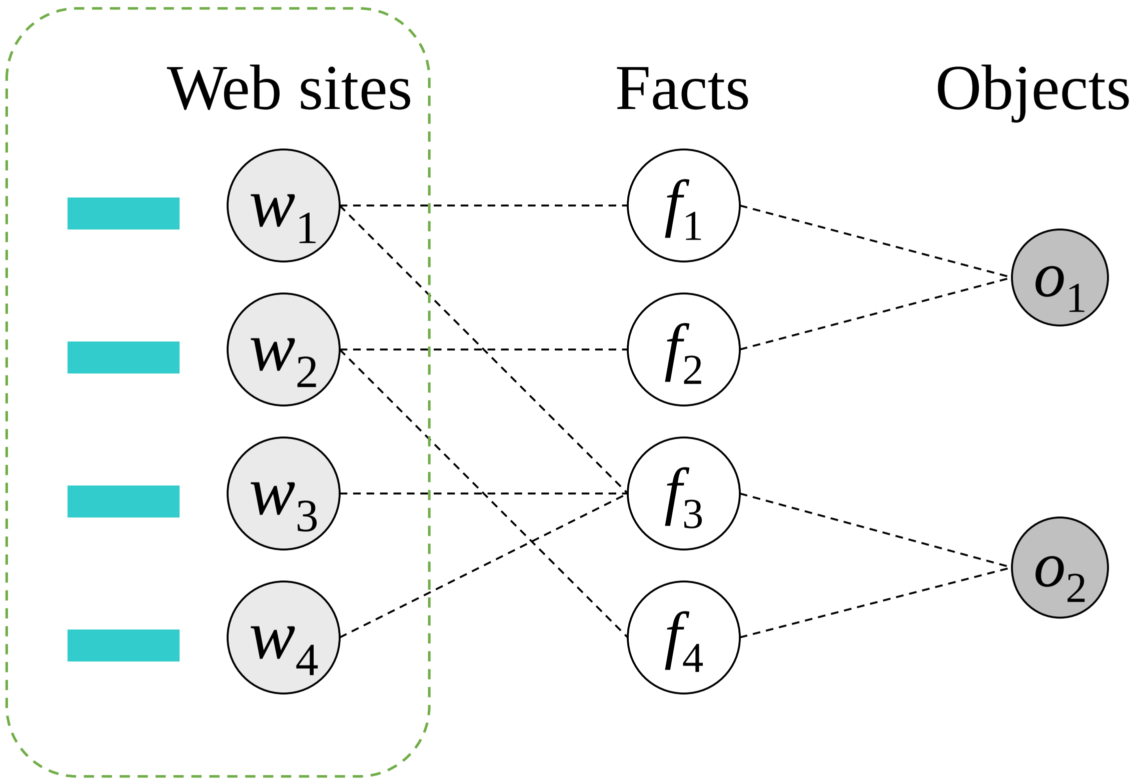
STA - TruthFinder

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**

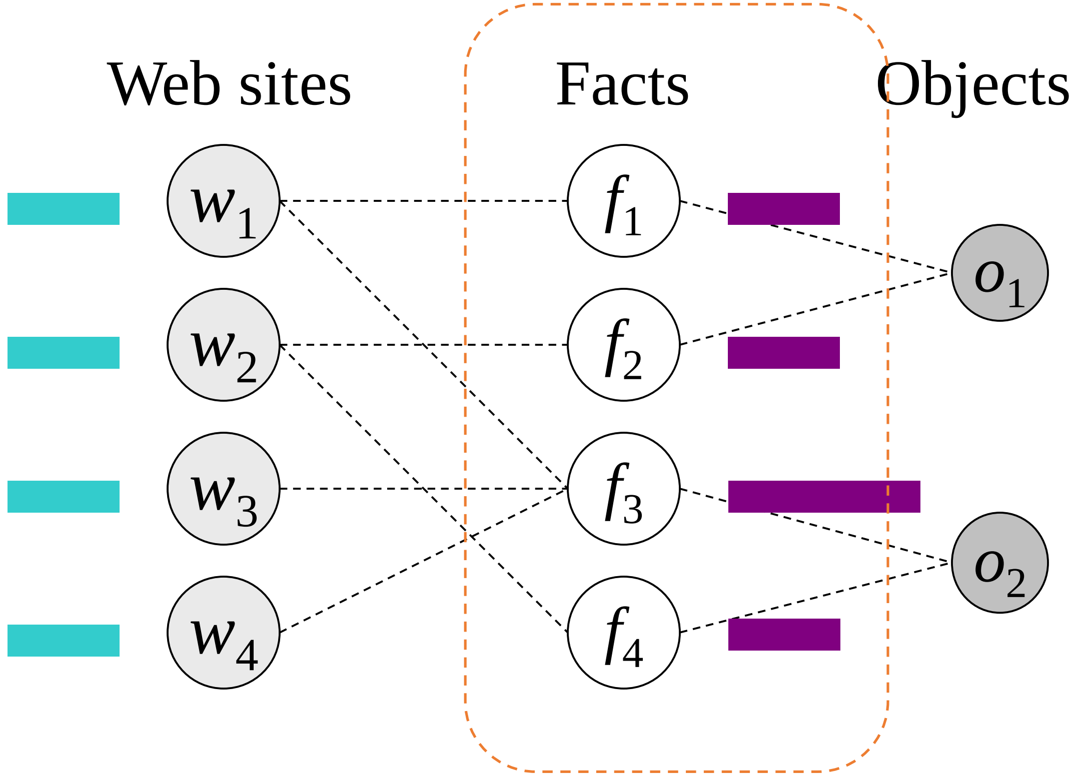
STA - TruthFinder

- *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
- Iterative steps
  - Initially, each web site is equally trustworthy
  - Based on the four heuristics, infer fact confidence from web site trustworthiness, and then backwards
  - Repeat until achieving stable state

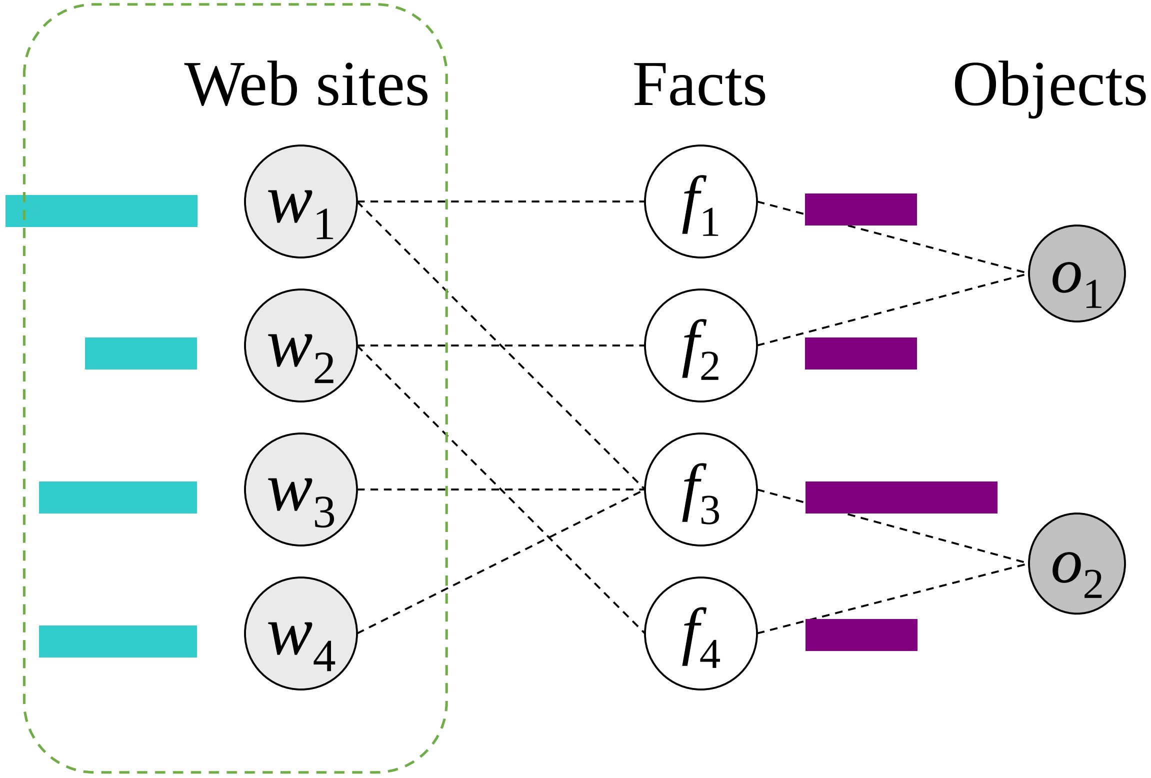
STA - TruthFinder



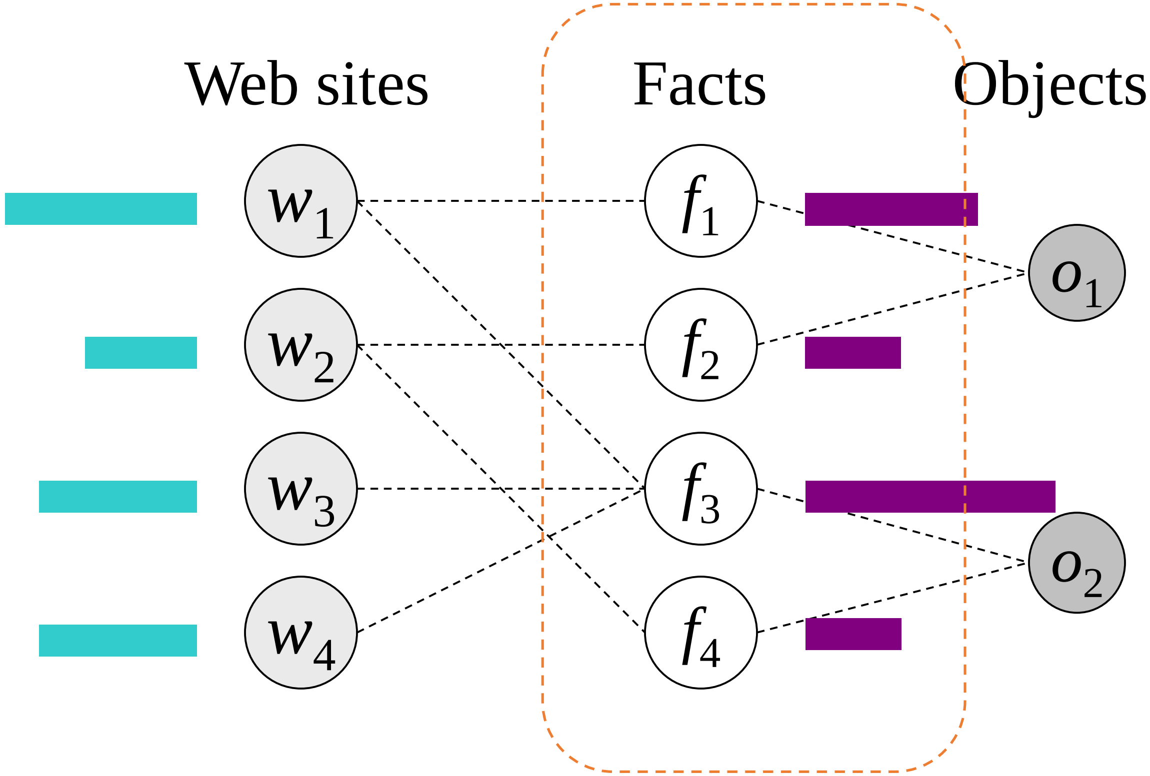
STA - TruthFinder



STA - TruthFinder



STA - TruthFinder



38

STA - TruthFinder

- **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

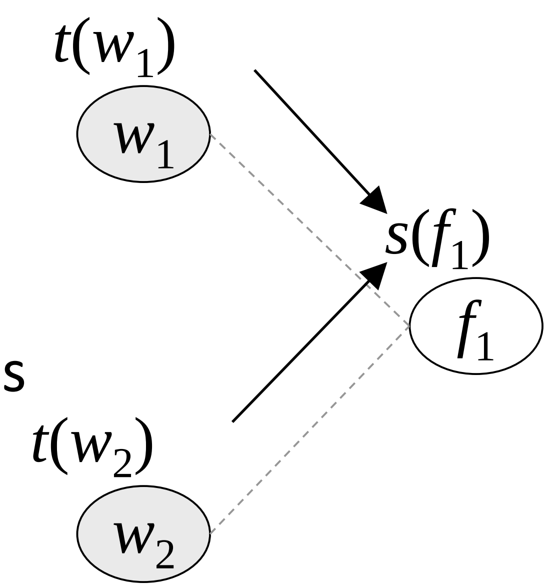
Set of facts provided by w

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

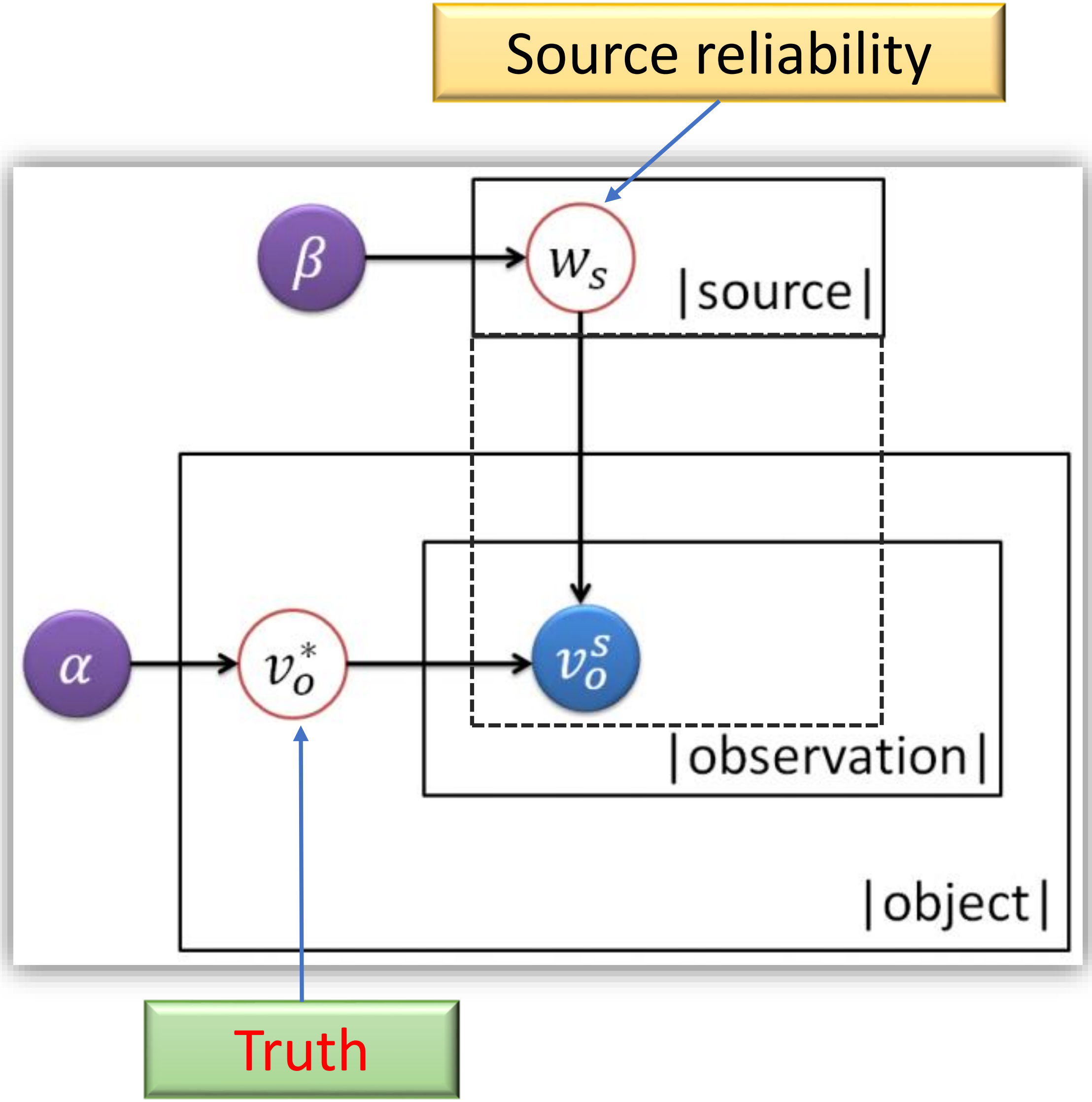
Probability that w is wrong

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

Set of websites providing f



Probabilistic Graphical Model (PGM)



Probabilistic Graphical Model (PGM)

- Models
  - GTM [Zhao&Han, QDB'12]
  - LTM [Zhao et al., VLDB'12]
  - MSS [Qi et al., WWW'13]
  - LCA [Pasternack&Roth, WWW'13]
  - TEM [Zhi et al., KDD'15]

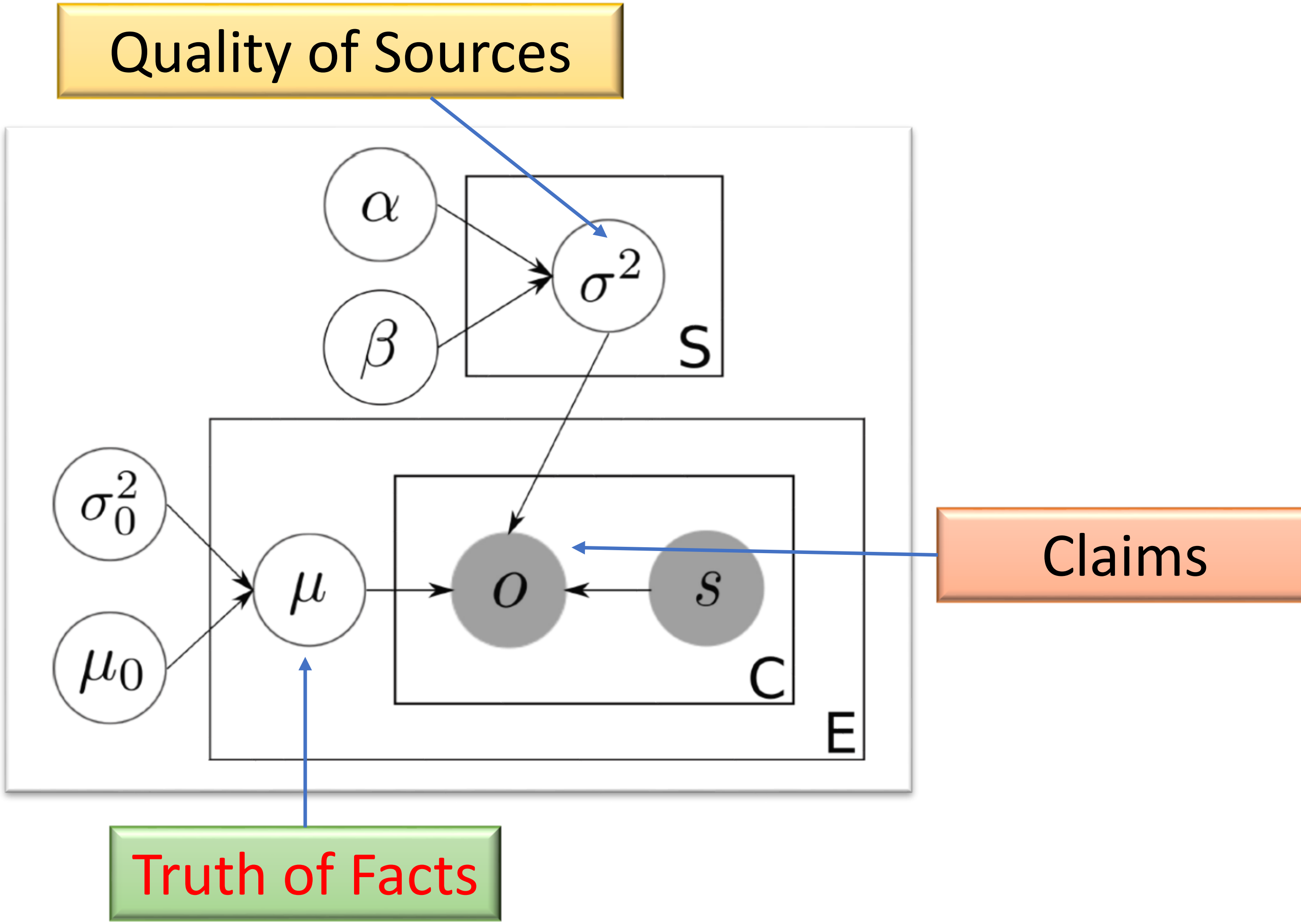
  …

PGM – Gaussian Truth Model (GTM)

- **Real-valued** Truths and Claims
  - Population of a city is numerical
- The quality of sources is modeled as how **close** their claims are to the truth
  - Distance is better than accuracy for numerical data
- Sources and objects are **independent** respectively

[Zhao&Han, QDB'12]

PGM – Gaussian Truth Model (GTM)



PGM – Gaussian Truth Model (GTM)

- For each source k
  - Generate its quality from a prior inverse Gamma distribution : $\sigma_s^2 \sim Inv - Gamma(\alpha, \beta)$
- For each fact f
  - Generate its prior truth from a prior Gaussian distribution: $\mu_e \sim Gaussian(\mu_0, \sigma_0^2)$
- For each claim c of fact f, generate claim of c.
  - Generate it from a Gaussian distribution with truth as mean and the quality as variance: $o_c \sim Gaussian\left(\mu_e, \sigma_{s_c}^2\right)$

Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. **Truth Discovery Scenarios**
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References

Number of Truths for One Object

- Single truth
  - Each object has one and only one truth
  - The claims from sources contain the truth
  - Complementary vote
- Multiple truth
  - Each object may have more than true fact
  - Each source may provide more than one fact for each object
- Existence of truths
  - The true fact for an object may be not presented by any sources

Single Truth

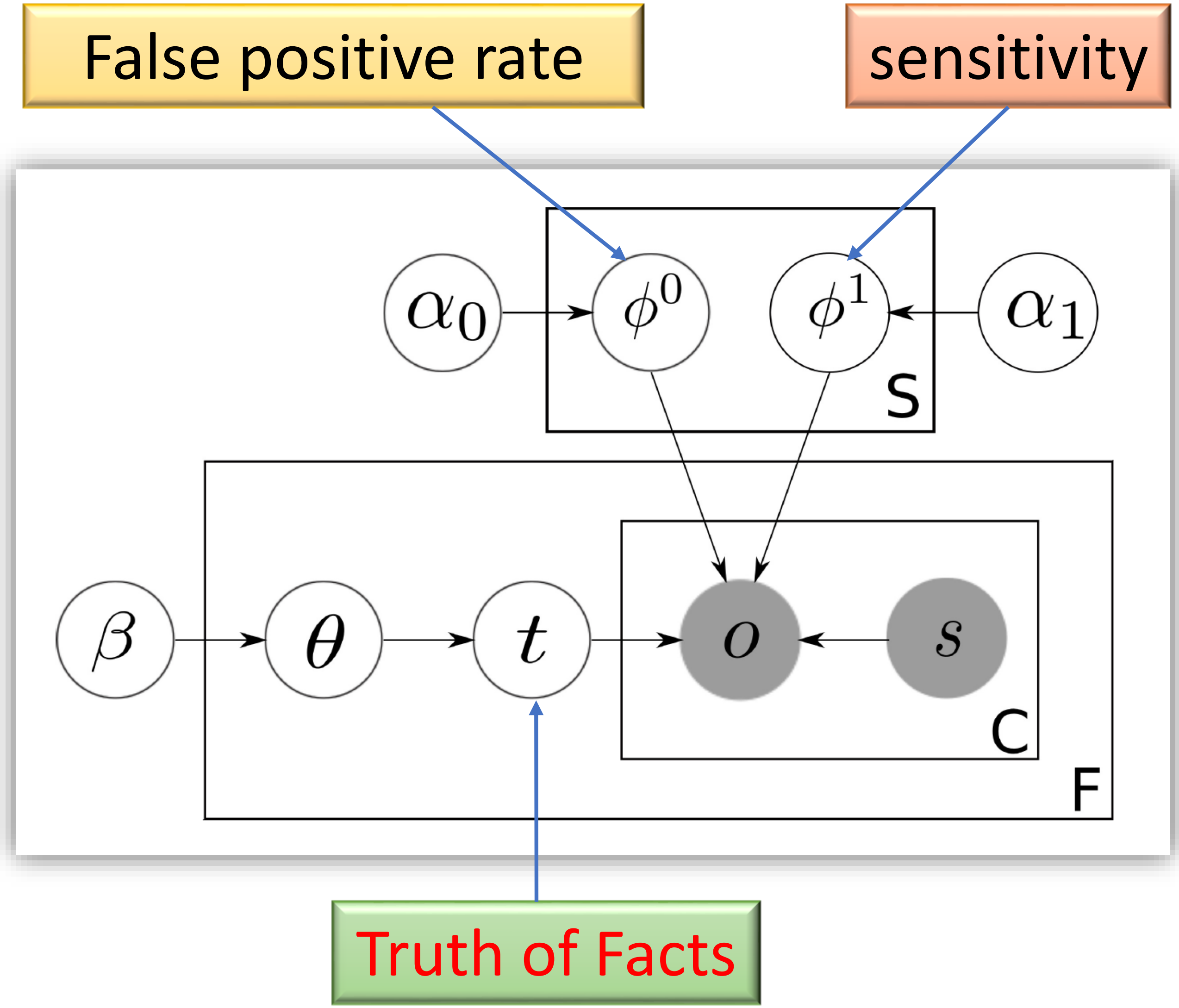
- Example
  - A person's birthday
  - Population of a city
  - Address of a shop
- Complementary vote
  - If a source makes a claim on an object, that source considers all the other claims as false
- Positive vote only [Wang et al., ToSN'14]
  - An event only receive positive claims, but no negative claims. E.g., people only report that they observe an event.

Multiple Truth- Latent Truth Model (LTM)

- **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
- Source reliability
  - False positive: making a wrong claim
  - Sensitivity: missing a claim
- Modeled in PGM

[Zhao et al., VLDB'12]

Multiple Truth- Latent Truth Model (LTM)



Multiple Truth- Latent Truth Model (LTM)

- 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 claim 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: $o_c \sim Bernoulli(\phi_{s_c}^1)$

Existence of Truth

- Truth Existence problem: when the true answers are excluded from the candidate answers provided by all sources.
  - *Has-truth questions*: correct answers exist among the candidate answers provided by all sources.
  - *No-truth questions*: true answers are not included in the candidate answers provided by all sources.
- Without any prior knowledge, the no-truth questions are hard to distinguish from the has-truth ones.
  - These no-truth questions degrade the precision of the answer integration system.
- Example: Slot Filling Task

[Yu et al., COLING'14][Zhi et al., KDD'15]

Existence of Truth

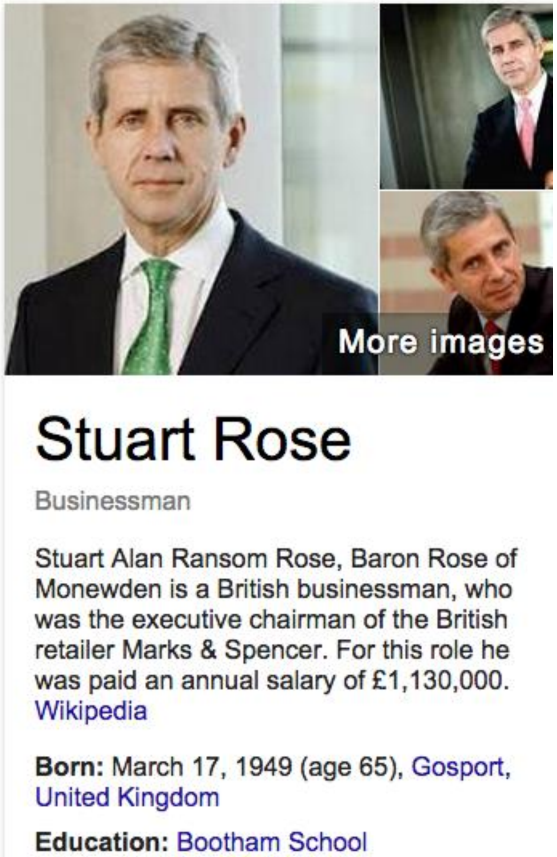
Example: Slot Filling Task

Table 1: Example Questions of Slot Filling Task

	Question
q_1	What's the age of Ramazan Bashardost?
q_2	What's the country of birth of Ramazan Bashardost?
q_3	What's the province of birth of Ramazan Bashardost?
q_4	What's the age of Marc Bolland?
q_5	What's the country of birth of Marc Bolland?
q_6	What's the age of Stuart Rose?
q_7	What's the country of birth of Stuart Rose?
q_8	What's the province of death of Stuart Rose?

Hard to detect

Invalid



Existence of Truth

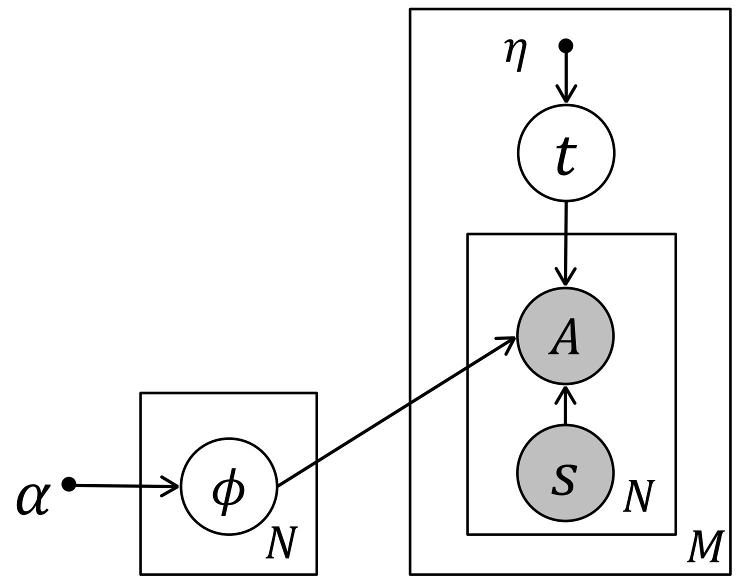
Source	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
s_1	43			50				
s_2	:11			:31			UK	
s_3	627	Pakistan		50				
s_4		Afghanistan	Ghazni					
s_5	43			50				
s_6	43		Ghazni	50				
s_7	43	Afghanistan	Ghazni	50				
s_8	IL	Khost		Mar	London	Marks	Russia	Holly
s_9	/25	Pakistan	Pakistan			actor	Spence	Spence
s_{10}			Kabul					
s_{11}	43			50				
s_{12}		Afghanistan	Ghazni					
s_{13}	9		Ghazni	50	Holland			
Truth	**43**	**Afghanistan**	**Ghazni**	**50**	**Empty**	**Empty**	**Empty**	**Empty**

Has-truth questions

No-truth questions

Existence of Truth - Truth Existence Model (TEM)

- Probabilistic Graphical Model
  - Output
    - t: latent truths
    - ϕ : source quality
  - Input
    - A: observed answers
    - S: sources
  - Parameters (fixed)
    - Prior of source quality: α
    - Prior of truth: η $\quad \eta_{i0} = P(t_i = E), \eta_{in} = P(t_i = d_{in})$



- Maximum Likelihood Estimation
- Inference: EM

Source Dependency

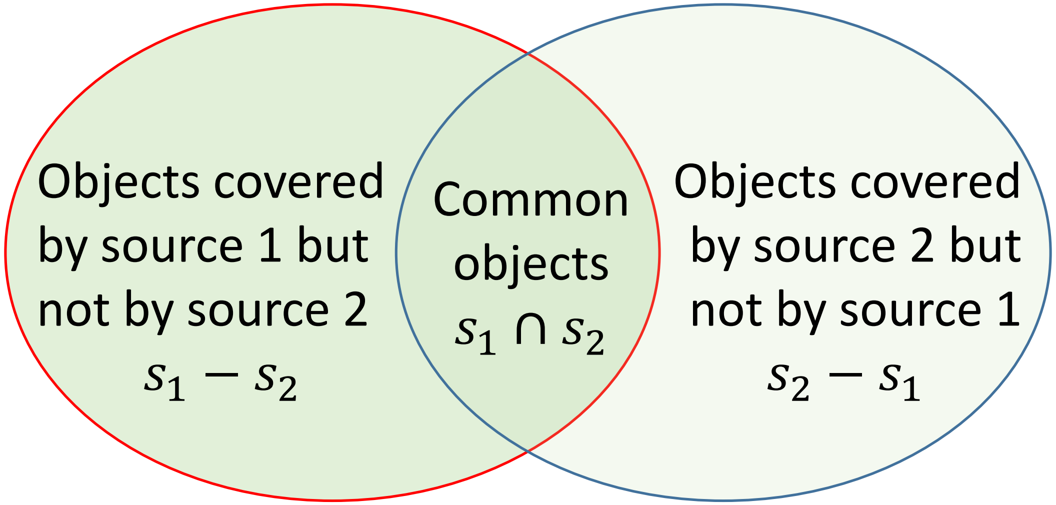
- Many truth discovery methods considers independent sources
  - Sources provide information independently
  - Source correlation can be hard to model
  - However, this assumption may be violated in real life
- Copy relationships between sources
  - Sources can copy information from one or more other sources
- General correlations of sources

Source Dependency

- Known relationships
  - Apollo-Social [Wang et al., IPSN'14]
    - For a claim, a source may copy from a related source with a certain probability
    - Used MLE to estimate a claim being correct
- Unknown relationships
  - Accu-Copy [Dong et al., VLDB'09a] [Dong et al., VLDB'09b]
  - MSS [Qi et al., WWW'13]
    - Modeled as a PGM
    - Related sources are grouped together and assigned with a group weight

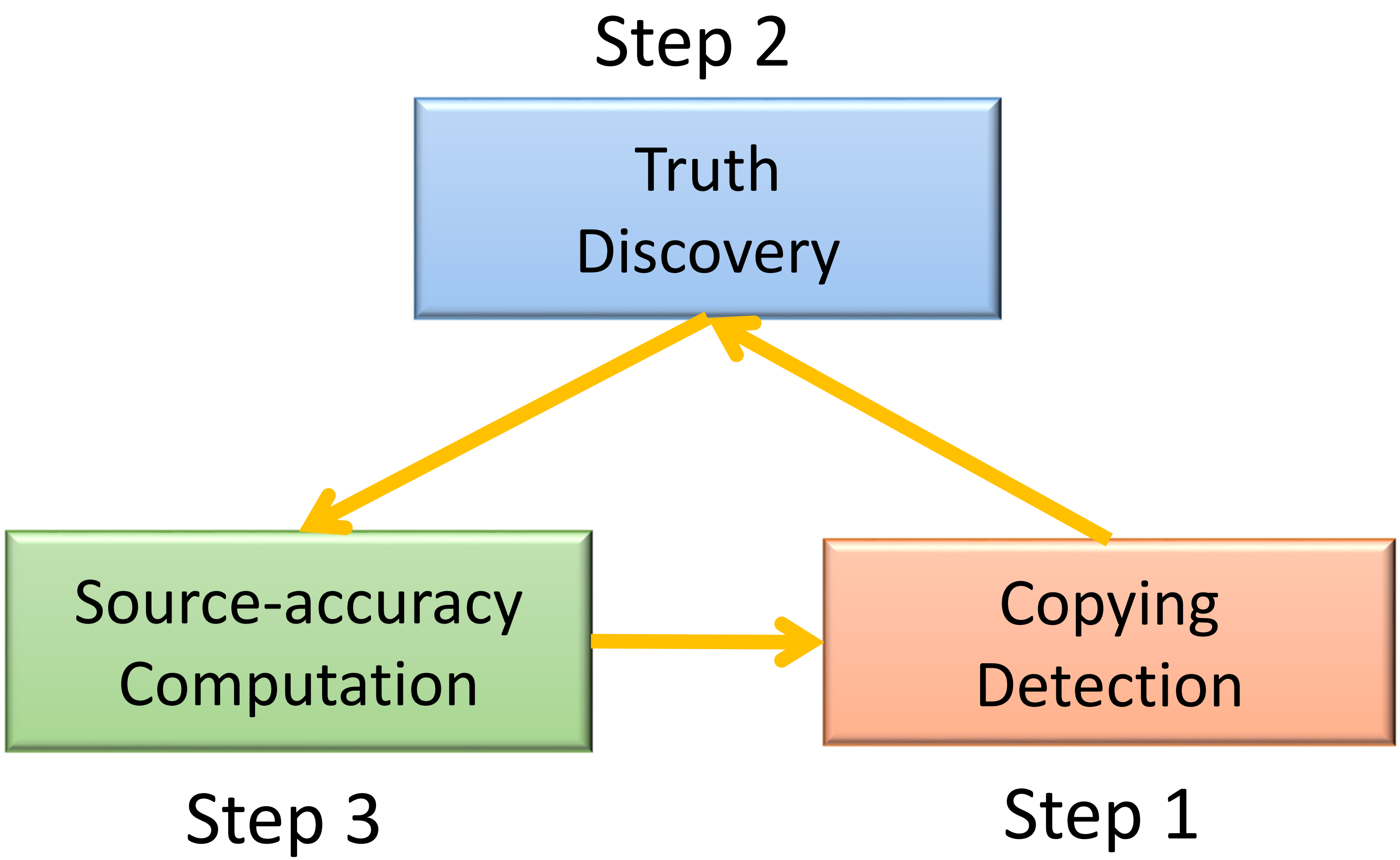
Copy Relationships between Sources

- High-level intuitions for copying detection
  - Common error implies copying relation
    - e.g., many same errors in $s_1 \cap s_2$ imply source 1 and 2 are related
  - Source reliability inconsistency implies copy direction
    - e.g., $s_1 \cap s_2$ and $s_1 - s_2$has similar accuracy, but $s_1 \cap s_2$ and $s_2 - s_1$ has different accuracy, so source 2 may be a copier.



Copy Relationships between Sources

- Incorporate copying detection in truth discovery



[Dong et al., VLDB'09a] [Dong et al., VLDB'09b]

General Source Correlation

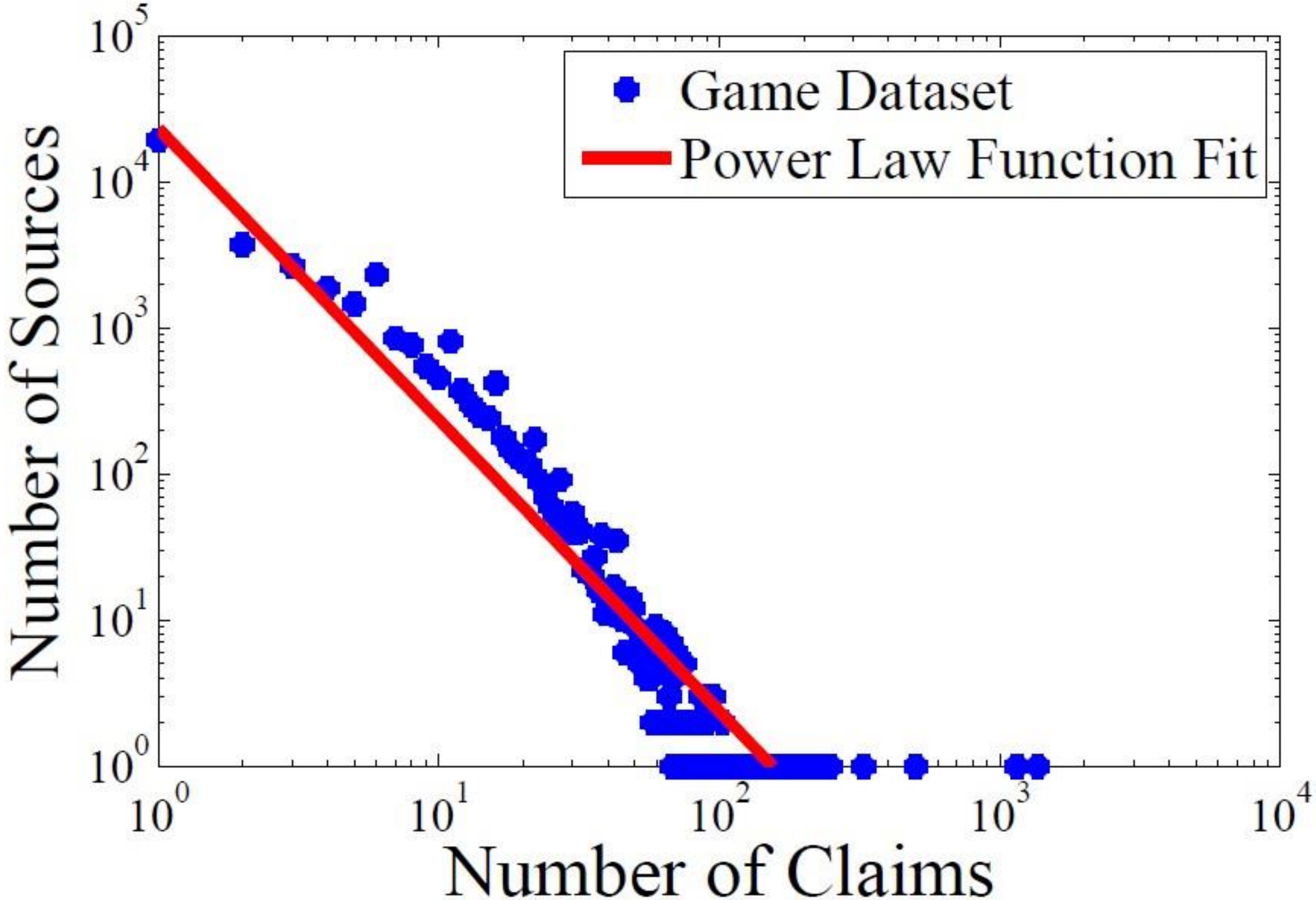
- More general source correlations
  - Sources may provide data from complementary domains (negative correlation)
  - Sources may focus on different types of information (negative correlation)
  - Sources may apply common rules in extraction (positive correlation)
- How to detect
  - Hypothesis test of independence using joint precision and joint recall

[Pochampally et al., SIGMOD'14]

Information Density

- Dense information
  - Each source provides plenty of claims
  - Each object receives plenty of information from sources
- Long-tail phenomenon on sources side
  - Many sources provide limited information
  - Only a few sources provide sufficient information
- Auxiliary information
  - Text of question/answers
  - Fine-grained source reliability estimation

Long-tail Phenomenon on Sources Side



Long-tail Phenomenon on Sources Side - CATD

- Challenge when most sources make a few claims
  - Sources weights are usually estimated as proportional to the accuracy of the sources
  - If long-tail phenomenon occurs, most source weights are not properly estimated.
- A confidence-aware approach
  - not only estimates source reliability
  - but also considers the confidence interval of the estimation
- An optimization based approach

[Li et al., VLDB’15]

Long-tail Phenomenon on Sources Side - CATD

- Assume that sources are independent and error made by source s: $\epsilon_s \sim N(0,\ \sigma_s^2)$
- $\epsilon_{aggregate} = \frac{\sum_{s\in S} w_s \epsilon_s}{\sum_{s\in S} w_s} \sim N\left(0, \frac{\sum_{s\in S} w_s^2 \sigma_s^2}{\left(\sum_{s\in S} w_s\right)^2}\right)$

Without loss of generality, we constrain $\sum_{s\in S} w_s = 1$

- **Optimization**

$$
\begin{aligned}
\min_{\{w_s\}} \quad & \sum_{s\in\mathcal{S}} w_s^2 \boxed{\sigma_s^2} \\
\text{s.t.} \quad & \sum_{s\in\mathcal{S}} w_s = 1, \\
& w_s \geqslant 0, \forall s \in \mathcal{S}.
\end{aligned}
$$

Long-tail Phenomenon on Sources Side - CATD

Sample variance:

$$\widehat{\sigma_s^2} = \frac{1}{|N_s|} \sum_{n \in N_s} \left(x_n^s - x_n^{*(0)} \right)^2$$

where $x_n^{*(0)}$ is the initial truth.

The estimation is not accurate with small number of samples.

Find a range of values that can act as good estimates.

Calculate confidence interval based on

$$\frac{|N_s|\widehat{\sigma_s^2}}{\sigma_s^2} \sim \chi^2(|N_s|)$$

Long-tail Phenomenon on Sources Side - CATD

- Consider the possibly worst scenario of σ_s^2
- Use the upper bound of the 95% confidence interval of σ_s^2

$$u_s^2 = \frac{\sum_{n \in N_s} \left(x_n^s - x_n^{*(0)} \right)^2}{\chi^2_{(0.05, |N_s|)}}$$

Long-tail Phenomenon on Sources Side - CATD

$$\min_{\{w_s\}} \quad \sum_{s \in \mathcal{S}} w_s^2 u_s^2$$

$$\text{s.t.} \quad \sum_{s \in \mathcal{S}} w_s = 1, w_s \geqslant 0, \forall s \in \mathcal{S}.$$

- Closed-form solution:

$$w_s \propto \frac{1}{u_s^2} = \frac{\chi^2_{(0.05,|N_s|)}}{\sum_{n \in N_s} \left(x_n^s - x_n^{*(0)} \right)^2}$$

Long-tail Phenomenon on Sources Side - CATD

Example on calculating confidence interval

Source ID	# Claims	$\hat{\sigma}_s^2$	Confidence Interval (95%)
Source A	200	0.1	(0.0830, 0.1229)
Source B	200	3	(2.4890, 3.6871)
Source C	2	0.1	(0.0271, 3.9498)
Source D	2	3	(0.8133, 118.49)

Long-tail Phenomenon on Sources Side - CATD

Example on calculating source weight

Source ID	$\hat{\sigma}_s^2$	u_s^2	Source Weight (based on $\hat{\sigma}_s^2$)	Source Weight (based on u_s^2)
Source A	0.1	0.1229	0.4839	0.9385
Source B	3	3.6871	0.0161	0.0313
Source C	0.1	3.9498	0.4839	0.0292
Source D	3	118.49	0.0161	0.0010

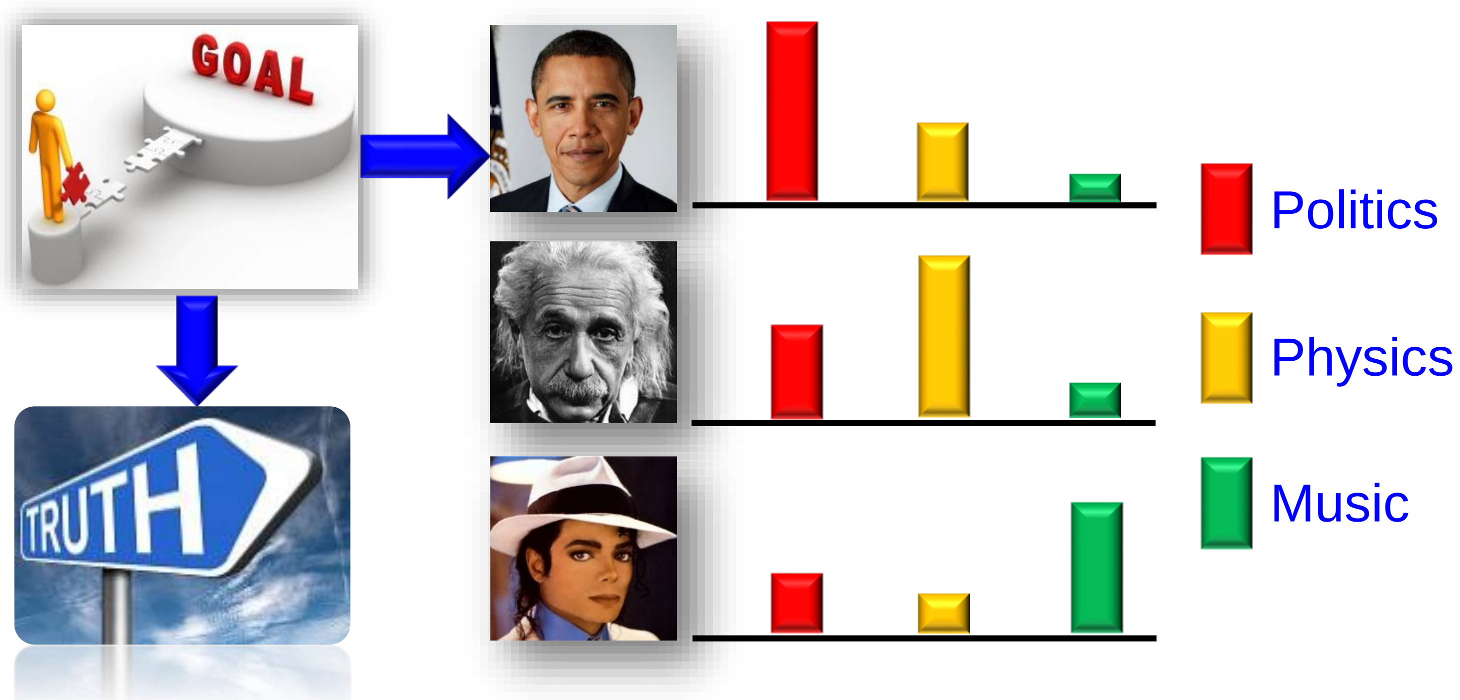
Long-tail Phenomenon on Sources Side - CATD

Question level	Error rate of Majority Voting	Error rate of CATD
1	0.0297	**0.0132**
2	0.0305	**0.0271**
3	0.0414	**0.0276**
4	0.0507	**0.0290**
5	0.0672	**0.0435**
6	0.1101	**0.0596**
7	0.1016	**0.0481**
8	0.3043	**0.1304**
9	0.3737	**0.1414**
10	0.5227	**0.2045**

Higher level indicates harder questions

Fine-Grained Truth Discovery - FaitCrowd

- To learn **fine-grained (topical-level) user expertise** and the **truths** from conflicting crowd-contributed answers.
- Topic is learned from question&answer texts



[Ma et al., KDD'15]

Fine-Grained Truth Discovery - FaitCrowd

- **Input**
  - Question Set
  - User Set
  - Answer Set
  - Question Content
- **Output**
  - Questions' Topic
  - Topical-Level Users' Expertise
  - Truths

Question	User u1	User u2	User u3	Word	
q1	1	2	1	a	b
q2	2	1	2	b	c
q3	1	2	2	a	c
q4	1	2	2	d	e
q5	2		1	e	f
q6	1	2	2	d	f

Topic	Question		
K1	q1	q2	q3
K2	q4	q5	q6

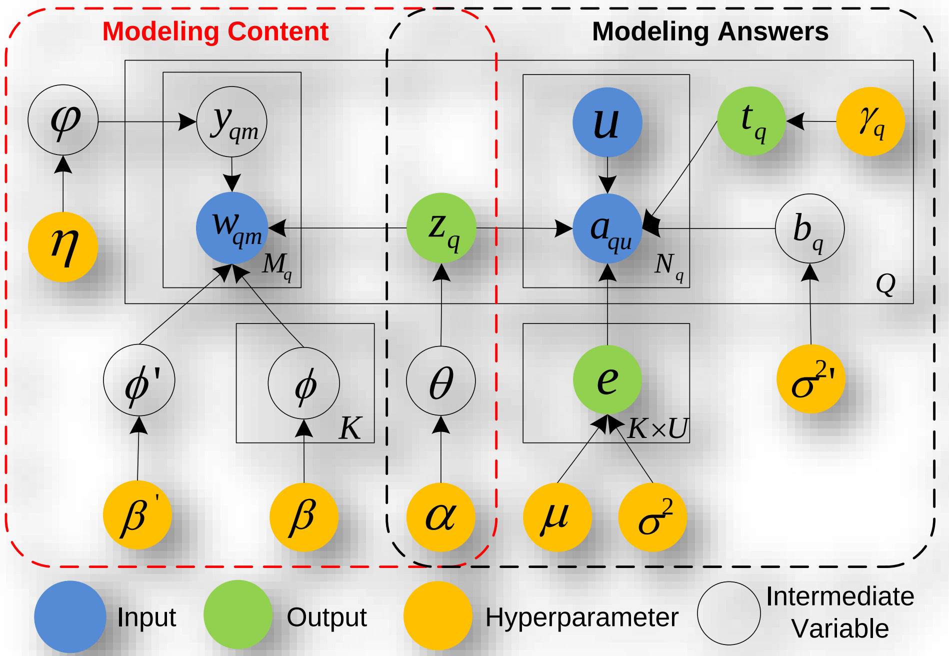
User		u1	u2	u3
Expertise	K1	2.34	2.70E-4	1.00
	K2	1.30E-4	2.34	2.35

Question	q1	q2	q3	q4	q5	q6
Truth	1	2	1	2	1	2

Question	q1	q2	q3	q4	q5	q6
Ground Truth	1	2	1	2	1	2

Fine-Grained Truth Discovery - FaitCrowd

- **Overview**

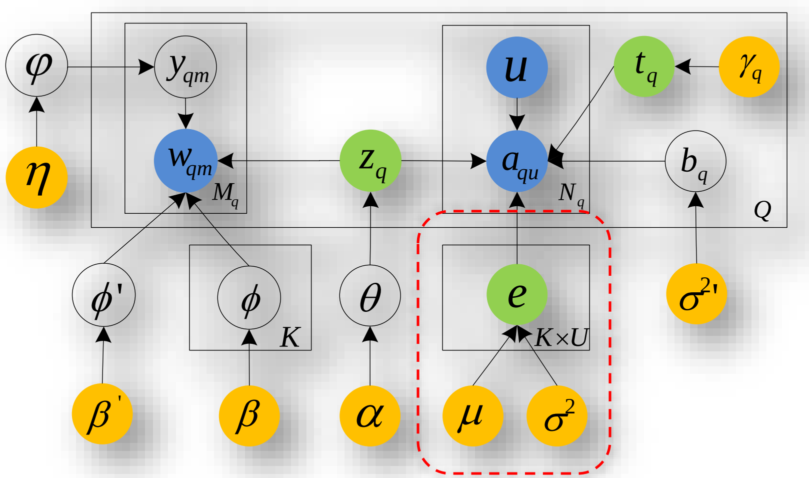


- Jointly modeling question content and users' answers by introducing latent topics.
- Modeling question content can help estimate reasonable user reliability, and in turn, modeling answers leads to the discovery of meaningful topics.
- Learning topics, topic-level user expertise and truths simultaneously.

Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

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



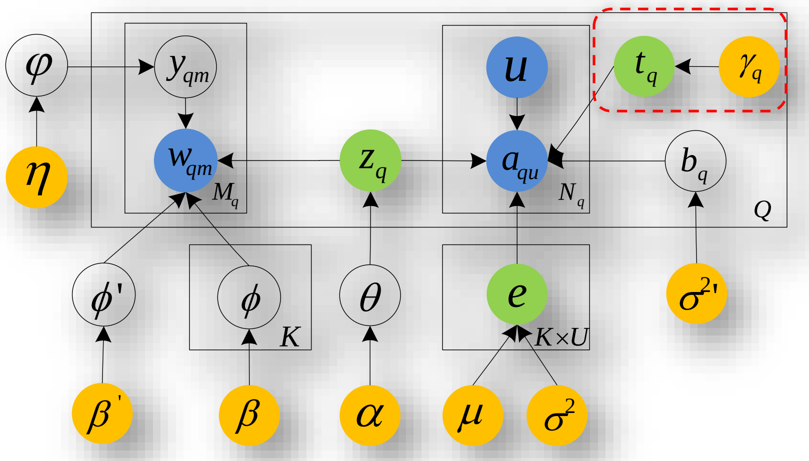
Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

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

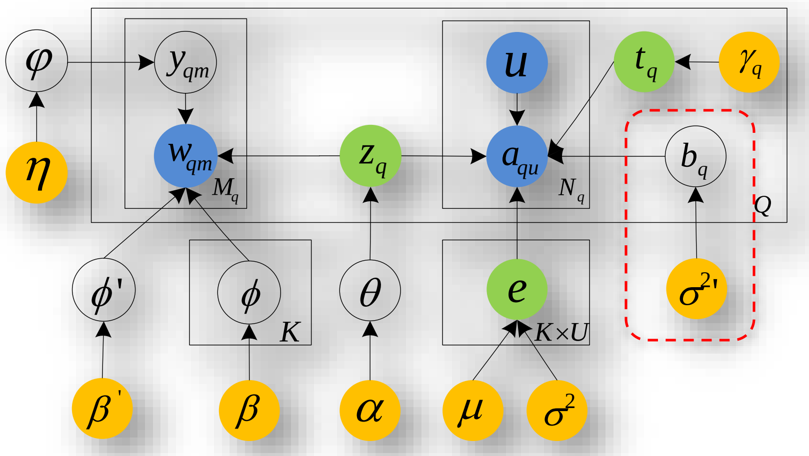
    - Draw the truth

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



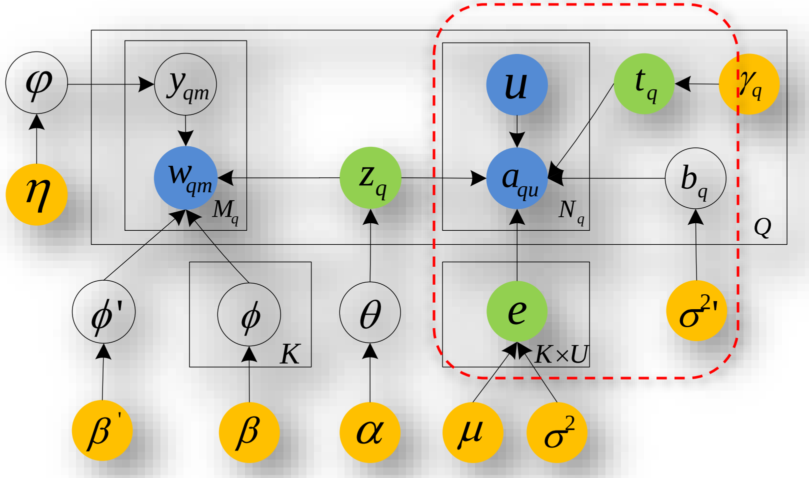
Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise
      $$e_{z_q u} \sim N(\mu, \sigma^2)$$
    - Draw the truth
      $$t_q \sim U(\gamma_q)$$
    - Draw the bias
      $$b_q \sim N(0, \sigma^{2\prime})$$



Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise
      $$e_{z_q u} \sim N(\mu, \sigma^2)$$
    - Draw the truth
      $$t_q \sim U(\gamma_q)$$
    - Draw the bias
      $$b_q \sim N(0, \sigma^{2'})$$
    - Draw a user's answer
      $$a_{qu} | t_q \sim logistic(e_{z_q u}, b_q)$$



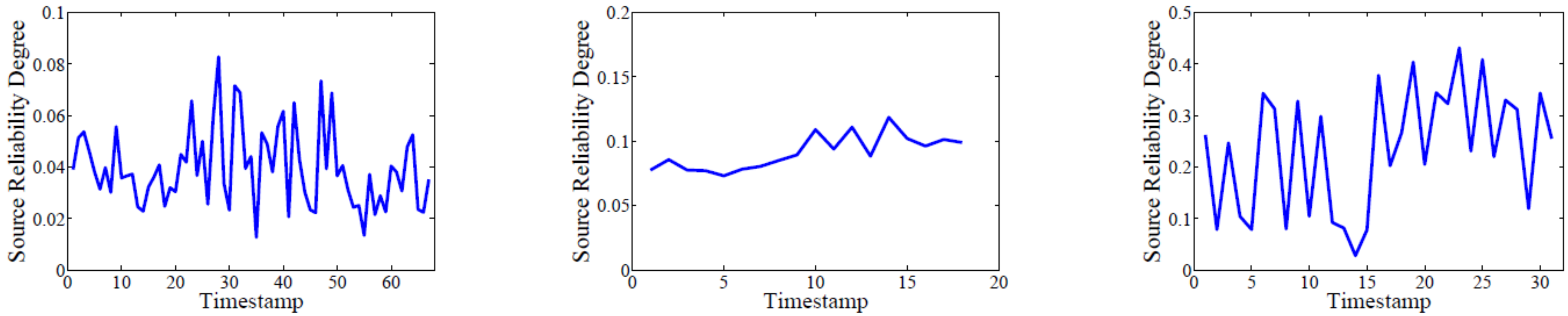
$$e_{z_q u} \uparrow \text{ and } b_q \downarrow \longrightarrow p(a_{qu} = t_q | t_q) \uparrow$$
$$e_{z_q u} \downarrow \text{ and } b_q \uparrow \longrightarrow p(a_{qu} = t_q | t_q) \downarrow$$

Fine-Grained Truth Discovery - FaitCrowd

Question level	Majority Voting	CATD	FaitCrowd
1	0.0297	0.0132	**0.0132**
2	0.0305	0.0271	**0.0271**
3	0.0414	0.0276	**0.0241**
4	0.0507	0.0290	**0.0254**
5	0.0672	0.0435	**0.0395**
6	0.1101	0.0596	**0.0550**
7	0.1016	0.0481	**0.0481**
8	0.3043	0.1304	**0.0870**
9	0.3737	0.1414	**0.1010**
10	0.5227	0.2045	**0.1136**

Real Time Truth Discovery - DynaTD

- Source reliability evolves over time



(a) Weather (b) Stock (c) Flight

- Update source reliability based on continuously arriving data:

$$p(w_s|e_{1:T}^{s}) \propto p(e_T^s|w_s)p(w_s|e_{1:T-1}^{s})$$

[Li et al., KDD'15]

Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. **Veracity Analysis from Features of Sources and Claims**
5. Applications
6. Open Questions and Resources
7. References

Veracity Analysis from Features of Sources and Claims

- Rumor detection
  - Find the rumor
  - Find the source of the rumor
- Source trustworthiness analysis
  - Graph based model
  - Learning based model

Rumor Detection on Twitter

- Clues for Detecting Rumors
  - Burst
  - High retweet ratio
  - Clue words

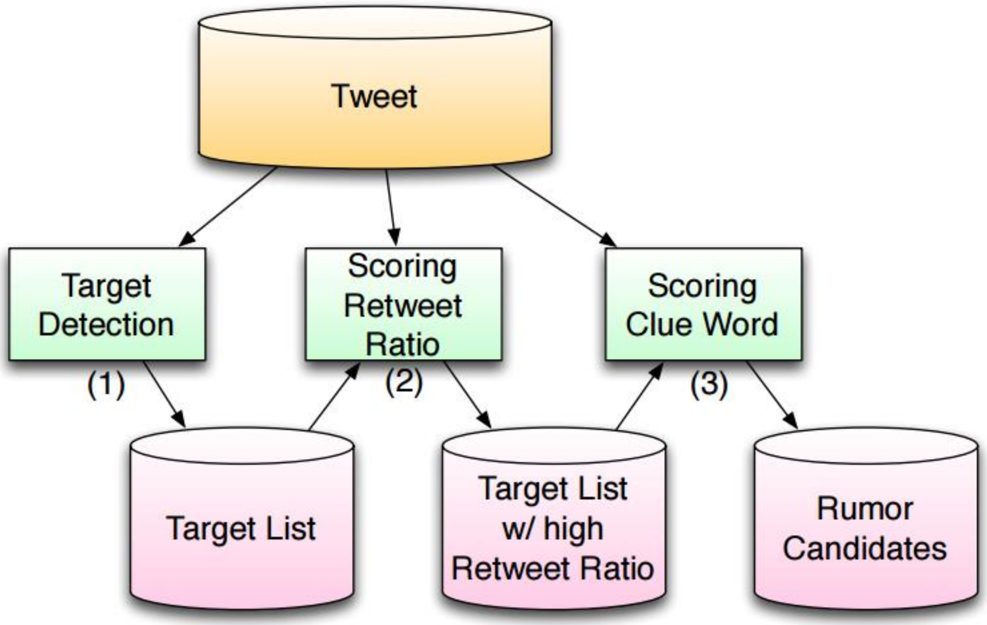


Fig. 7. Diagram of process flow

[Takahashi&Igata, SCIS'12]

Rumor Detection – Find the Rumor

- Content-based features
  - Lexical patterns
  - Part-of-speech patterns
- Network-based features
  - Tweeting and retweeting history
- Microblog-specific memes
  - Hashtags
  - URLs
  - Mentions

[Qazvinian et al., EMNLP’11][Ratkiewicz et al., CoRR’10]

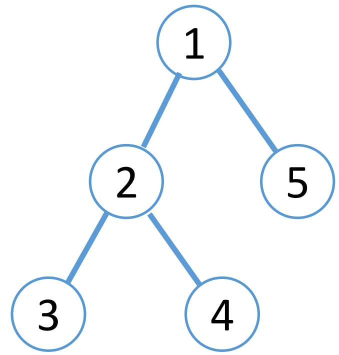
Rumor Detection on Sina Weibo

- Content-based features
  - Has multimedia, sentiment, has URL, time span
- Network-based features
  - Is retweeted, number of comments, number of retweets
- Client
  - Client program used
- Account
  - Gender of user, number of followers, user name type, …
- Location
  - Event location

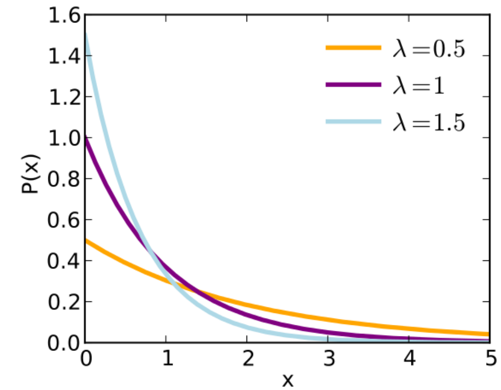
[Yang et al., MDS'12]

Rumor Detection – Find the Source

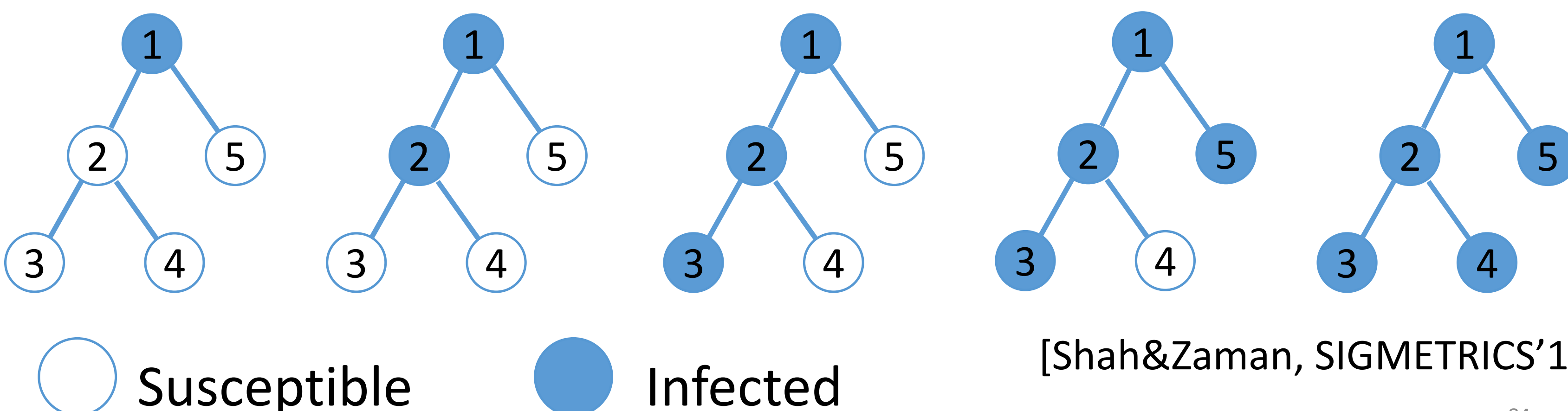
- Graph G
- If u infected, v not, and u-v, u will infect v after delay ~ exp(λ)
- Note: everyone will be infected, just a matter of time.



G



exp(λ)



[Shah&Zaman, SIGMETRICS'12]

Centrality Measures

- How "important" or central is a node u?
- Rank or measure with topological properties
  - Degree
  - Eigenvector
  - Pagerank
  - Betweenness
    - The fraction of all shortest paths that a node u is on
  - Closeness
    - Average of shortest distances from u to other nodes
    - Equal to rumor centrality for trees

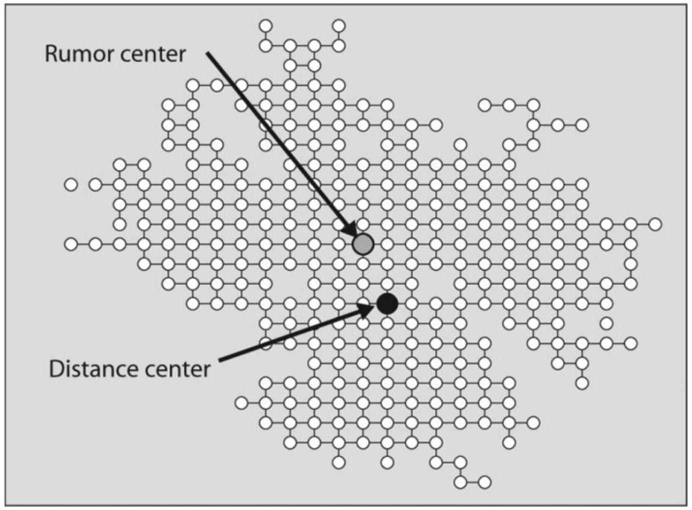
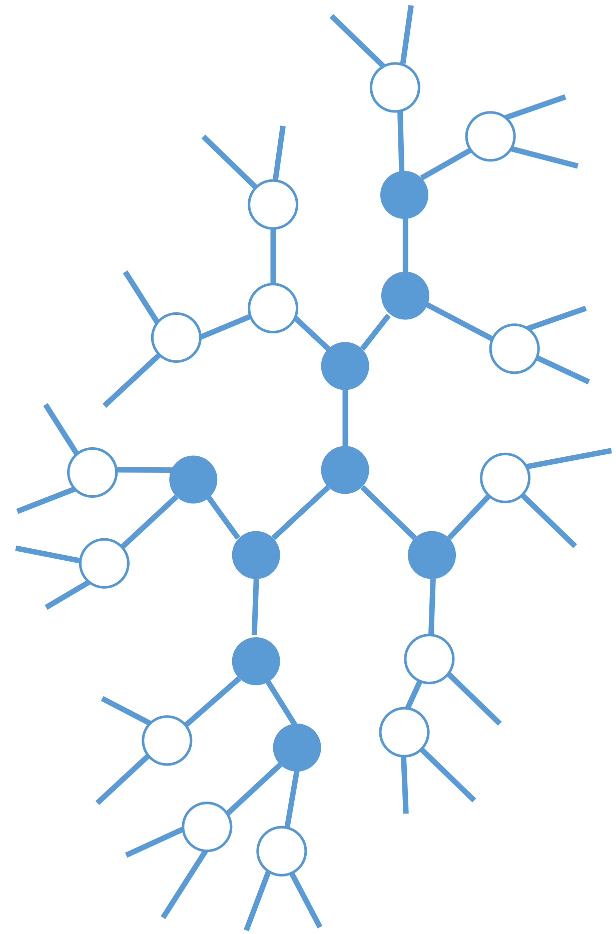


Fig. 7. A network where the distance center does not equal the general graph rumor center.

Rumor Source Detection – Rumor Centrality

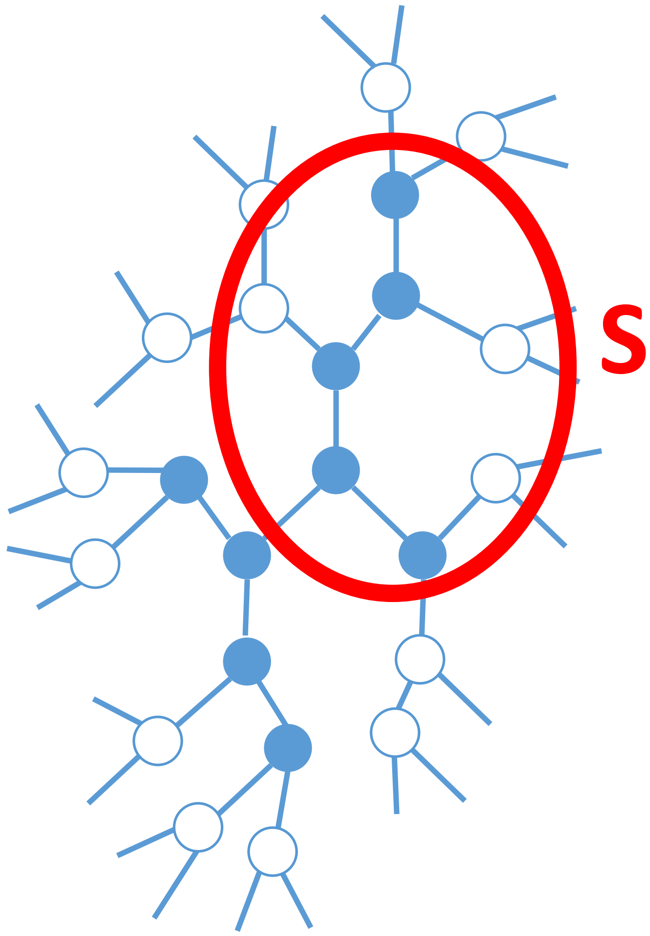
- Known infinite regular tree G, degree d > 1
- exp(λ) transmission times
  - Each edge has iid random draw
  - Value is the same for either direction
- At an unknown time t, you observe the state of the network.
- Which node was the source of the infection?
- Idea: Compute rumor centrality for each node in infected subgraph; take highest ranking node



Graph G at time t

Rumor Source Detection – Rumor Suspects

- Here you also have an a priori set of suspects S
- Which suspect was the source of the infection?
- Idea: Compute rumor centrality like before, but take highest ranking node in S



[Dong et al., ISIT'13]

Graph G at time t

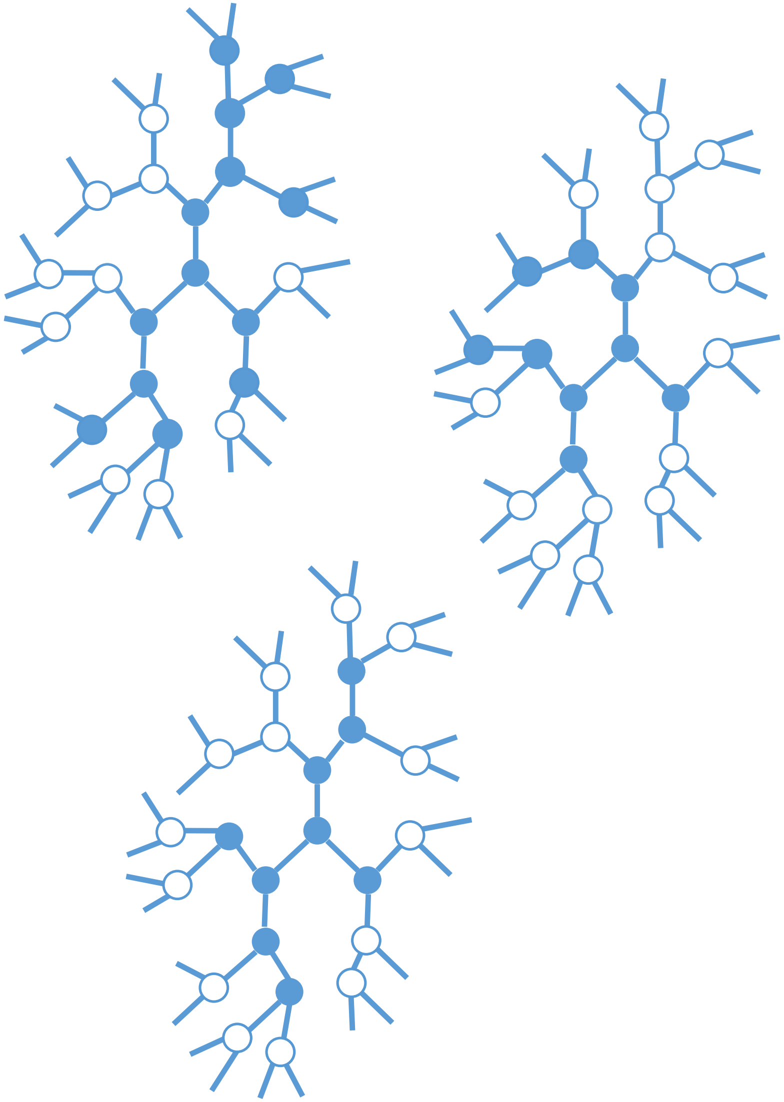
Rumor Source Detection – Multiple Observations

- Here you have multiple observations of independent rumor spreads, with the ***same*** source.

- Idea: Compute rumor centrality for each graph, take product

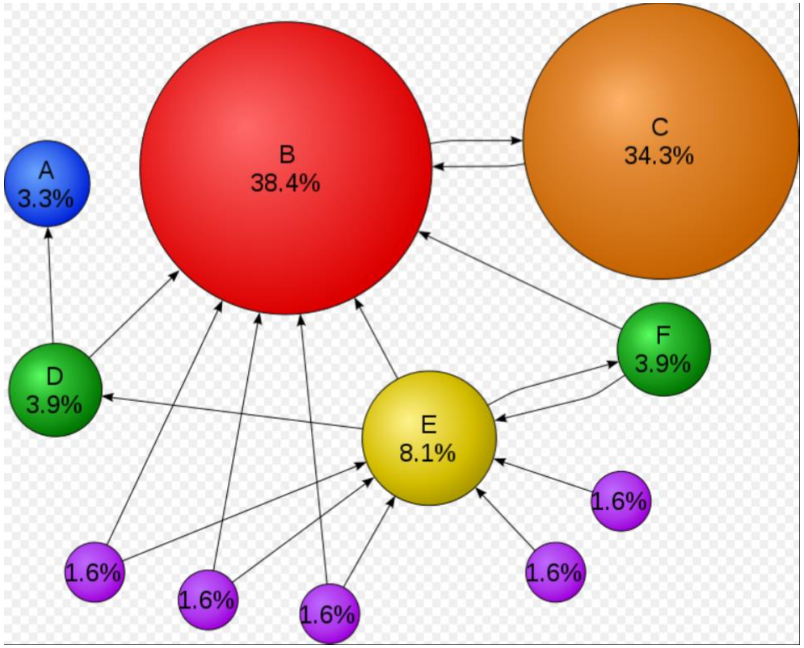
$$\widehat{s} := \arg\max_{s \in \cap_{i=1}^{m} G_i} \prod_{i=1}^{m} R(s, G_i)$$

[Wang et al., SIGMETRICS'14]

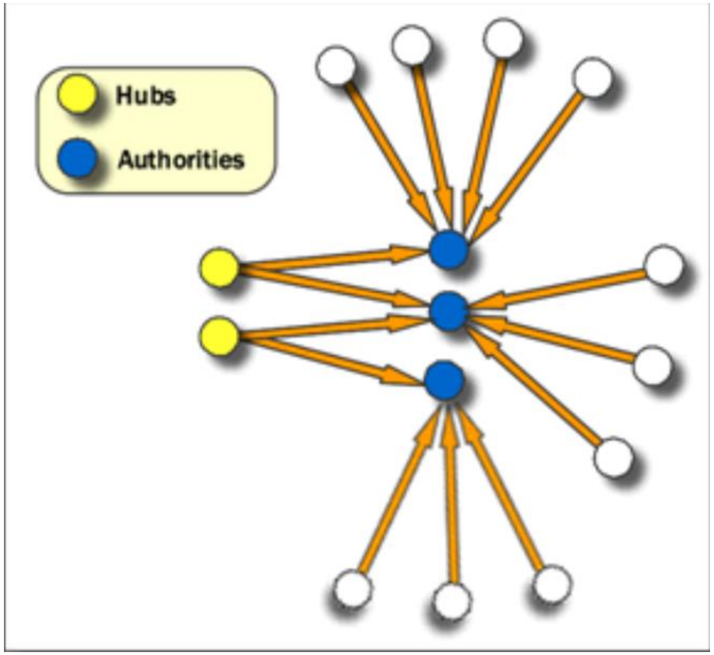


Source Trustworthiness – Graph-Based

- Intuition
  - A page has a high trustworthiness if its backlinks are trustworthy
- Only use source linkage



[Page et al., 1999]



[Kleinberg, JACM'99]

Source Trustworthiness – EigenTrust

- Problem in P2P:
  - Inauthentic files distributed by malicious nodes
- Objective:
  - Identify the source of inauthentic files and bias against downloading from them
- Basic Idea
  - Each peer has a *Global Reputation* given by the local trust values assigned by other peers

[Kamvar et al., WWW'03]

Source Trustworthiness – EigenTrust

- Local trust value c_{ij}
  - The opinion peer i has of peer j, based on past experiences
  - Each time peer i downloads an authentic/inauthentic file from peer j, c_{ij} increases/decreases.
- Global trust value t_i
  - The trust that the entire system places in peer i

What their opinion of peer k

$$t_{ik} = \sum_j c_{ij} c_{jk}$$

Ask friend j

Weight your friend's opinion by how much you trust them

Source Trustworthiness – Learning-Based

- Trust prediction: classification problem
  - Trust: positive class
  - Not trust: negative class
- Features
  - Extracted from sources to represent pairs of users

Source Trustworthiness – User Pair Trust

- Developed extensive list of possible predictive variables for trust between users
  - User factors
  - Interaction factors
- Epinions
  - Write reviews
  - Rate reviews
  - Post comments
- Used several ML tools
  - Decision tree
  - Naïve Bayes
  - SVM
  - Logistic regression
- Interaction factors are important to predict trust

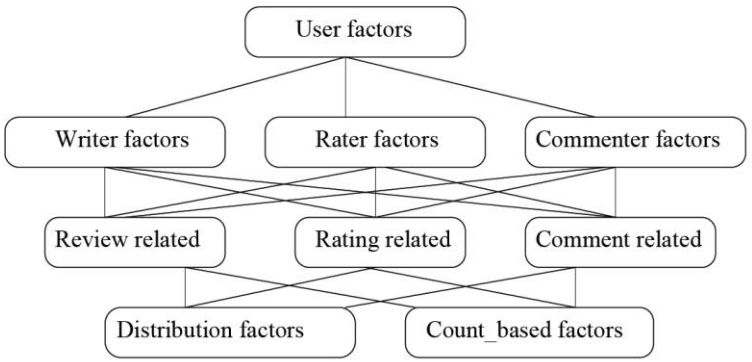


Figure 2: An taxonomy of user factors

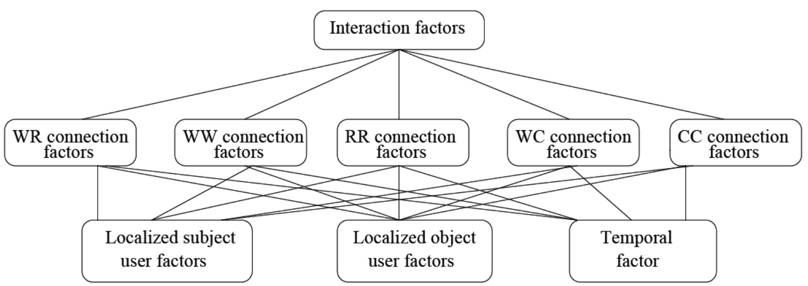


Figure 3: An taxonomy of interaction factors

[Liu et al., EC'08]

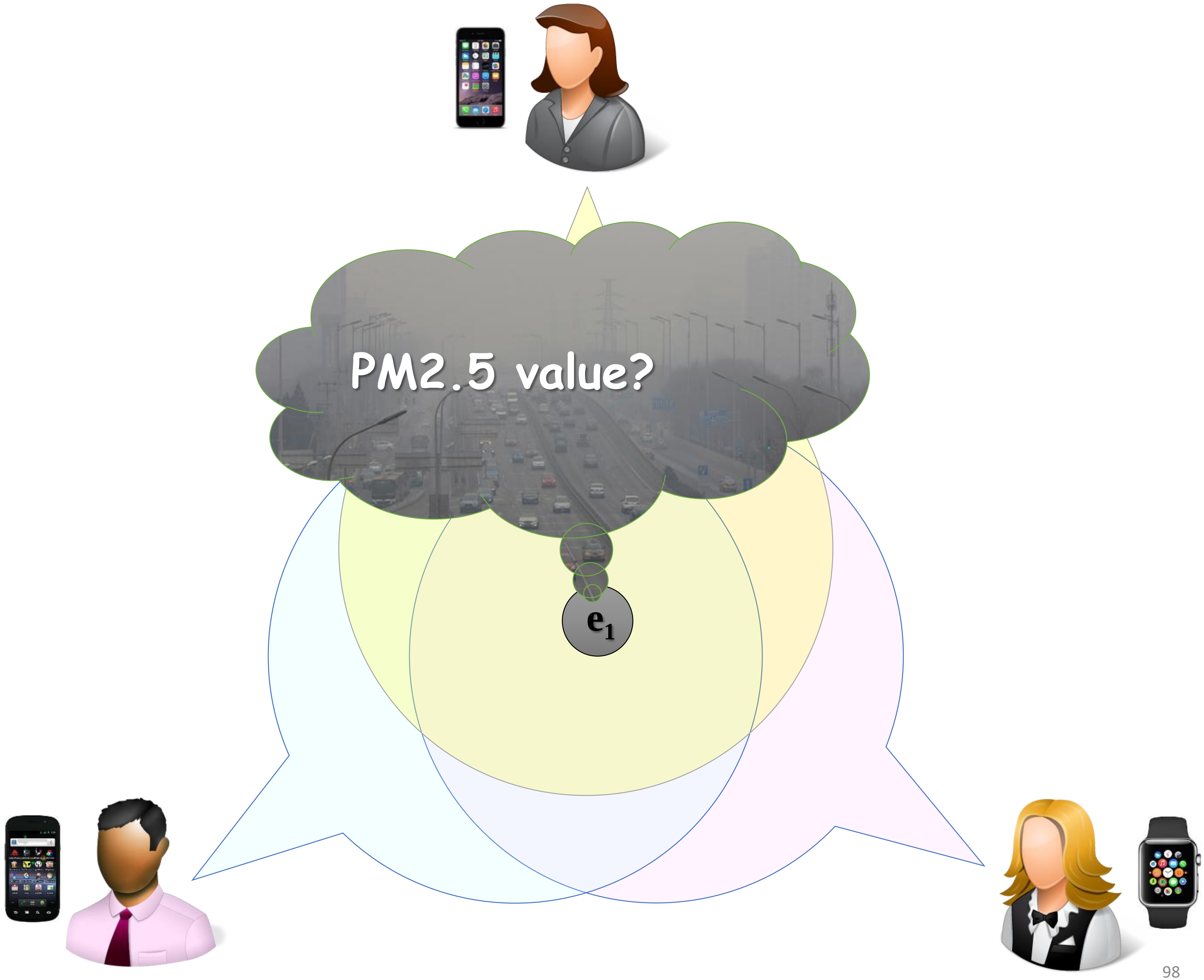
Overview

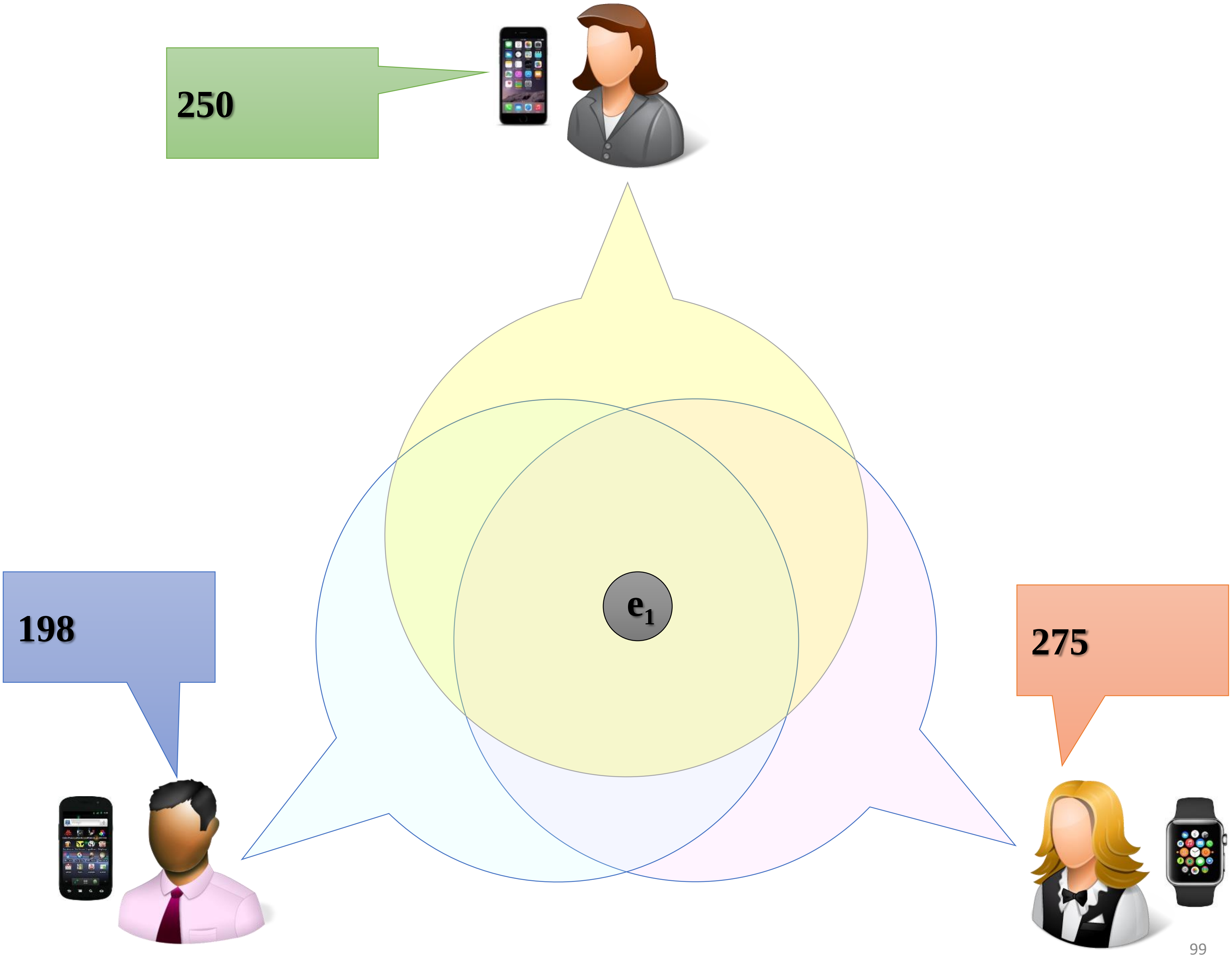
1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. **Applications**
6. Open Questions and Resources
7. References

Applications

- Knowledge base construction
  - Slot filling
- Social media data analysis
  - Rumor/fraud detection, rumor propagation
  - Claim aggregation
- Mobile sensing
  - Environmental monitoring
- Wisdom of the crowd
  - Community question answering systems

Mobile Sensing


PM2.5 value?
e_1


250
198
275
e1

Health-Oriented Community Question Answering Systems

Quality of Question-Answer Thread

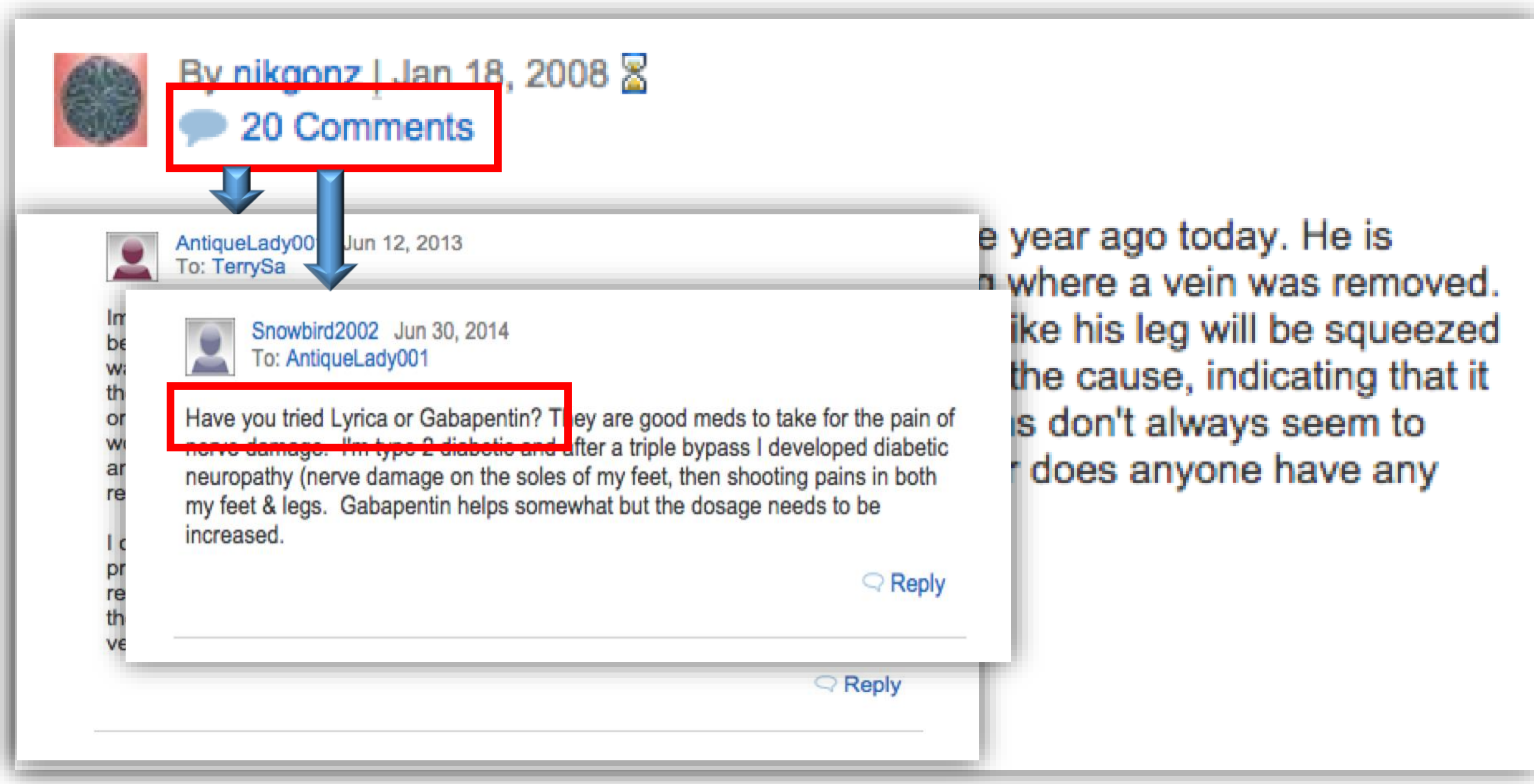
By nikgonz | Jan 18, 2008
20 Comments

My husband had quintuple heart bypass surgery one year ago today. He is experiencing increasing amount of pain in his left leg where a vein was removed. He describes it as a squeezing pain below the knee; like his leg will be squeezed in half at times. Doctors don't seem to be able to find the cause, indicating that it may be nerve damage. Pain management medications don't always seem to ease the pain either. Anyone else experiencing this or does anyone have any insight?
Thanks!

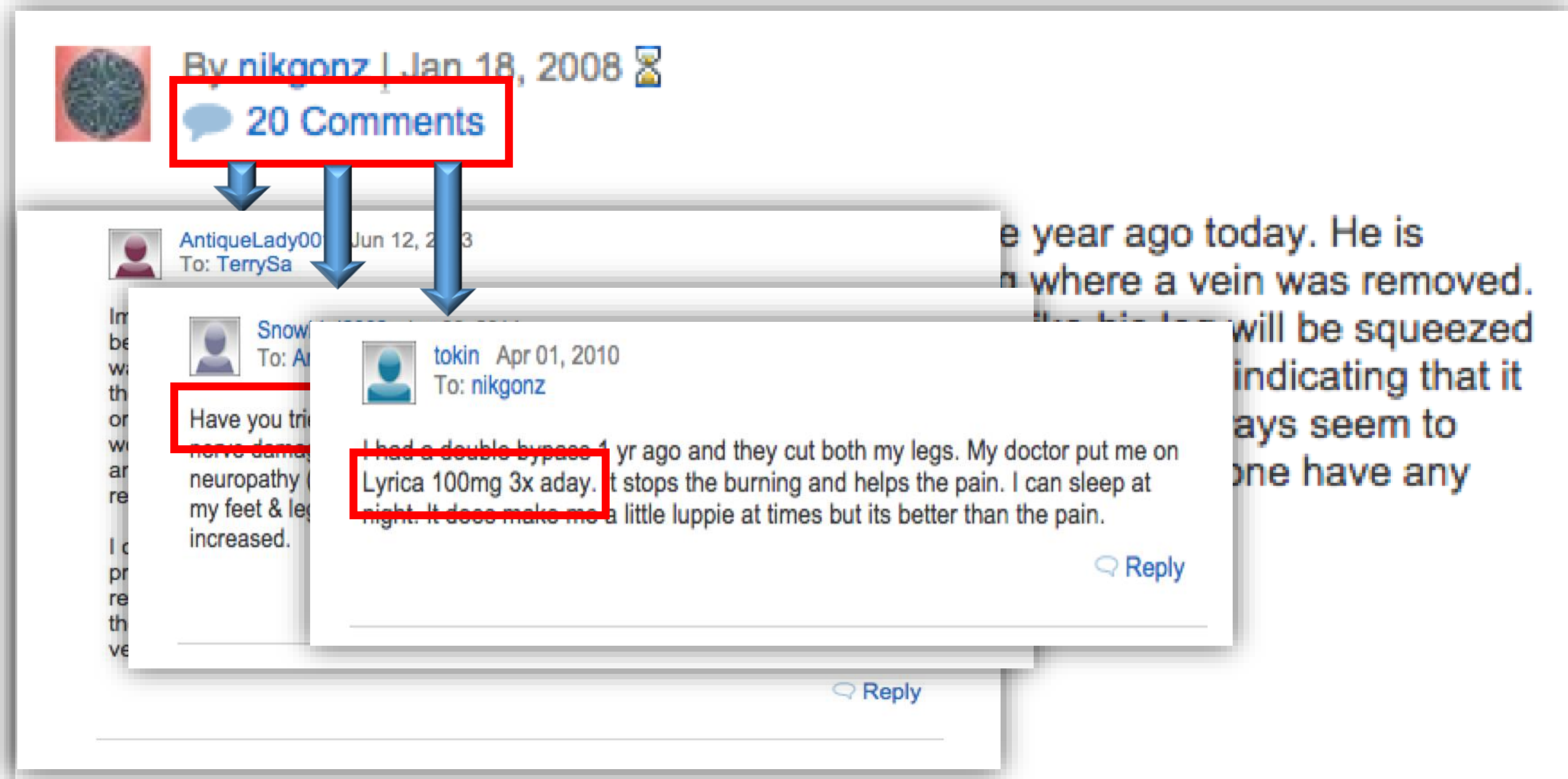
Tags: leg pain, Heart surgery

Quality of Question-Answer Thread

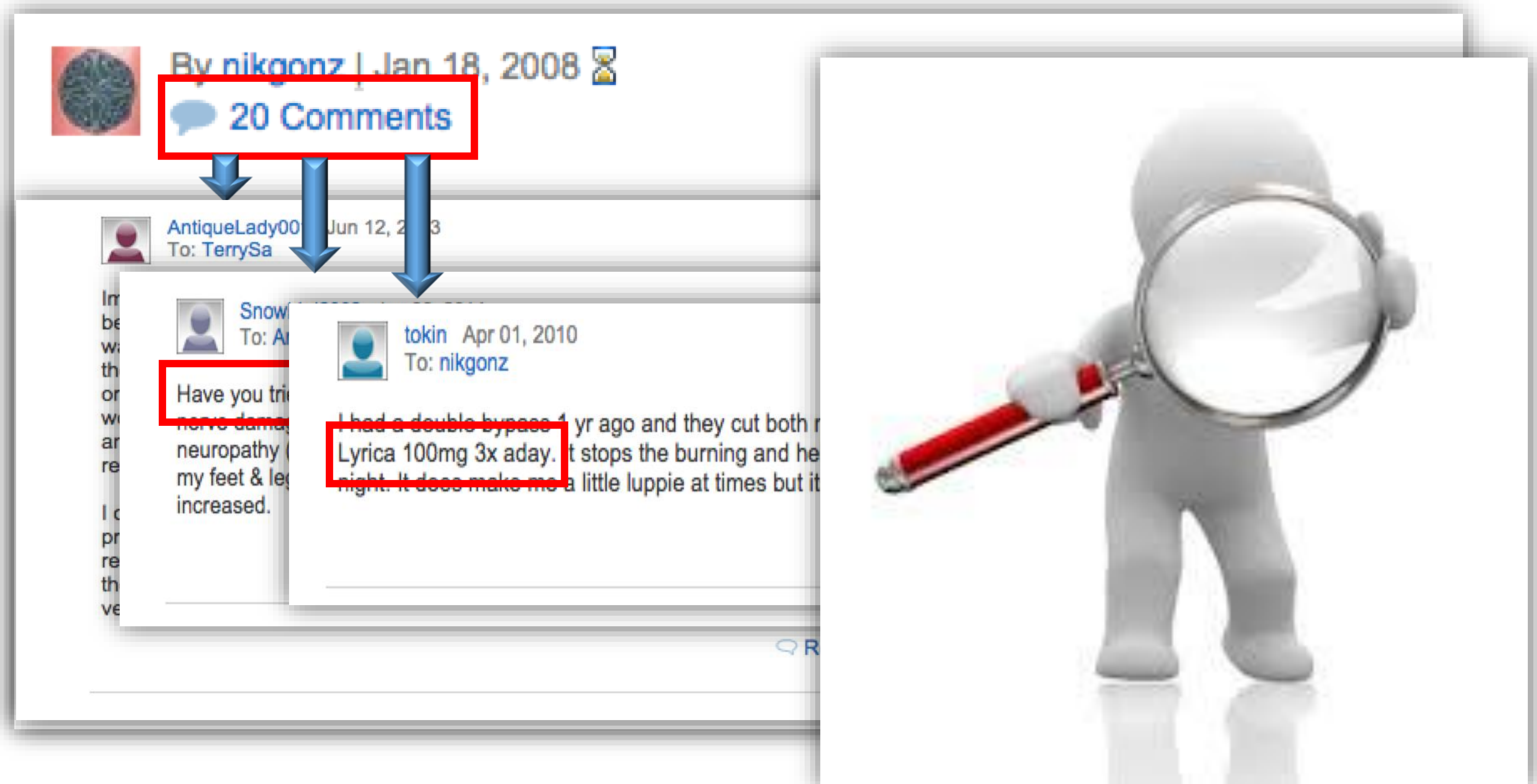
Quality of Question-Answer Thread



Quality of Question-Answer Thread



Quality of Question-Answer Thread



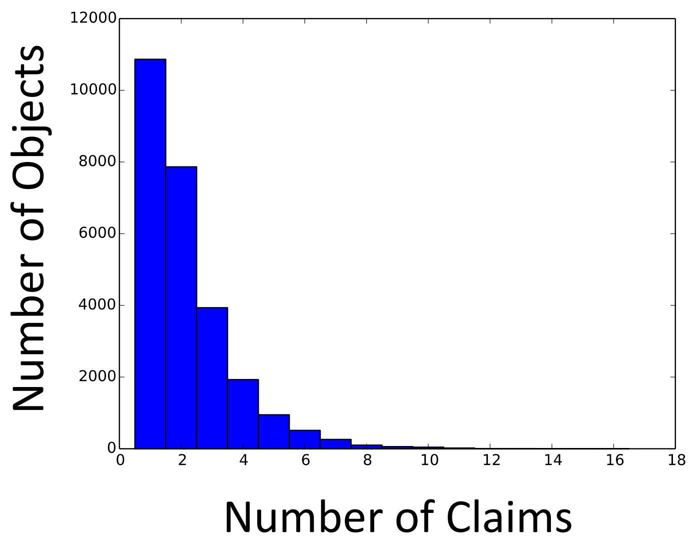
Quality of Question-Answer Thread

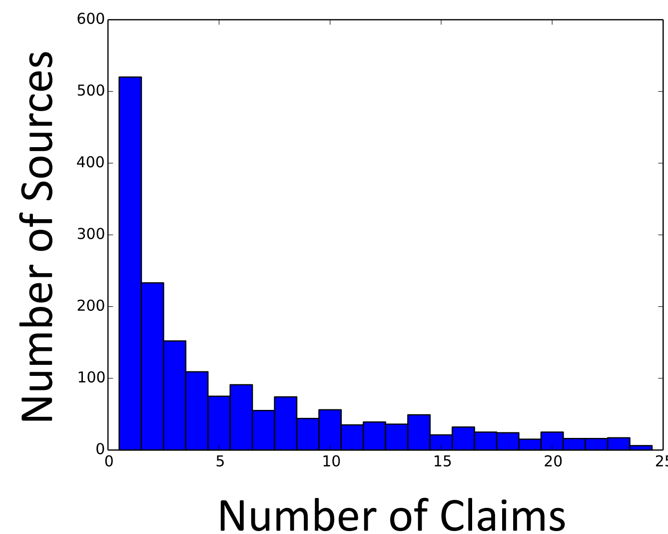
Challenge (1): Noisy Input

- Raw textual data, unstructured
- Error introduced by extractor

Challenge (2): Long-tail Phenomenon

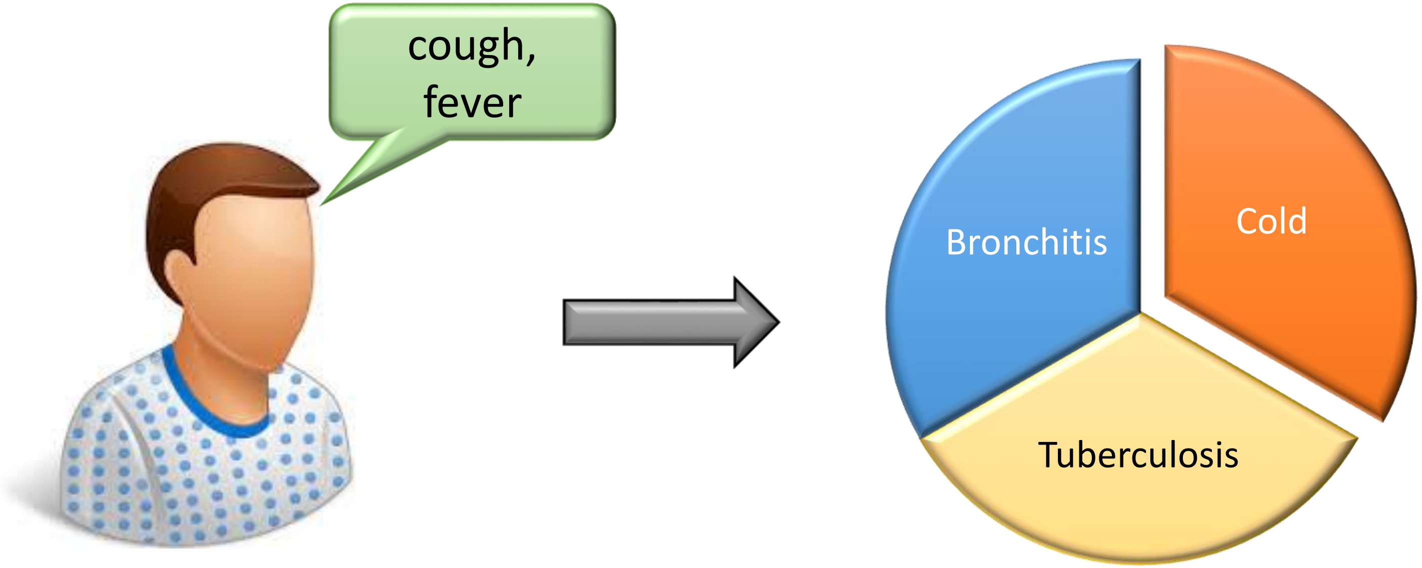
- Long-tail on both object and source sides
  - Most questions have few answers





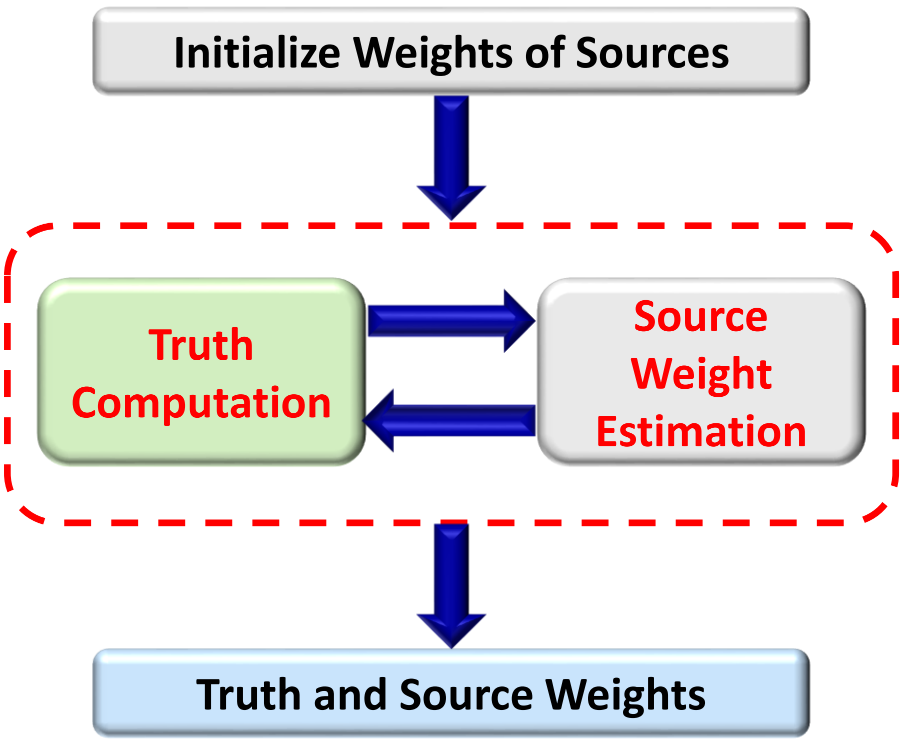
Challenge (3): Multiple Linked Truths

- Truths can be multiple, and they are correlated with each other



Challenge (4): Efficiency Issue

- Truth Discovery
  - iterative procedure

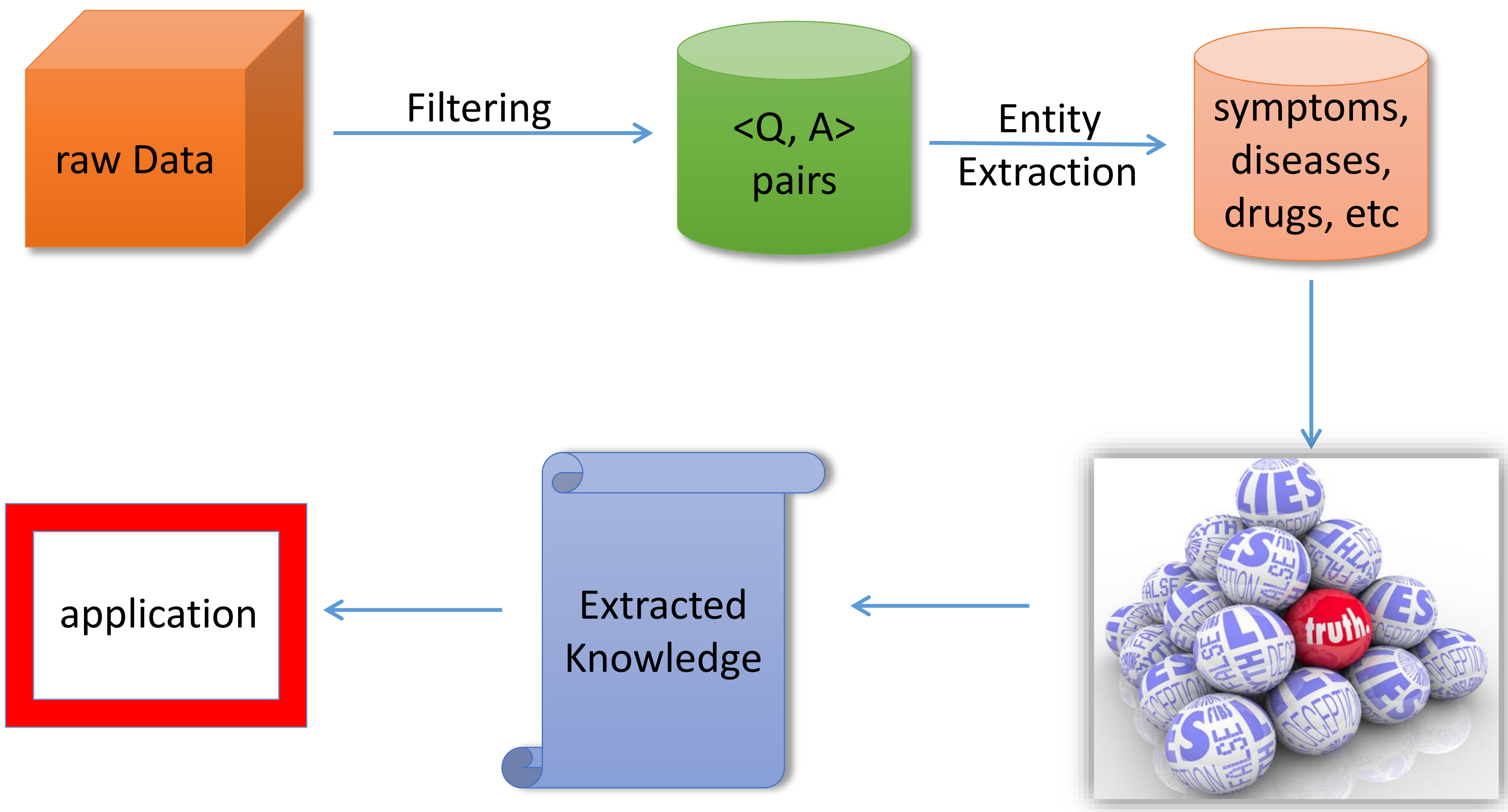


- Medical QA
  - large-scale data

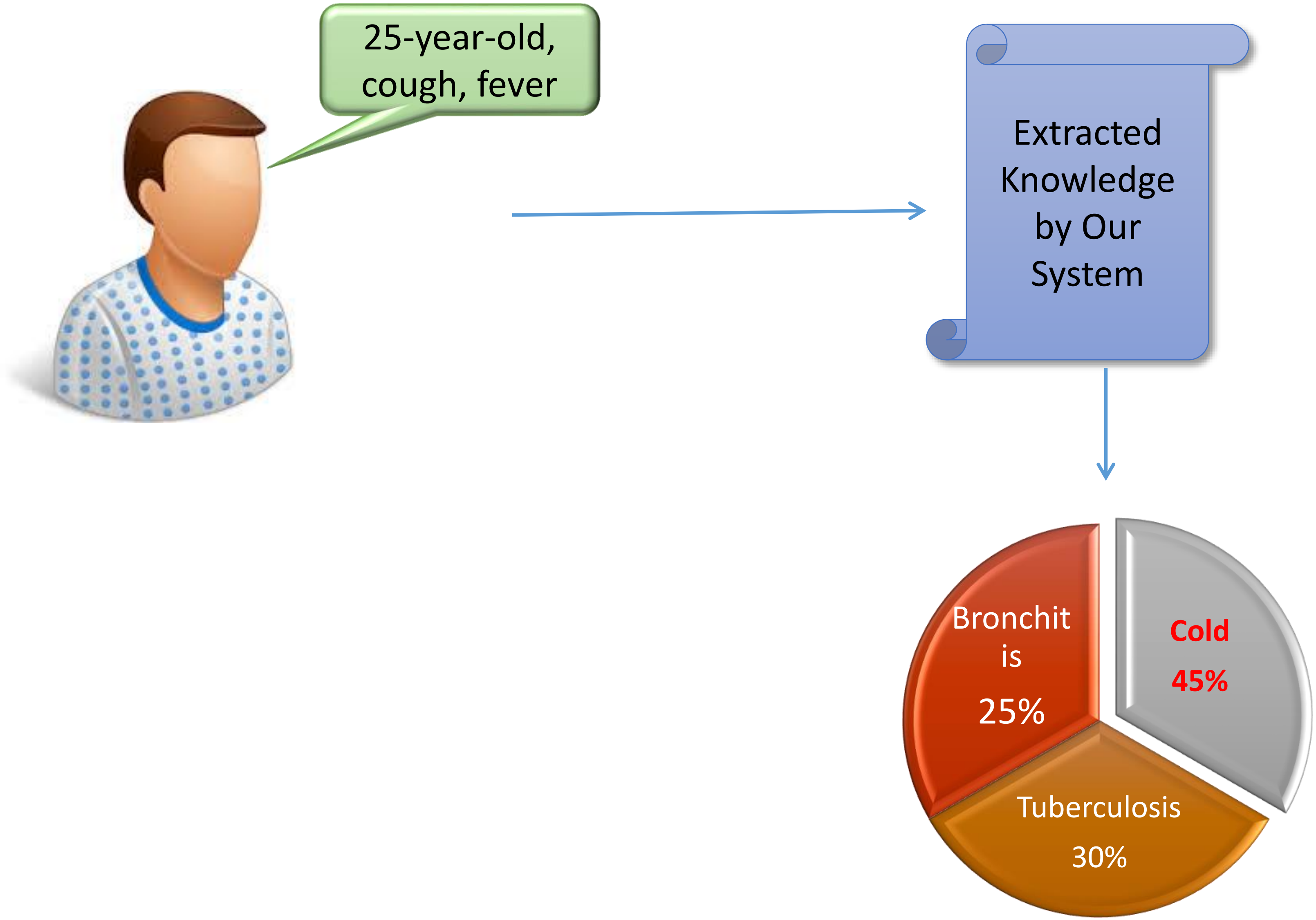
One Chinese Medical Q&A forum:

- millions of registered patients
- hundreds of thousands of doctors
- thousands of new questions per day

Overview of Our System

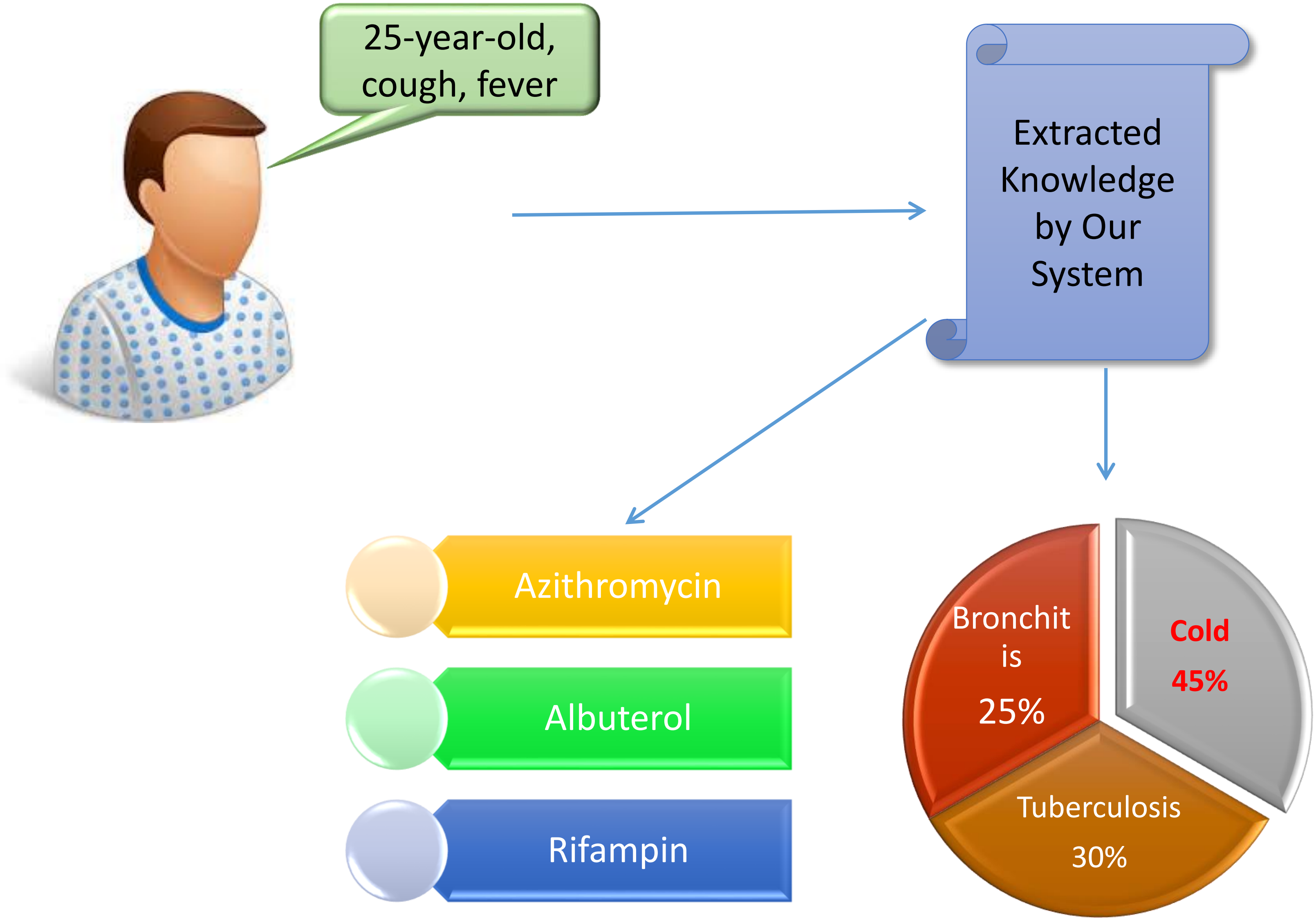


Q&A System



25-year-old, cough, fever
Extracted Knowledge by Our System
Bronchitis 25%
Cold 45%
Tuberculosis 30%

Q&A System



25-year-old, cough, fever
Extracted Knowledge by Our System
Azithromycin
Albuterol
Rifampin
Bronchitis 25%
Cold 45%
Tuberculosis 30%

Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. **Open Questions and Resources**
7. References

Open Questions

- Data with complex types and structures
- Theoretical analysis
- Efficiency of veracity analysis
- Interpretation and evaluation
- Application-specific challenges

Available Resources

- Survey for truth discovery
  - [Gupta&Han, 2011]
  - [Li et al., VLDB'12]
  - [Waguih et al., 2014]
  - [Waguih et al., ICDE'15]
  - [Li et al., 2016]
- Survey for source trustworthiness analysis
  - [Tang&Liu, WWW'14]

Available Resources

- Truth discovery data and code
  - http://lunadong.com/fusionDataSets.htm
  - http://cogcomp.cs.illinois.edu/page/resource_view/16
  - http://www.cse.buffalo.edu/~jing/software.htm

- These slides are available at

http://www.cse.buffalo.edu/~jing/talks.htm

- KDD’16 Tutorial

Enabling the Discovery of Reliable Information from Passively and Actively Crowdsourced Data

-Budget allocation

-Privacy preservation

-Crowd sensing

-.......

References

[Li et al., VLDB'14] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving Conflicts in heterogeneous data by truth discovery and source reliability estimation. *In Proc. of the ACM SIGMOD International Conference on Management of Data,* pages 1187–1198, 2014.

[Wang et al., ToSN'14] D. Wang, L. Kaplan, and T. F. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. ACM Transactions on Sensor Networks (ToSN'14), 10(2):30, 2014.

[Pasternack&Roth, COLING'10] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proc. of the International Conference on Computational Linguistics (COLING'10),* pages 877–885, 2010.

[Galland et al., WSDM'10] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. of the ACM International Conference on Web Search and Data Mining (WSDM'10),* pages 131–140, 2010.

[Yin et al., TKDE'08] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6): 796–808, 2008.

[Zhao&Han, QDB'12] B. Zhao, and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. of the VLDB workshop on Quality in Databases (QDB'12)*, 2012.

[Zhao et al., VLDB'12] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, Feb. 2012.

[Qi et al., WWW'13] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proc. of the International Conference on World Wide Web (WWW'13),* pages 1041–1052, 2013.

[Pasternack&Roth, WWW'13] J. Pasternack and D. Roth. Latent credibility analysis. In Proc. of the International Conference on World Wide Web (WWW'13), pages 1009–1020, 2013.

[Zhi et al., KDD'15] 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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, 2015.

[Yu et al., *COLING’14*] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. of the International Conference on Computational Linguistics (COLING’14)*, 2014.

[Wang et al., IPSN’14] D.Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: An estimation-theoretic perspective. In *Proc. of the International Conference on Information Processing in Sensor Networks (IPSN'14),* pages 35–46, 2014.

[Dong et al., VLDB’09a] X. L. Dong, L. Berti-Equille,and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB,* pages 550–561, 2009.

[Dong et al., VLDB’09b] X. L. Dong, L. Berti-Equille,and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB,* pages 550–561, 2009.

[Pochampally et al., SIGMOD’14] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 433–444, 2014.

[Li et al., VLDB’12] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[Li et al., VLDB’15] 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. *PVLDB*, 8(4), 2015.

[Ma et al., KDD’15] 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 the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’15)*, 2015.

[Li et al., KDD’15] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the Discovery of Evolving Truth. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15),* 2015.

[Qazvinian et al., EMNLP’11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP’11)*, pages 1589–1599, 2011.

[Ratkiewicz et al., CoRR’10] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonc¸alves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. CoRR, 2010.

[Takahashi&Igata, SCIS'12] T. Takahashi, and N Igata. Rumor detection on twitter. *2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS),* 2012.

[Yang et al., MDS'12] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics (MDS'12)*, 2012.

[Shah&Zaman, SIGMETRICS'12] D. Shah, and T. Zaman. Rumor centrality: a universal source detector. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40, No. 1, pages 199-210, 2012.

[Dong et al., ISIT'13] W. Dong, W. Zhang, and CW. Tan. Rooting out the rumor culprit from suspects. In *Proc. Of the IEEE International Symposium on Information Theory Proceedings (ISIT'13)*, 2013.

[Wang et al., SIGMETRICS'14] Z. Wang, W. Dong, W. Zhang, and CW. Tan. Rumor source detection with multiple observations: fundamental limits and algorithms. In *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pages 1-13, 2014.

[Page et al., 1999] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[Kleinberg, JACM'99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[Kamvar et al., WWW'03] S.D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proc. of the 12th international conference on World Wide Web (WWW'03)*, 2003.

[Liu et al., EC'08] H. Liu, E.-P. Lim, H.W. Lauw, M.-T. Le, A. Sun, J. Srivastava, and Y. Kim. Predicting trusts among users of online communities: an epinions case study. In *Proc. of the 9th ACM conference on Electronic commerce*, pages 310–319, 2008.

[Mukherjee et al., KDD'14] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*, pages 65–74, 2014.

[Gupta&Han, 2011] M. Gupta and J. Han. Heterogeneous network-based trust analysis: A survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

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

[Waguih et al., ICDE'15] 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 the IEEE International Conference on Data Engineering (ICDE'15)*, 2015.

[Li et al., 2016] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), pp.1-16.

[Tang&Liu, WWW'14] J. Tang and H. Liu. Trust in social computing. In *Proc. of the Companion Publication of the International Conference on World Wide Web Companion (WWW'14)*, pages 207–208, 2014.