Veracity of Big Data

Laure Berti-Equille and Javier Borge-Holthoefer

Qatar Computing Research Institute

{lberti,jborge}@qf.org.qa

CIKM2015
19–23 October 2015 Melbourne Australia

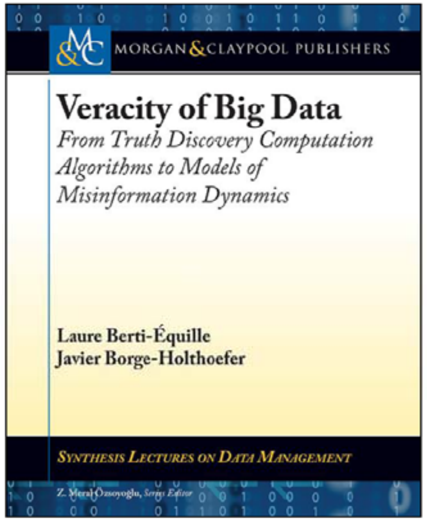
Disclaimer

Aim of the tutorial: Get the big picture

The algorithms of the basic 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/tutorial_cikm2015/index.html



MORGAN & CLAYPOOL PUBLISHERS

Veracity of Big Data

From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille
Javier Borge-Holthoefer

Synthesis Lectures on Data Management

Z. Meral Özsoyoğlu, *Series Editor*

Many **sources** of information available online

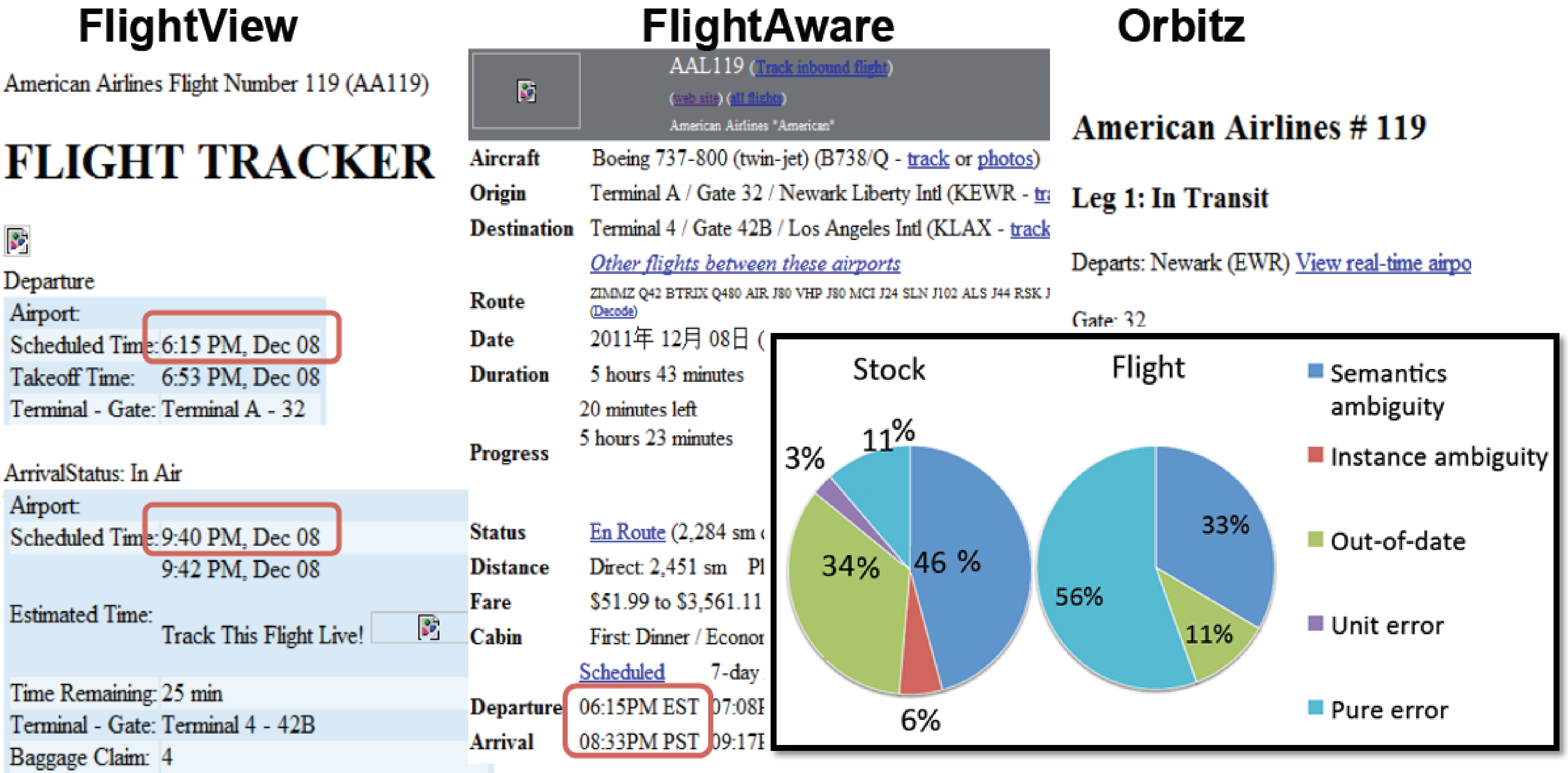
facebook
twitter
New York Times
NEW YORK, SUNDAY, JULY 7, 2013
Gaz
OF THE PEOPLE
Camp Help
a community of cancer survivors s
BLOG FOR A CURE
members
posts
61 Brain Cancer B
Find Cancer Fighters in this c
stronger. Looking for more de
Select Cancer Type: Brain Cancer Blogs
Poobare
Type of Cancer: Brain Cancer (Stage IV glioblastoma/a difference))
Cancer Symptoms: Seizures in left arm and left leg, he
Stage of Cancer: Stage 4
Surgery: First removeal was on April 13, 2011 - he was cancer or anything that would ever grow back and...
Chemotherapy: Started temodar on November 9, 201 there has not been much happen in terms of side affect
Bio: I am the Mom to four children, 2 boys and twin gi oldest step son has stage IV...
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Toolbox
Print/export
This article doc
This article has
The neutrality
This article ma
Operation Pillar of Defense (Hebrew: עמוד ענן
in the Gaza Strip from 14 to 21 November 2012. It s
The stated aims of the operation were to
disrupt the capabilities of militant organizations.
The Israeli government said the operation began in re
Israel-Gaza border. The Hamas gover
the blockade of the Gaza Strip and the occupation of
safe, your towns will not be safe."
The IDF stated it targeted more than 1,500 military si
centers, weapons manufacturing, and storage buildin
Create account
Log in
Search

Are all these sources equally

- accurate
- up-to-date
- and trustworthy?

Accurate?

Deep Web data quality is low



FlightView

American Airlines Flight Number 119 (AA119)

FLIGHT TRACKER

Departure

Airport:	
Scheduled Time:	6:15 PM, Dec 08
Takeoff Time:	6:53 PM, Dec 08
Terminal - Gate:	Terminal A - 32

ArrivalStatus: In Air

Airport:	
Scheduled Time:	9:40 PM, Dec 08
Estimated Time:	9:42 PM, Dec 08 Track This Flight Live!
Time Remaining:	25 min
Terminal - Gate:	Terminal 4 - 42B
Baggage Claim:	4

FlightAware

AAL119 (Track inbound flight)

(web site) (all flights)

American Airlines "American"

Aircraft	Boeing 737-800 (twin-jet) (B738/Q - track or photos)
Origin	Terminal A / Gate 32 / Newark Liberty Intl (KEWR - tr
Destination	Terminal 4 / Gate 42B / Los Angeles Intl (KLAX - track Other flights between these airports
Route	ZIMMZ Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J (Decode)
Date	2011年 12月 08日 (
Duration	5 hours 43 minutes
Progress	20 minutes left 5 hours 23 minutes
Status	En Route (2,284 sm (
Distance	Direct: 2,451 sm Pl
Fare	$51.99 to $3,561.11
Cabin	First: Dinner / Econor
	Scheduled 7-day
Departure	06:15PM EST 07:08I
Arrival	08:33PM PST 09:17I

Orbitz

American Airlines # 119

Leg 1: In Transit

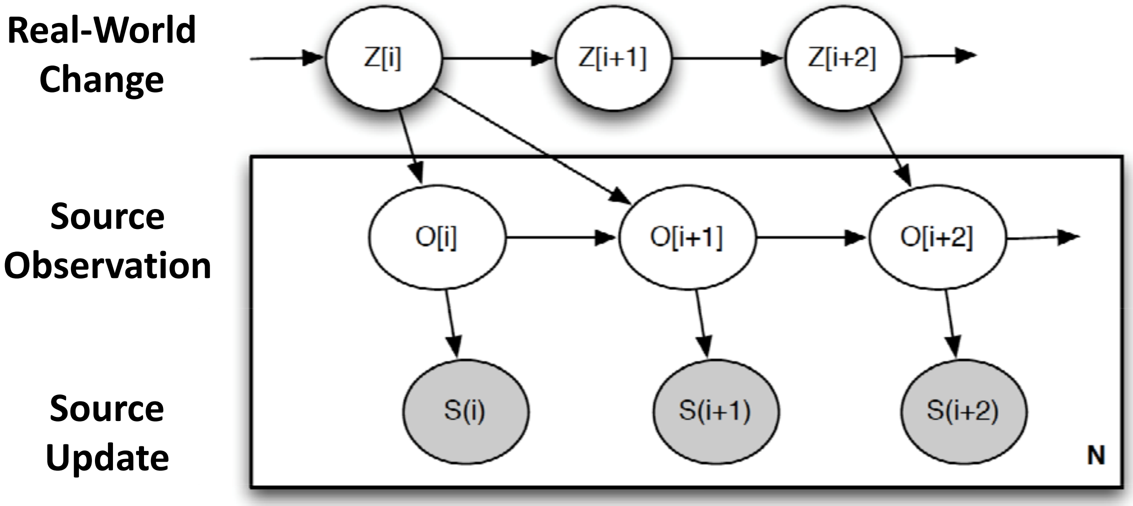
Departs: Newark (EWR) View real-time airpo

Gate: 32

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?

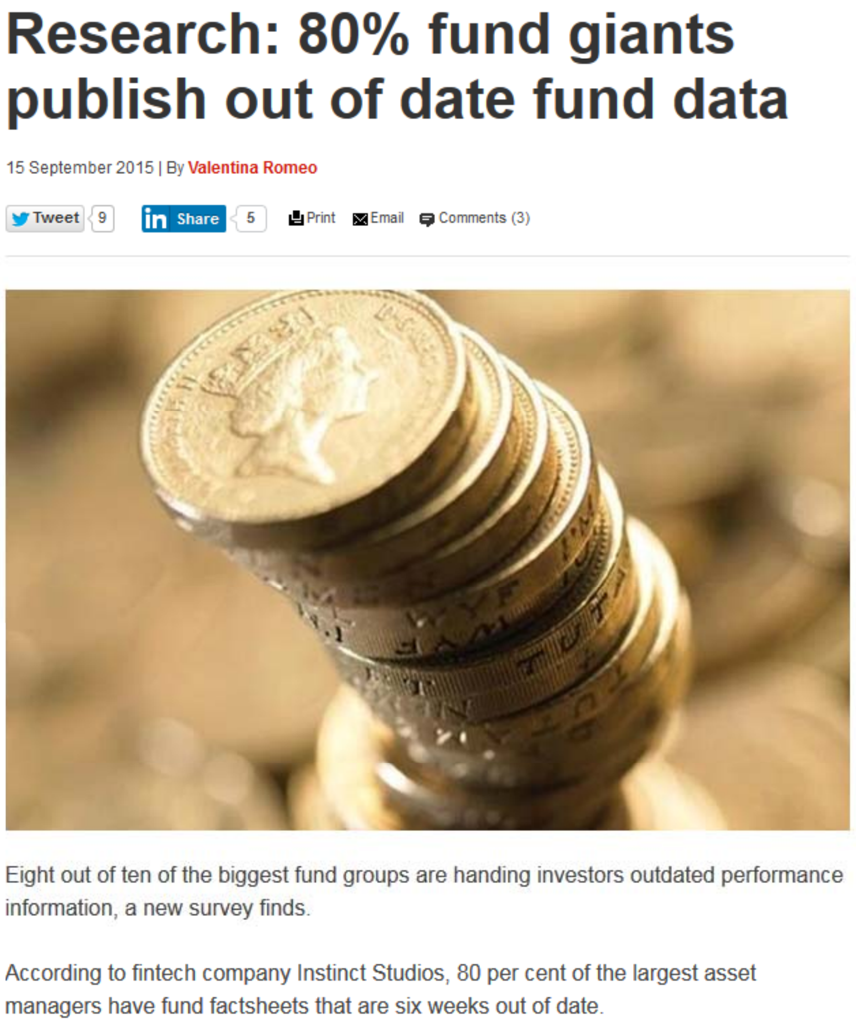
Real-world entities evolve over time, but sources can delay, or even miss, reporting some of the real-world updates.



Real-World Change

Source Observation

Source Update



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

15 September 2015 | By Valentina Romeo

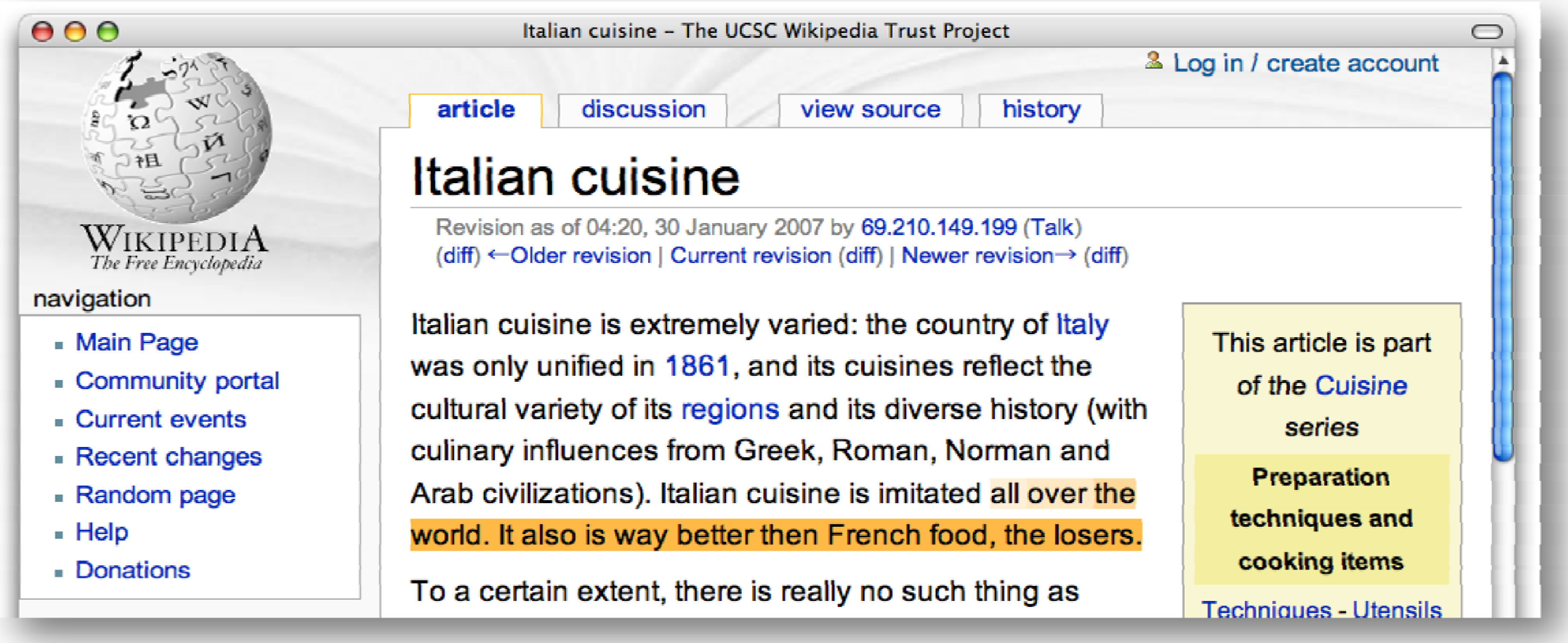
Eight out of ten of the biggest fund groups are handing investors outdated performance information, a new survey finds.

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



Italian cuisine

Revision as of 04:20, 30 January 2007 by 69.210.149.199 (Talk)
(diff) ←Older revision | Current revision (diff) | Newer revision→ (diff)

Italian cuisine is extremely varied: the country of Italy was only unified in 1861, and its cuisines reflect the cultural variety of its regions and its diverse history (with culinary influences from Greek, Roman, Norman and Arab civilizations). Italian cuisine is imitated all over the world. It also is way better then French food, the losers.

To a certain extent, there is really no such thing as

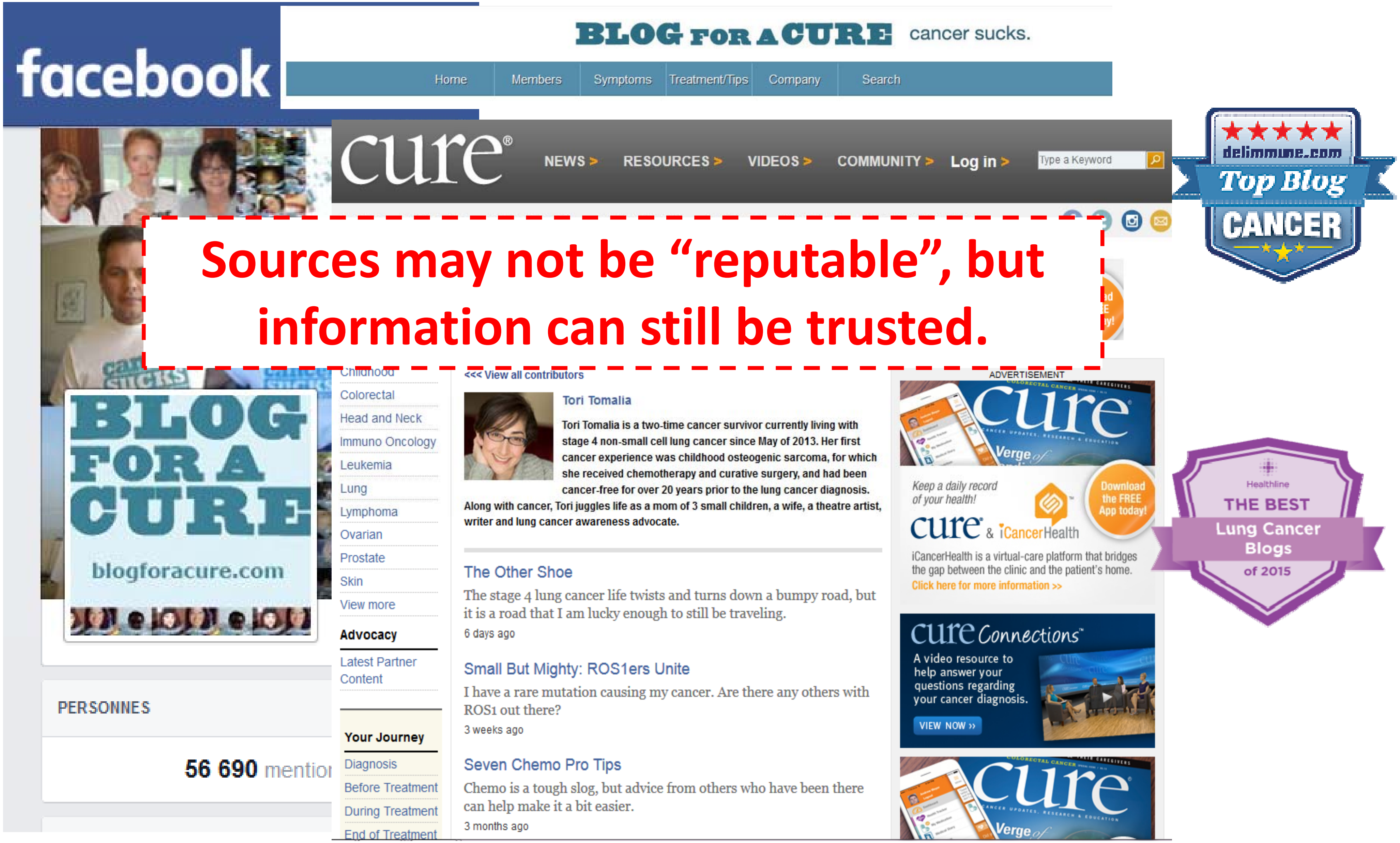
This article is part of the Cuisine series

Preparation techniques and cooking items

Techniques - Utensils

- *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.

<<< View all contributors

Tori Tomalia

Tori Tomalia is a two-time cancer survivor currently living with stage 4 non-small cell lung cancer since May of 2013. Her first cancer experience was childhood osteogenic sarcoma, for which she received chemotherapy and curative surgery, and had been cancer-free for over 20 years prior to the lung cancer diagnosis. Along with cancer, Tori juggles life as a mom of 3 small children, a wife, a theatre artist, writer and lung cancer awareness advocate.

The Other Shoe

The stage 4 lung cancer life twists and turns down a bumpy road, but it is a road that I am lucky enough to still be traveling.

6 days ago

Small But Mighty: ROS1ers Unite

I have a rare mutation causing my cancer. Are there any others with ROS1 out there?

3 weeks ago

Seven Chemo Pro Tips

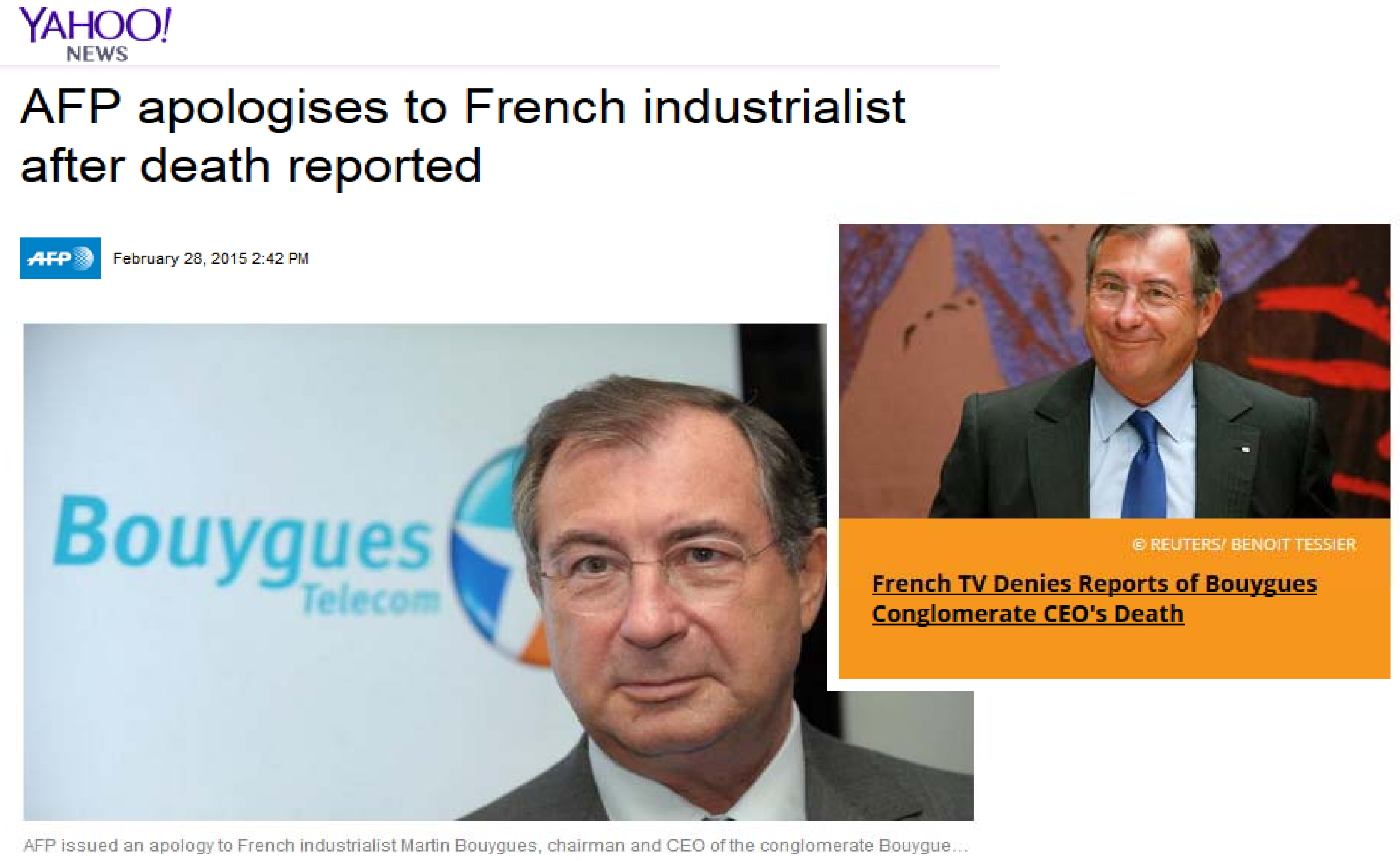
Chemo is a tough slog, but advice from others who have been there can help make it a bit easier.

3 months ago

iCancerHealth is a virtual-care platform that bridges the gap between the clinic and the patient's home.

A video resource to help answer your questions regarding your cancer diagnosis.

Authoritative sources can be wrong



YAHOO! NEWS

AFP apologises to French industrialist after death reported

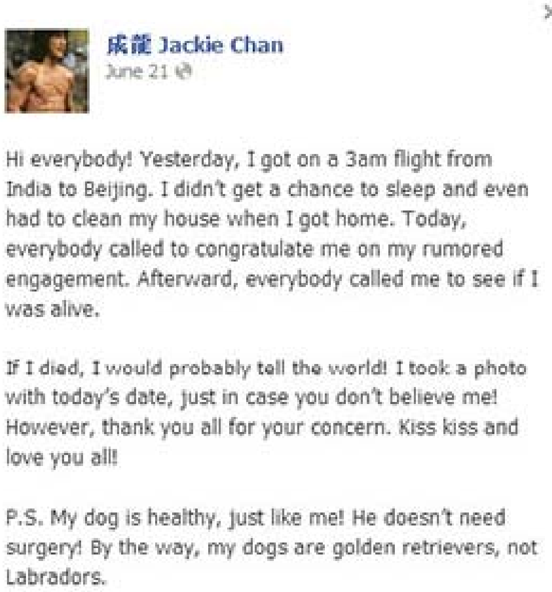
AFP February 28, 2015 2:42 PM

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

© REUTERS/ BENOIT TESSIER

French TV Denies Reports of Bouygues Conglomerate CEO's Death

Rumors: Celebrity Death Hoaxes

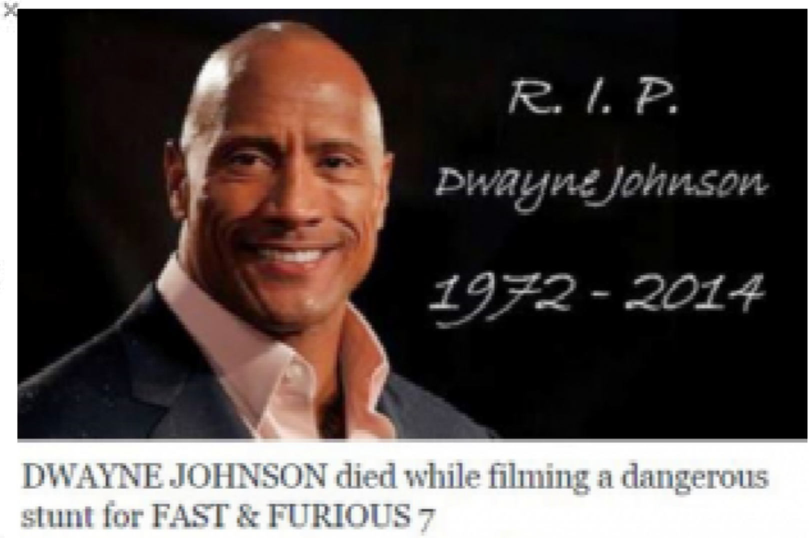


成龍 Jackie Chan
June 21

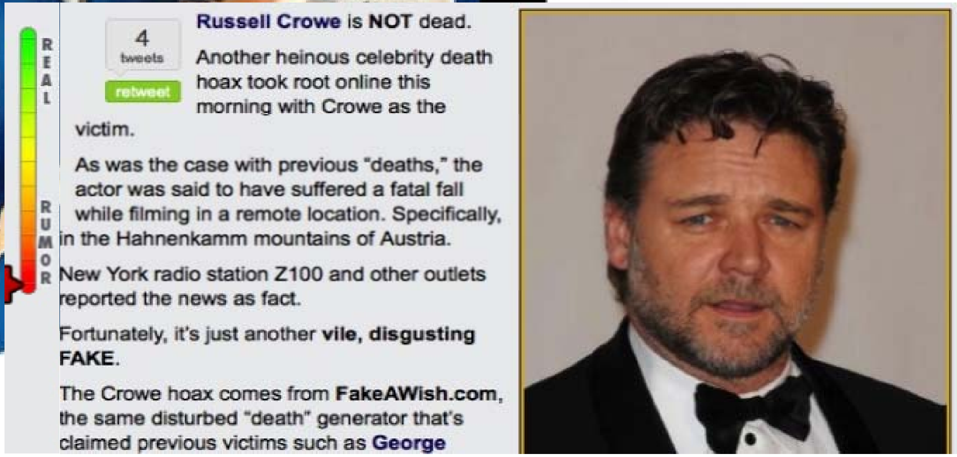
Hi everybody! Yesterday, I got on a 3am flight from India to Beijing. I didn't get a chance to sleep and even had to clean my house when I got home. Today, everybody called to congratulate me on my rumored engagement. Afterward, everybody called me to see if I was alive.

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



Russell Crowe is **NOT** dead.

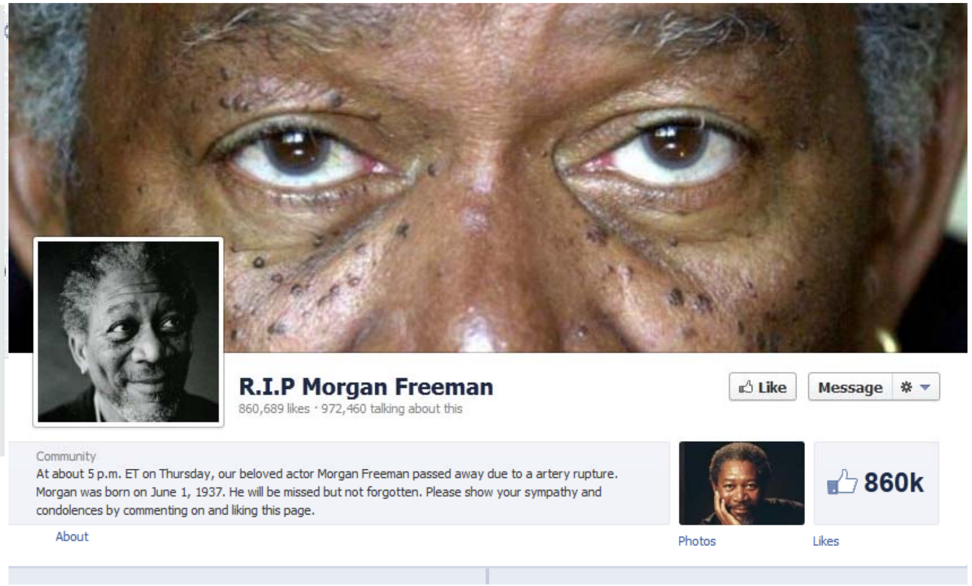
Another heinous celebrity death hoax took root online this morning with Crowe as the victim.

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**



R.I.P Morgan Freeman
860,689 likes · 972,460 talking about this

Community
At about 5 p.m. ET on Thursday, our beloved actor Morgan Freeman passed away due to a artery rupture. Morgan was born on June 1, 1937. He will be missed but not forgotten. Please show your sympathy and condolences by commenting on and liking this page.

(Manual) Fact Verification Web Sites (I)

PPRuNe
Professional Pilots
Rumour Network

FactCheck.org

TruthOrFiction.com

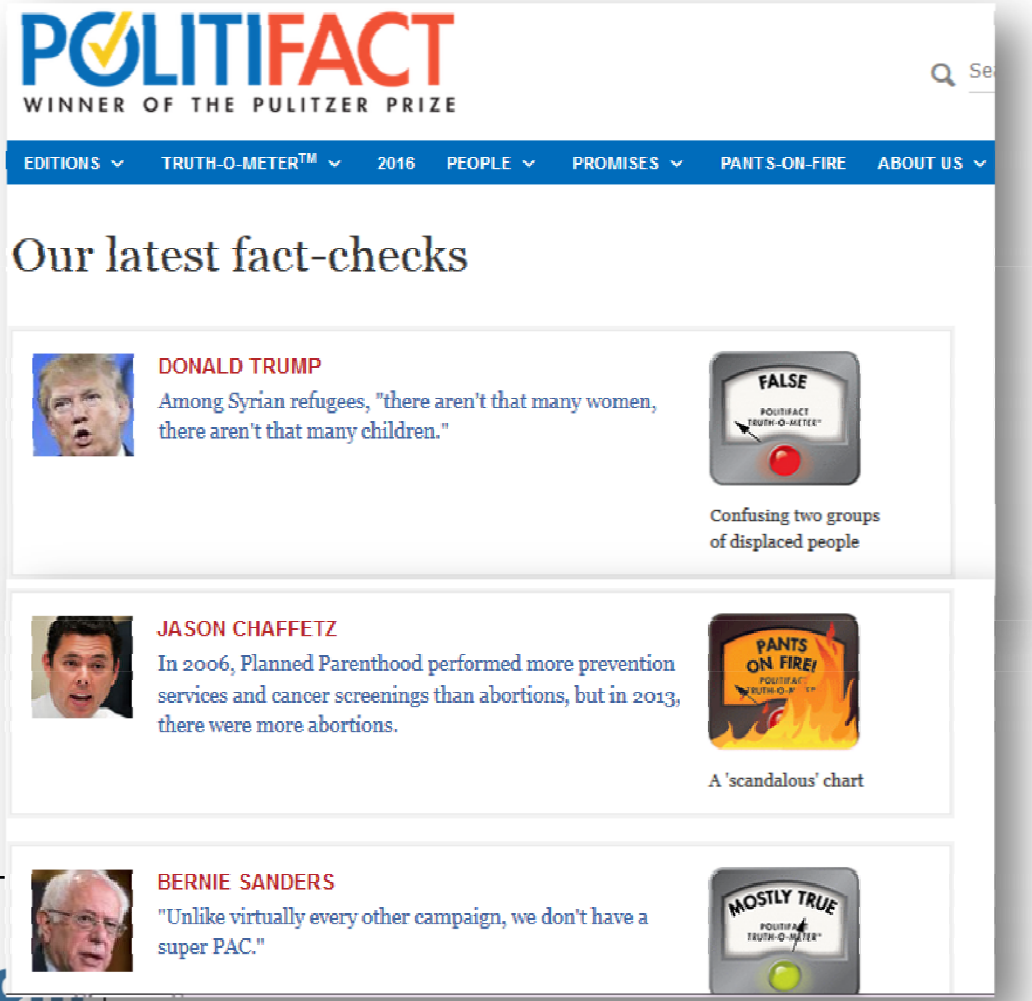
Snopes.com
Rumor Has It

EL POLÍGRAFO
Chequeo de datos | La política contrastada

OpenSecrets.org
Center for Responsive Politics

CHEQUEADO•COM
La verificación del discurso público

UYCheck.co
¿nos dicen la verdad?



POLITIFACT
WINNER OF THE PULITZER PRIZE

Our latest fact-checks

DONALD TRUMP
Among Syrian refugees, "there aren't that many women, there aren't that many children."
FALSE
Confusing two groups of displaced people

JASON CHAFFETZ
In 2006, Planned Parenthood performed more prevention services and cancer screenings than abortions, but in 2013, there were more abortions.
PANTS ON FIRE!
A 'scandalous' chart

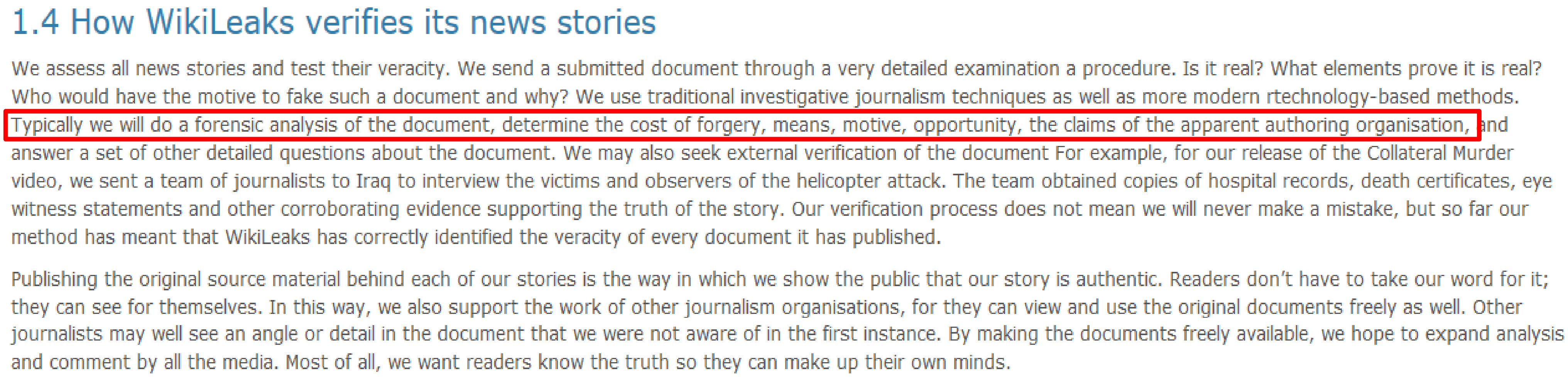
BERNIE SANDERS
"Unlike virtually every other campaign, we don't have a super PAC."
MOSTLY TRUE

(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



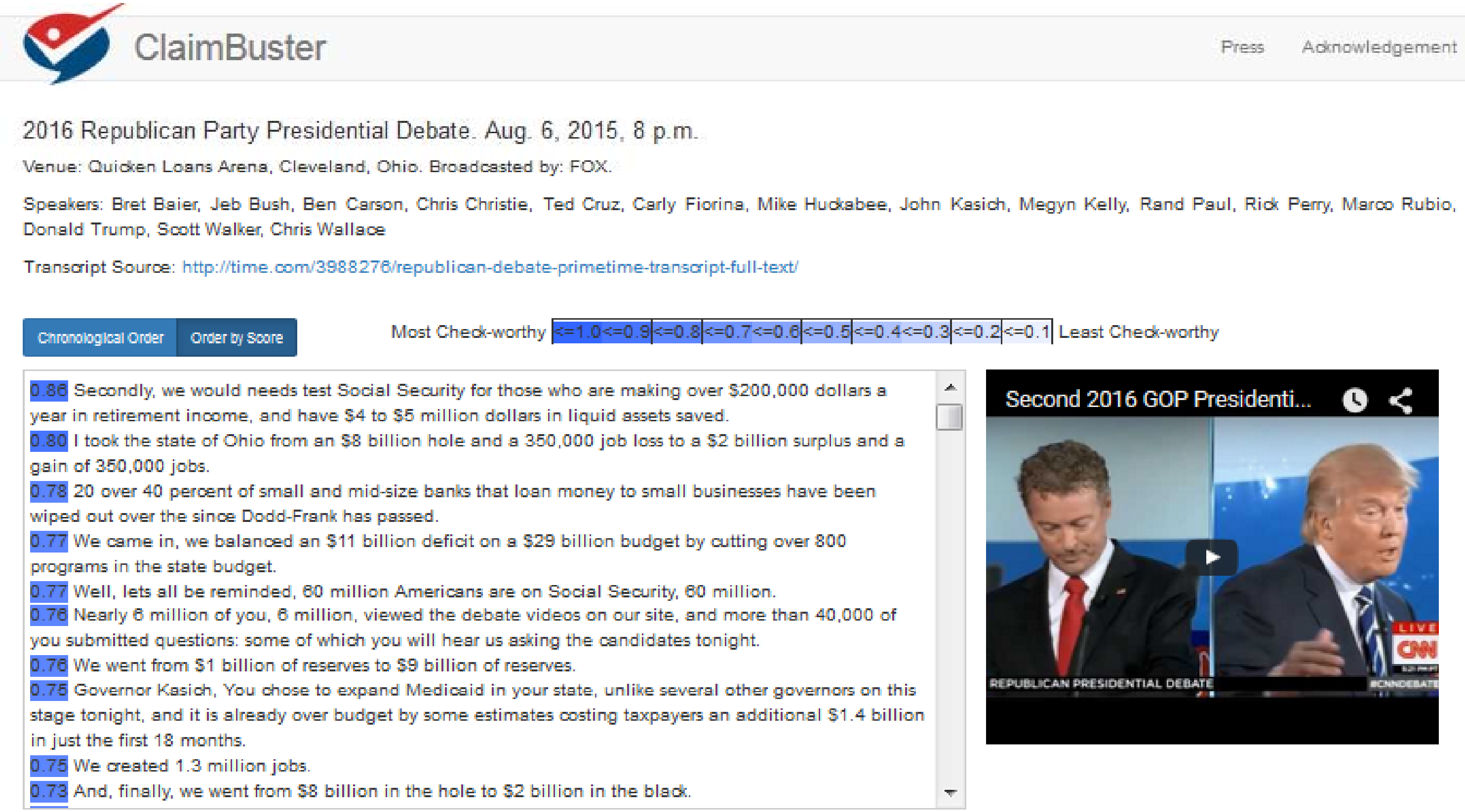
1.4 How WikiLeaks verifies its news stories

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

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

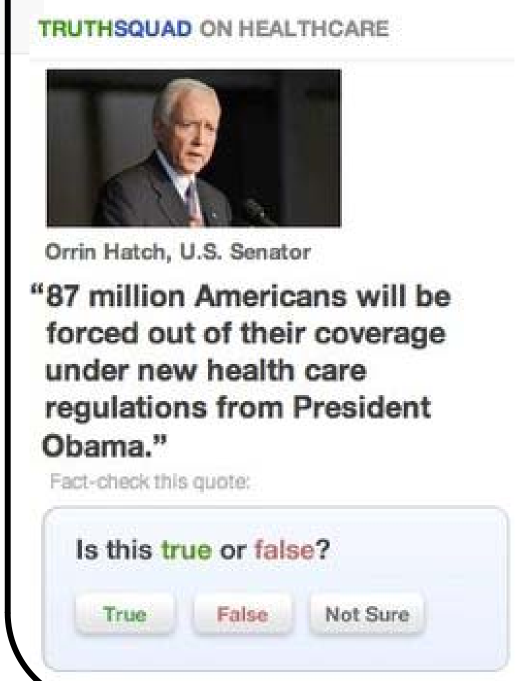
Transcript Source: http://time.com/3988276/republican-debate-primetime-transcript-full-text/

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?

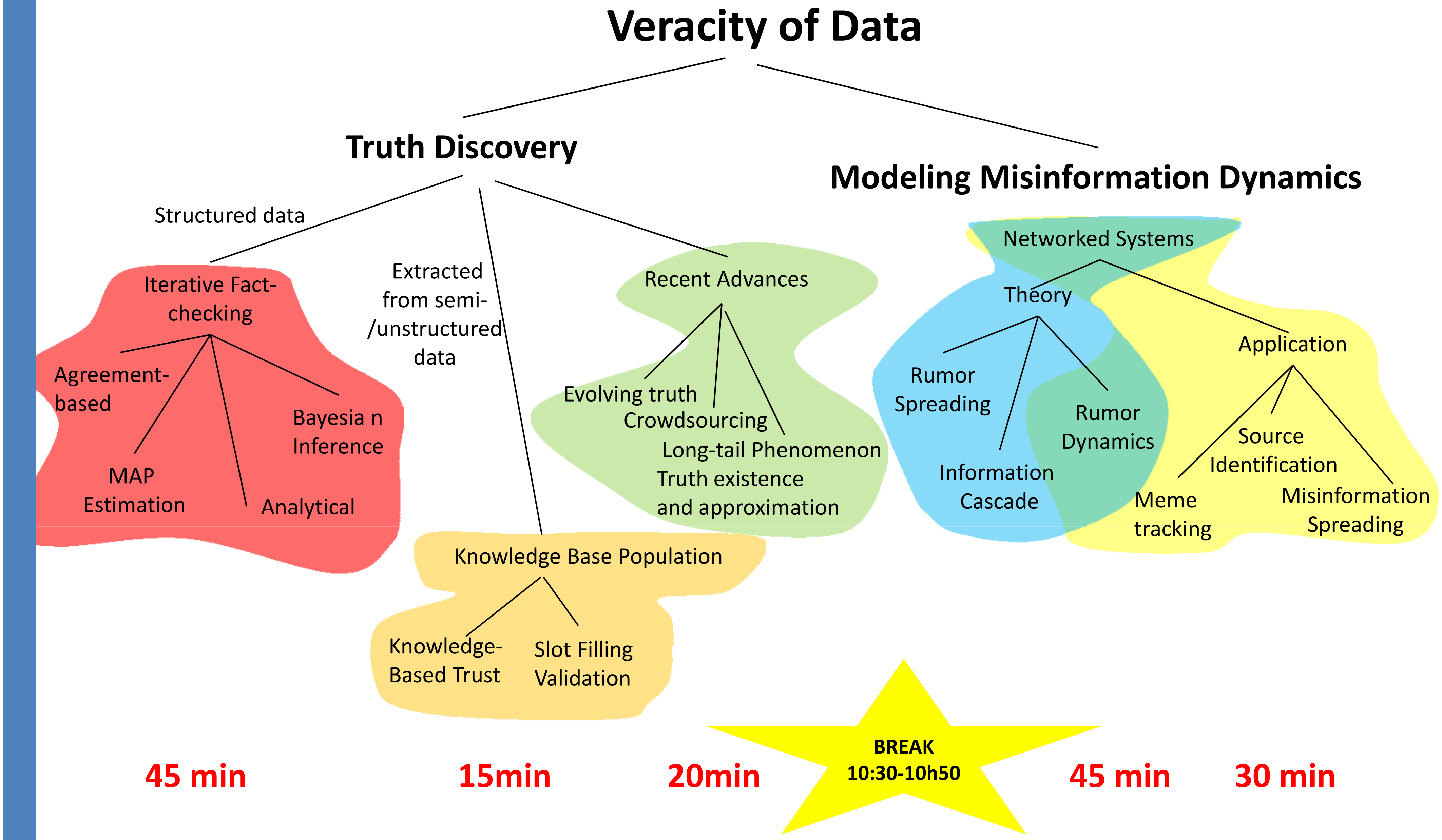
True False Not Sure

Truthsquad™

S. Cohen, J. T. Hamilton, and F. Turner. Computational journalism. CACM, 54(10):66–71, Oct. 2011.
S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In CIDR, 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



Veracity of Data

Truth Discovery

Structured data

Iterative Fact-checking

Agreement-based

Bayesia n Inference

MAP Estimation

Analytical

Extracted from semi-/unstructured data

Knowledge Base Population

Knowledge-Based Trust

Slot Filling Validation

Recent Advances

Evolving truth

Crowdsourcing

Long-tail Phenomenon

Truth existence and approximation

Modeling Misinformation Dynamics

Networked Systems

Theory

Rumor Spreading

Information Cascade

Rumor Dynamics

Application

Source Identification

Meme tracking

Misinformation Spreading

45 min

15min

20min

BREAK
10:30-10h50

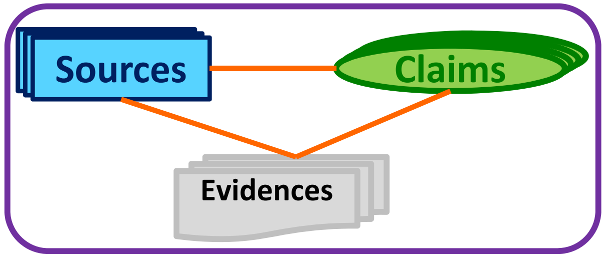
45 min

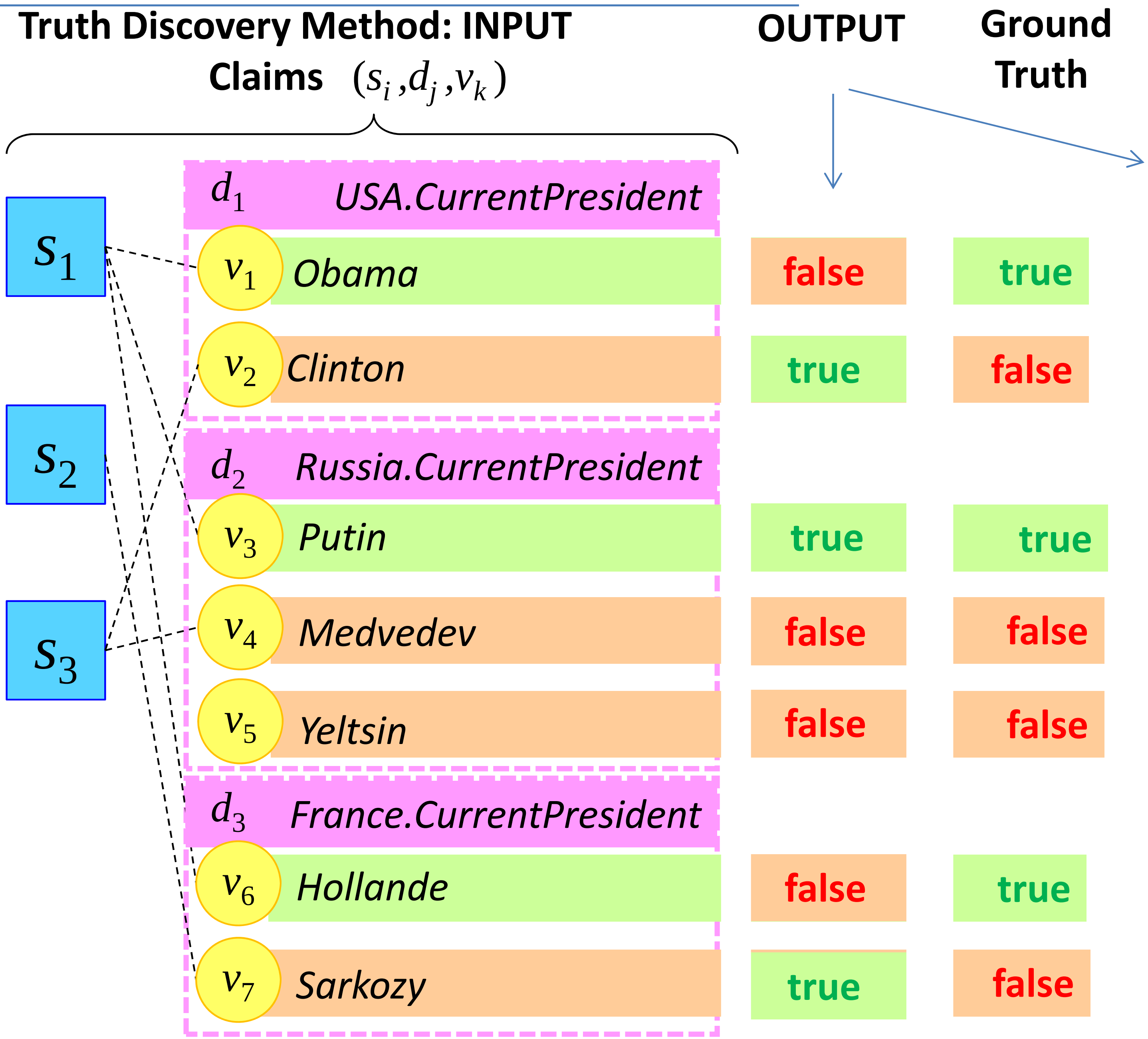
30 min

Outline

1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
4. Modeling Information Dynamics
5. Challenges

Terminology





		OUTPUT	Ground Truth
d_1 USA.CurrentPresident			
v_1	Obama	false	true
v_2	Clinton	true	false
d_2 Russia.CurrentPresident			
v_3	Putin	true	true
v_4	Medvedev	false	false
v_5	Yeltsin	false	false
d_3 France.CurrentPresident			
v_6	Hollande	false	true
v_7	Sarkozy	true	false

$C(v_k) \forall k$ Confidence of the values

$T(s_i) \forall i$ Trustworthiness of the sources

- s_i Source
- d_j Data item
- v_k Value
- ⬚ Mutual exclusive set
- true claim Fact
- false claim Allegation

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

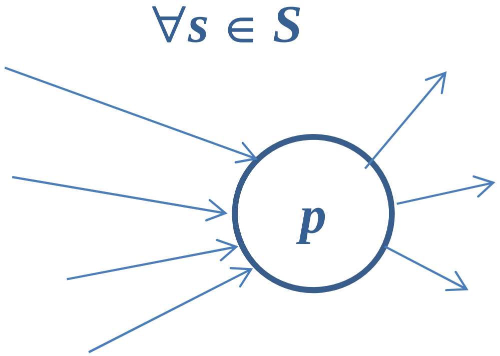
Compute the importance of a source in the Web graph based on the probability of landing on the source node by a random surfer

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

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



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

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

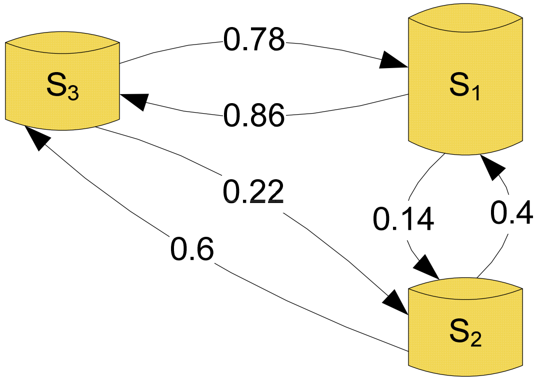
$$Auth^i(p) = \frac{1}{Z_a} \sum_{s \in S;\, s \to p} Hub^{i-1}(s)$$

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

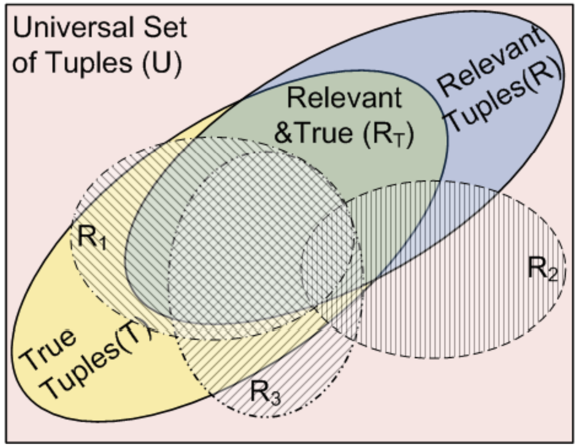


Probability of agreement of two independent false tuples

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

Probability of agreement of two independent true tuples

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



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

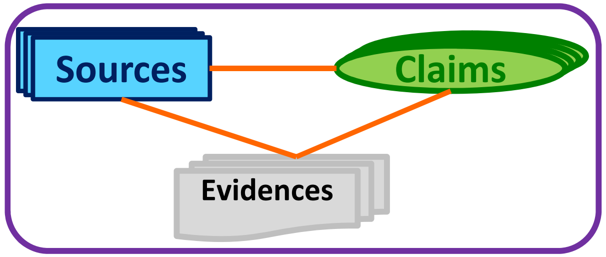
R. Balakrishnan, S. Kambhampati, SourceRank: Relevance and Trust Assessment for DeepWeb Sources Based on InterSource Agreement, In Proc. WWW 2009.

Agreement-Based Methods

Source Reputation Models

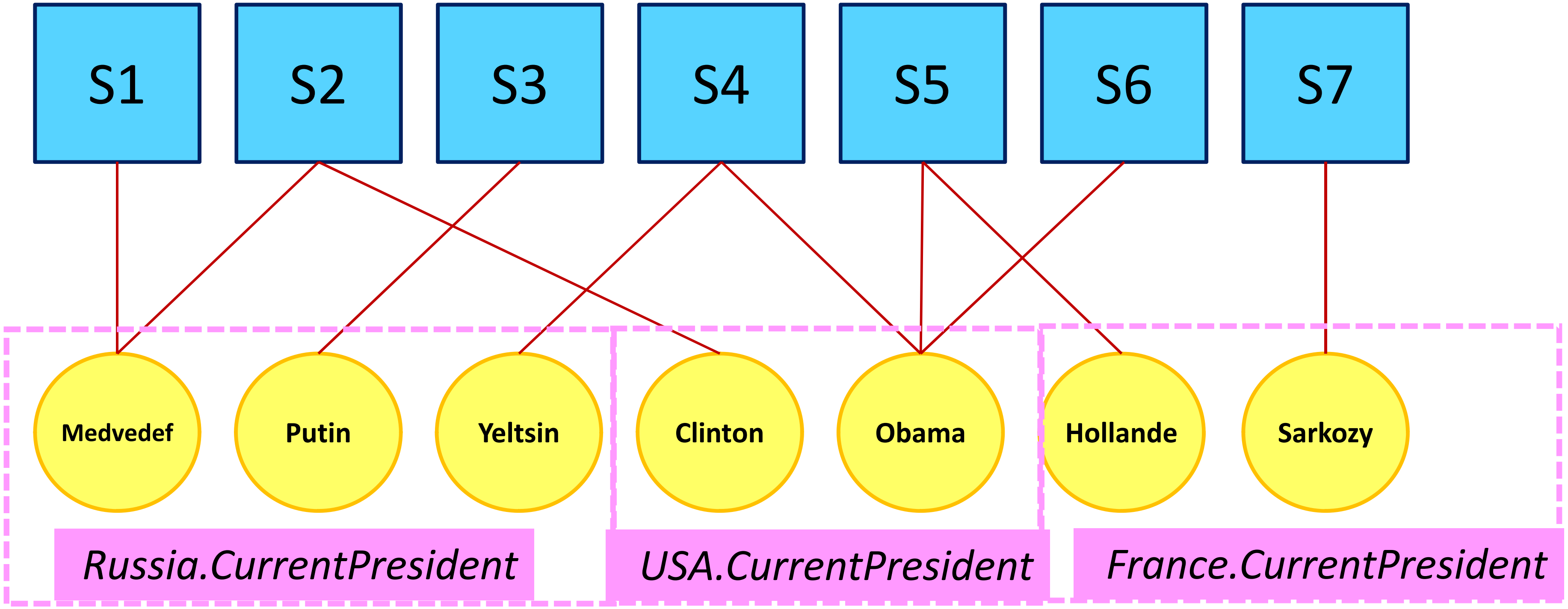
Only rely on source credibility is not enough

Source-Claim Iterative Models



Example

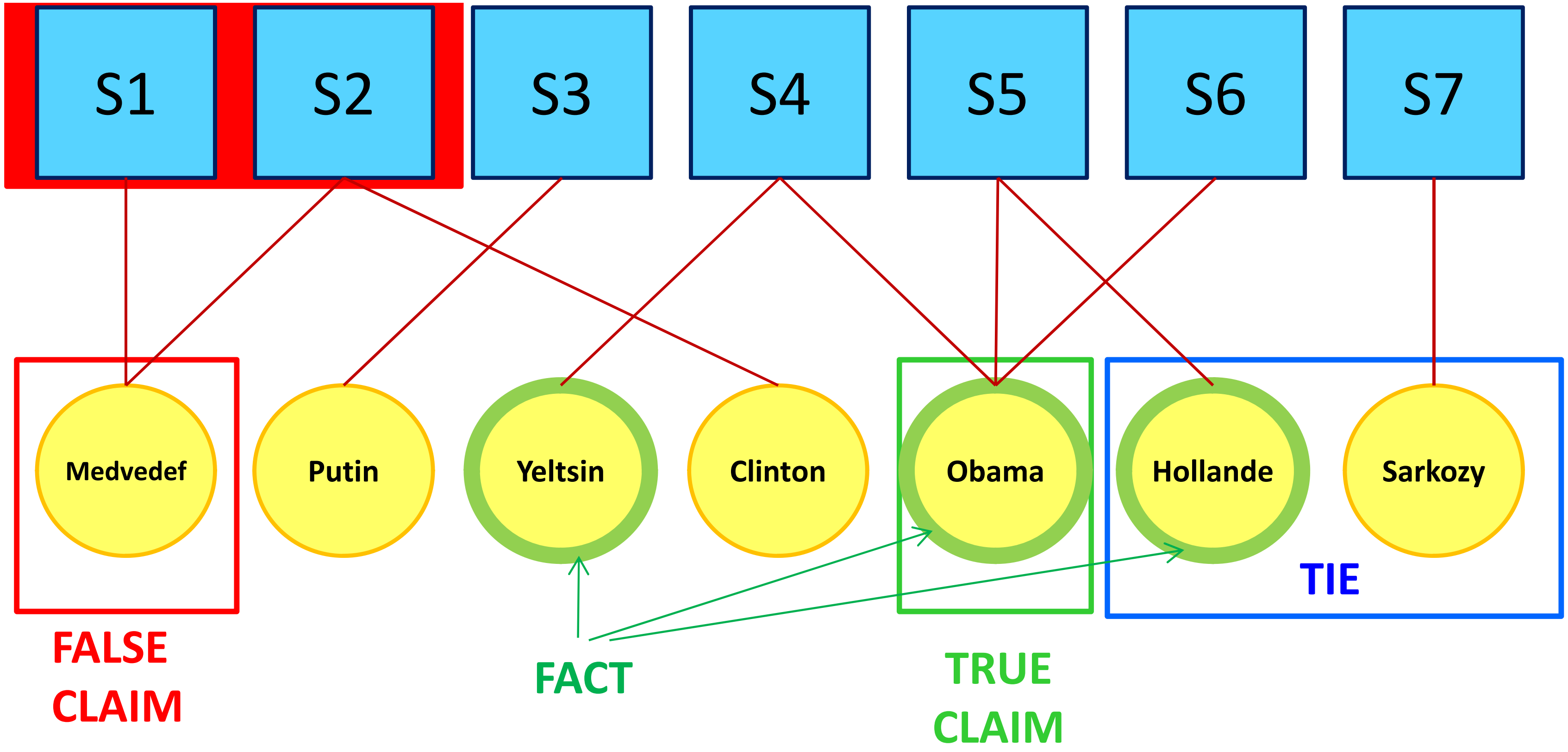
Seven sources disagree on the current president of Russia, Usa, and France
Can we discover the true values?



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)

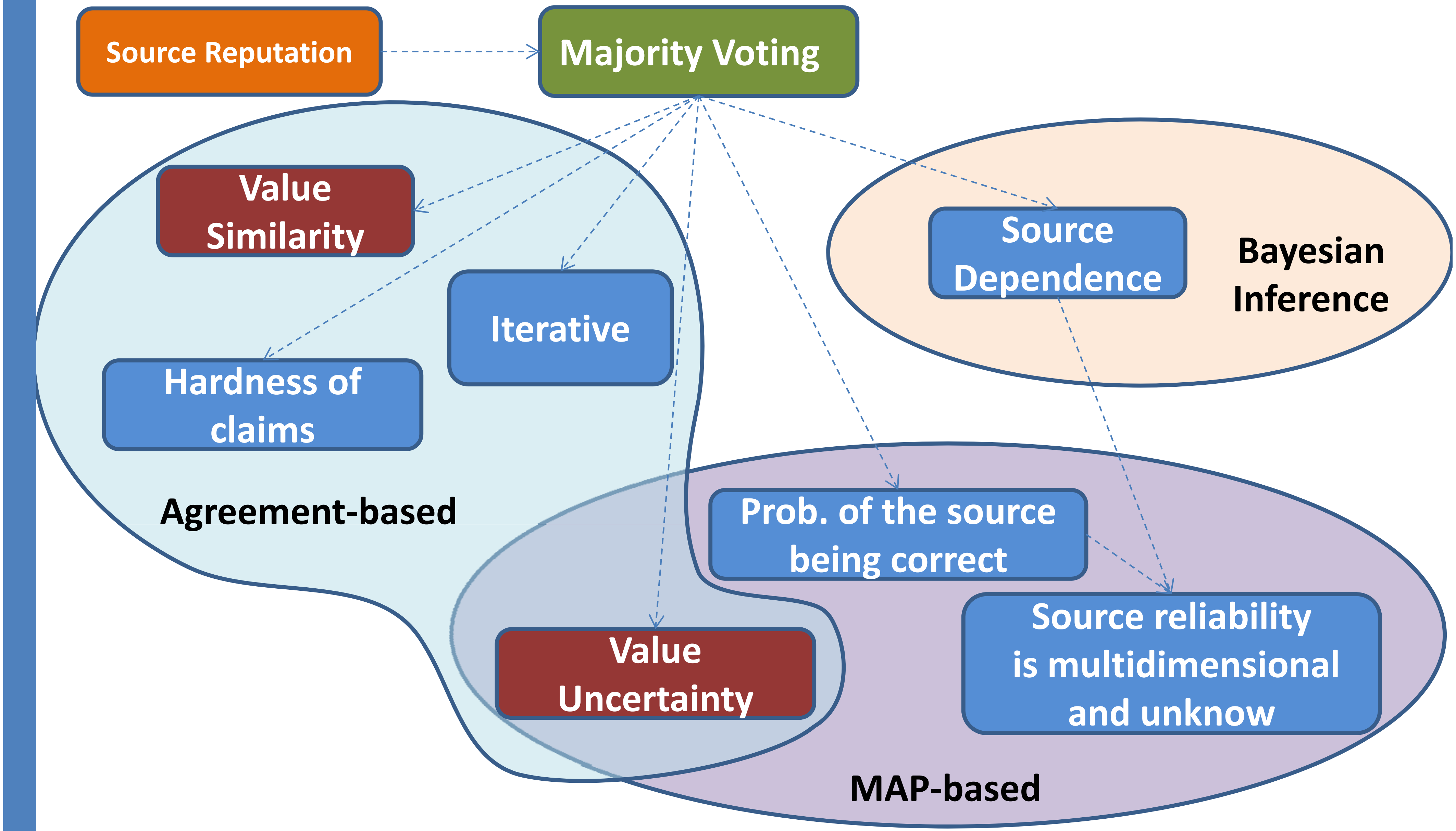
Originally written to provide theoritical basis of democracy

The majority vote will give an accurate value if at least $\lfloor S/2+1 \rfloor$ 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

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

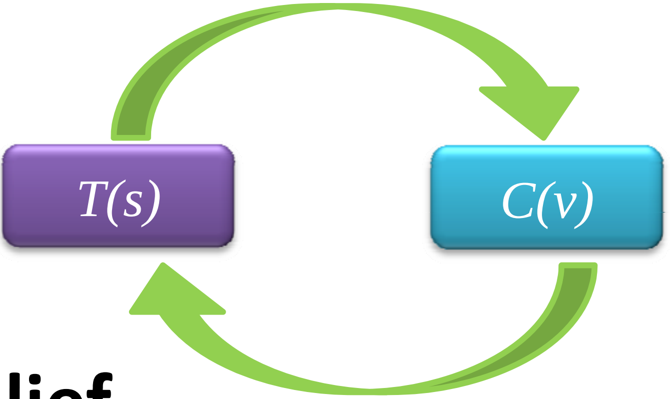
- If $p > 0.5$, then P_{MV} is monotonically increasing, $P_{MV} \to 1$ as $S \to \infty$
- If $p < 0.5$, then P_{MV} is decreasing and $P_{MV} \to 0$ as $S \to \infty$
- If $p = 0.5$, then $P_{MV} = 0.5$ for any S

Roadmap of Modeling Assumptions



Agreement-Based Methods

Source-Claim Iterative Models



Based on iterative computation of source trustworthiness and claim belief

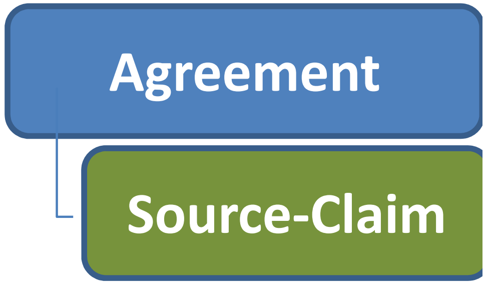
- Sums (adapted from HITS) *(1)*
- Average.Log, Investment, Pooled Investment *(1)*
- TruthFinder *(2)*
- Cosine, 2-Estimates, 3-Estimates *(3)*

(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.

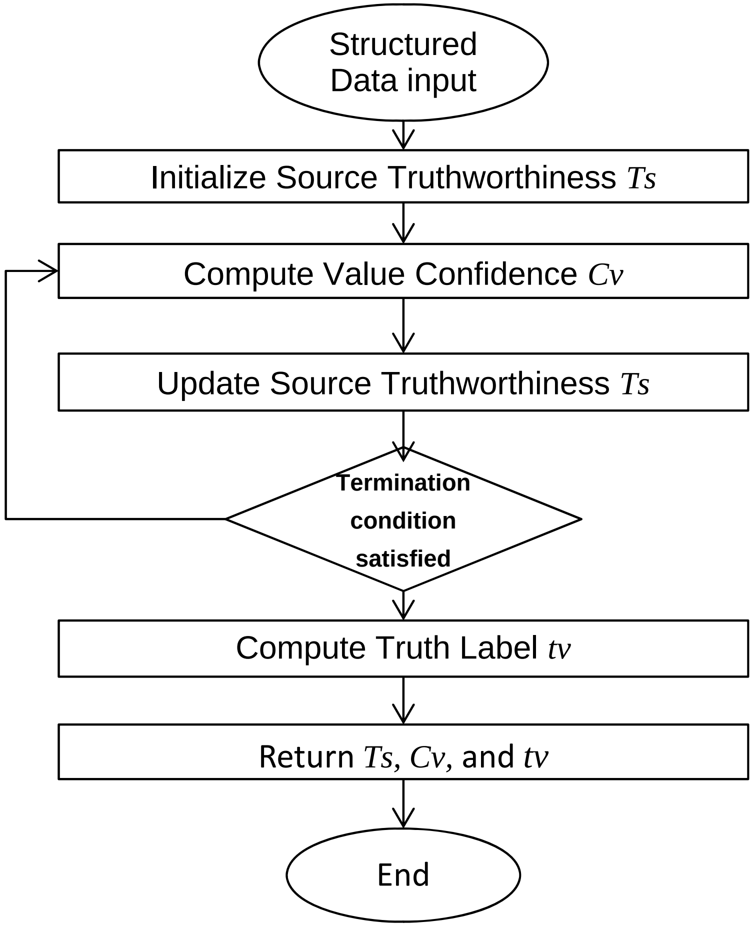
(2) X. Yin, J. Han, and P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. TKDE, 20(6):796–808, 2008.

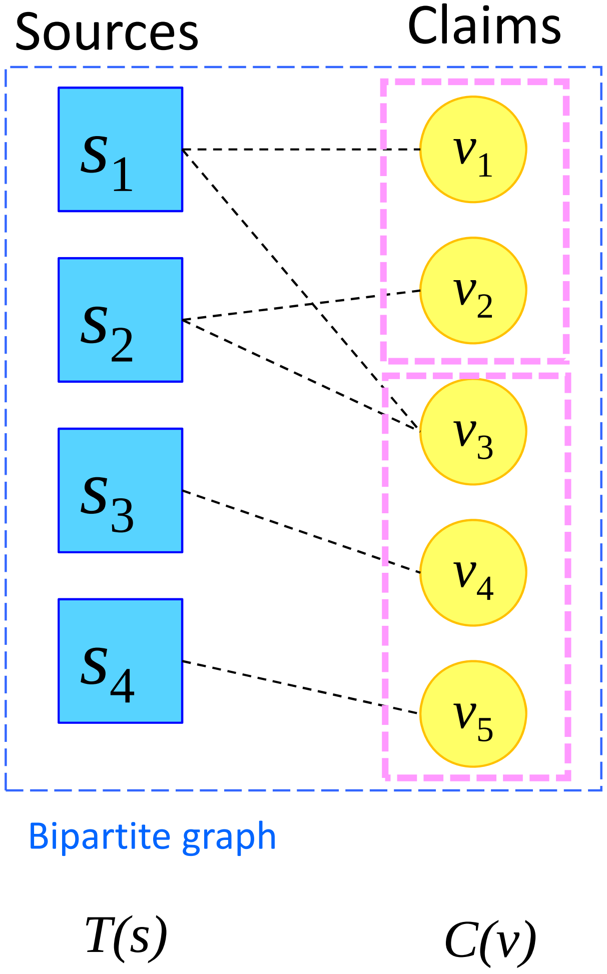
(3) A. Galland, S . Abiteboul, A. Marian, P. Senellart. Corroborating Information from Disagreeing Views. In Proc. of the ACM International Conference on Web Search and Data Mining (WSDM), pages 131–140, 2010.

Basic Principle



Iterative and transitive voting algorithm


Structured Data input
Initialize Source Truthworthiness Ts
Compute Value Confidence Cv
Update Source Truthworthiness Ts
Termination condition satisfied
Compute Truth Label tv
Return Ts, Cv, and tv
End



Example (cont'd)

Agreement

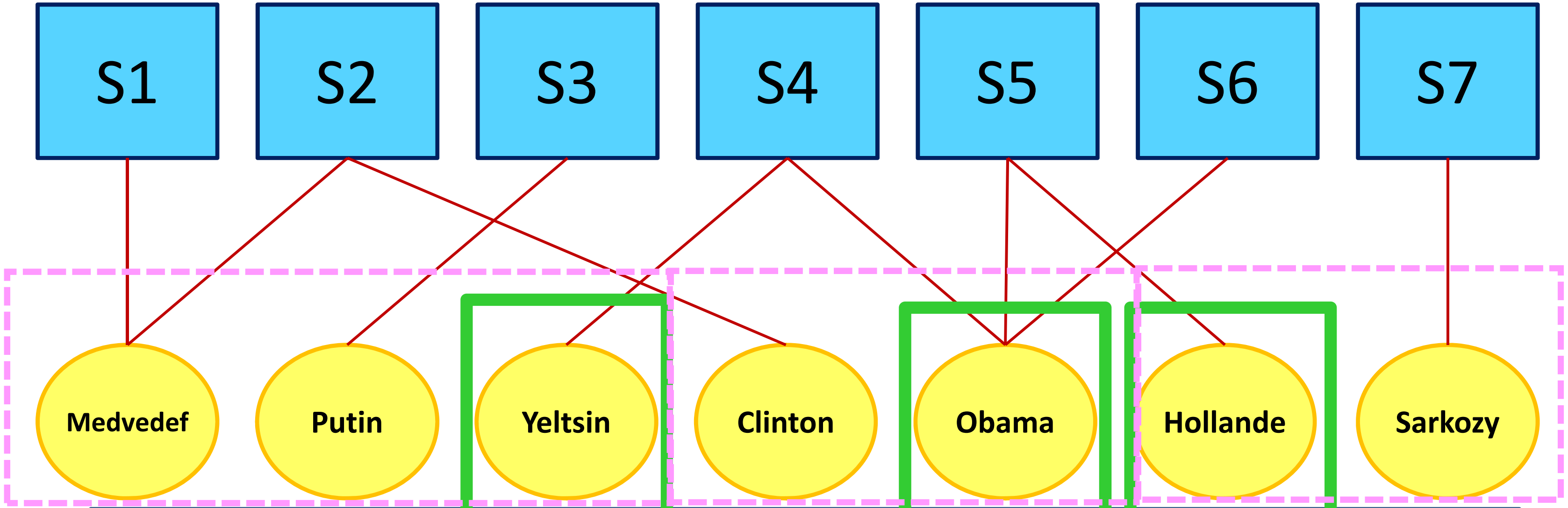
Source-Claim

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

Initialization: We believe in each claim equally

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

Source Trustworthiness T_s



	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

Value Confidence C_v

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

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

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

with $G(x)=x^{1.2}$

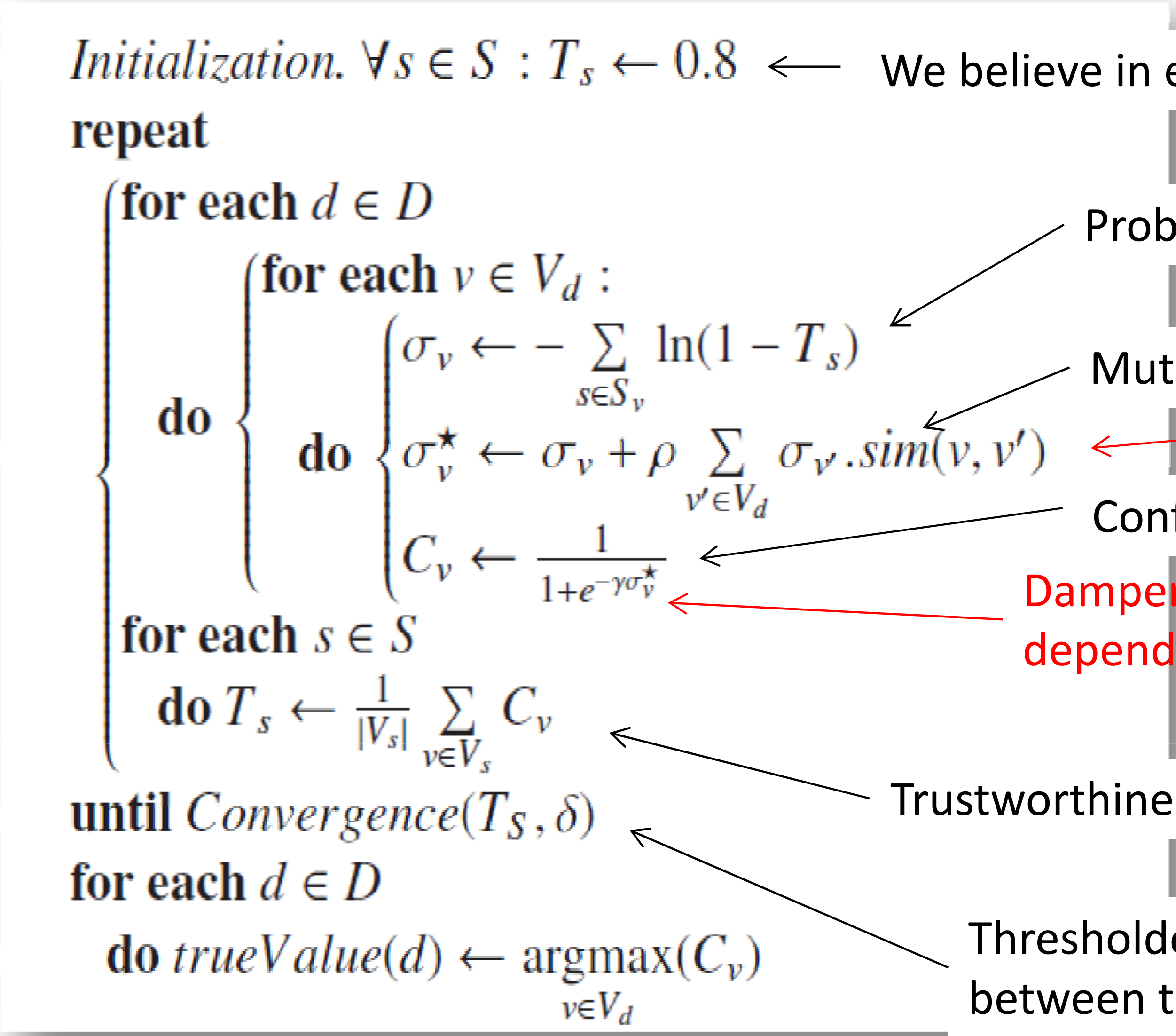
J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In COLING, pages 877–885. Association for Computational Linguistics, 2010.

TruthFinder

Value Similarity

Agreement

Source-Claim



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

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 Ts 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

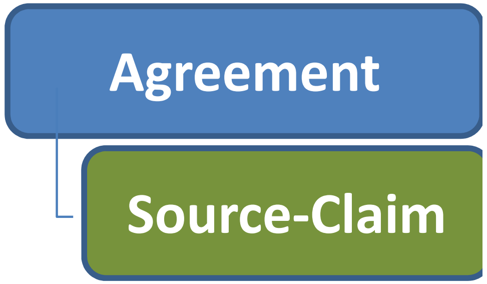
3. Prior Knowledge and Assumptions

- ❑ 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, δ, γ, ρ
□ Repeatability	Yes
□ Convergence and stopping criteria	δ 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

Maximum Likelihood Estimation

MAP

EM

Social Sensing

- Reliability that Participant *i* reports measured variable *j* :

$$t_i = P(C_j^{true} | S_i C_j)$$

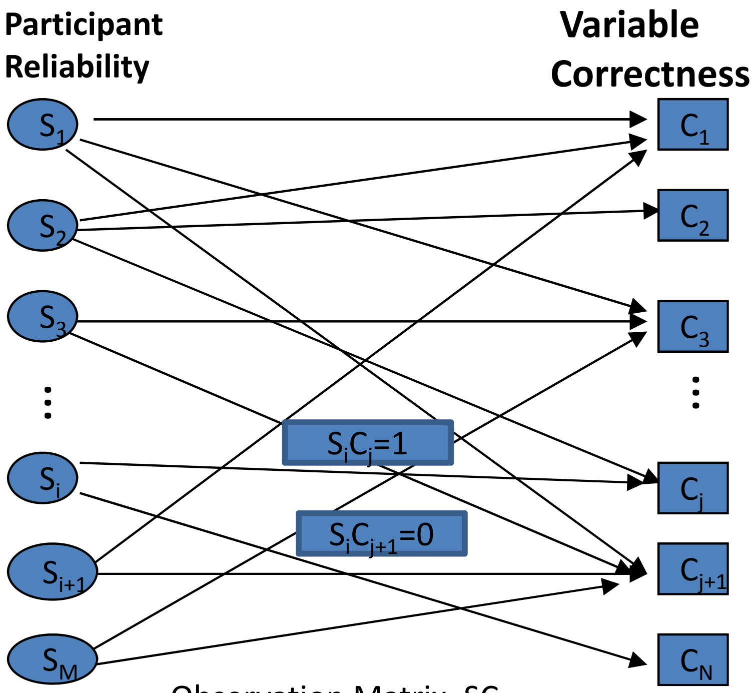
- Speak Rate of Participant *i*

$$s_i = P(S_i C_j)$$

- Source reliability parameters

$$a_i = \frac{t_i \times s_i}{d} \qquad b_i = \frac{(1-t_i) \times s_i}{1-d}$$

$Z=\{z_1, z_2, \ldots z_N\}$ where z_j =1 when assertion C_j is correct and 0 otherwise



Observation Matrix, SC

$$\theta = (a_1, \ldots, a_M ; b_1, \ldots, b_M)$$

Expectation Step (E-step)

Source reliability

$$Q(\theta | \theta^{(t)}) = E_{Z|SC,\theta^{(t)}} \left[\log \sum_z P(SC, z | \theta) \right]$$

Variable Correctness (hidden)

Maximization Step (M-step)

$$\theta^{(t+1)} = \arg\max_\theta \left(Q(\theta | \theta^{(t)}) \right)$$

D. Wang, L.M. Kaplan, H. Khac Le, and T. F. Abdelzaher. On Truth Discovery in Social Sensing: a Maximum Likelihood Estimation Approach. In Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN), p. 233–244, 2012.

MLE Signature

MAP

EM

1. Method Characteristics

❑ Initialization and parameter settings	T_s, s, d *(prior truth prob.)*
❑ Repeatability	Yes
❑ Convergence and stopping criteria	***K iterations***
❑ Complexity	*O(KSV)*
❑ Scalability	Yes

2. Input Data

❑ Type of value	**Boolean**
❑ Mono-/multi-valued claims	**Mono-valued**
❑ Similarity of claims	No
❑ Correlations between attributes or objects	No

3. Prior Knowledge

❑ Source Quality	Constant, source-specific
❑ Dependence of sources	No
❑ 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

Latent Credibility Analysis

MAP

EM

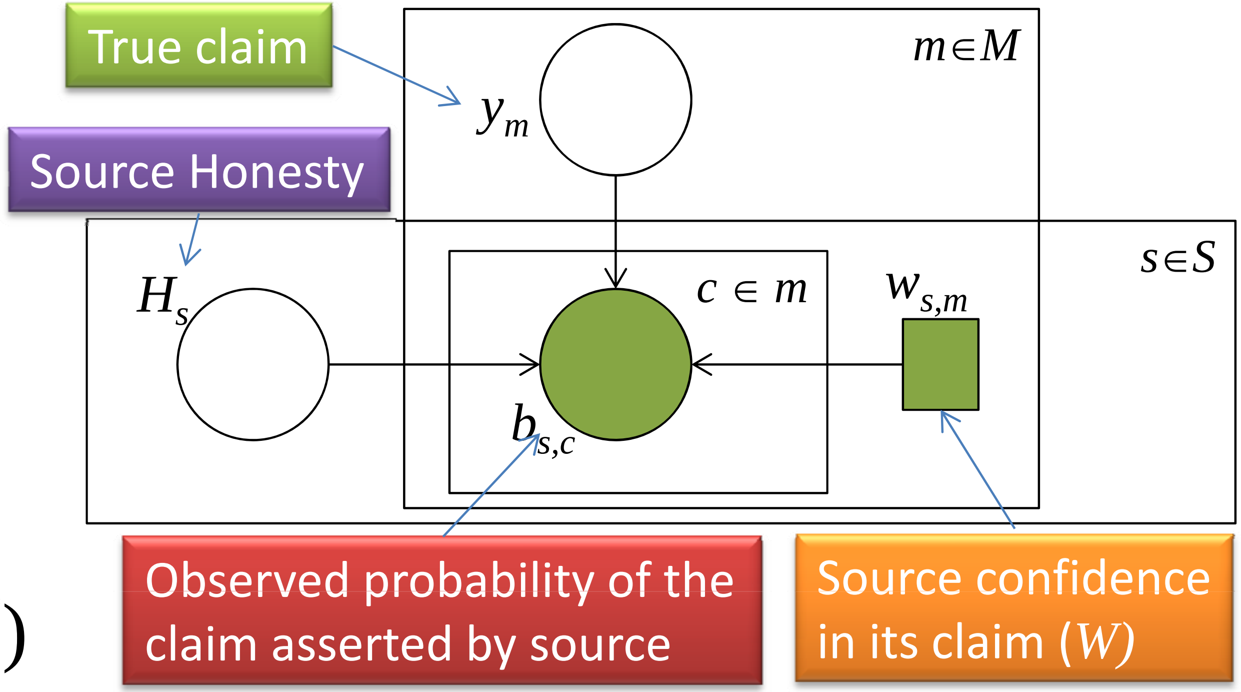
SimpleLCA, GuessLCA, MistakeLCA, LieLCA

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:

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



True claim

Source Honesty

Observed probability of the claim asserted by source

Source confidence in its claim (W)

Latent variables θ

- 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 World Wide Web (WWW '13), 2013.

LCA Signature

MAP

EM

1. Method Characteristics

□ Initialization and parameter settings	W, K, β_1 *(prior truth prob./claim)*
□ Repeatability	Yes
□ Convergence and stopping criteria	***K* iterations**
□ Complexity	$O(KSD)$
□ Scalability	Yes

2. Input Data

□ Type of value	String, categorical
□ Mono-/multi-valued claims	Multi-valued
□ Similarity of claims	Yes (as joint probability)
□ Correlations between attributes or objects	No

3. Prior Knowledge

□ Source Quality	Constant, source- and entity-specific
□ Dependence of sources	No
□ Hardness of certain claims	Yes

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

Latent Truth Model (LTM)

MAP

Gibbs Sampling

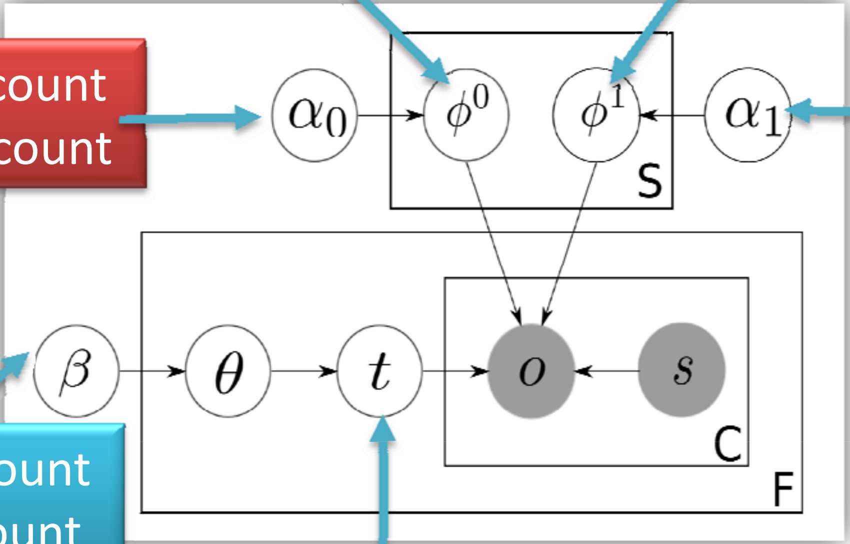
Collapsed Gibbs sampling to get MAP estimate for ***t***

False positive rate

sensitivity

α_{01} Prior false positive count
α_{00} Prior true negative count

α_{11} Prior true positive count
α_{10} Prior false negative count



β_0 Prior false count
β_1 Prior true count

Truth label

B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conicting 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	$(T_s$, ***K***, *Burn-in*, *Thin*, $\alpha_{00}, \beta_{00}, \alpha_{01}, \beta_{01}, \alpha_{10}, \beta_{10}, \alpha_{11}, \beta_{11})$
□ Repeatability	**No (Gibbs sampling)**
□ Convergence and stopping criteria	***K* iterations**
□ Complexity	$O(KSV)$
□ Scalability	Yes

2. Input Data

□ Type of value	String, categorical
□ Mono-/multi-valued claims	**Mono-valued (multiple claims/per source)**
□ Similarity of claims	No
□ Correlations between attributes or objects	No

3. Prior Knowledge

□ Source Quality	Incremental, source-specific, homogeneous/entity
□ Dependence of sources	No
□ Hardness of certain claims	No

4. Output

□ Single/multiple truth per data item	**Multiple true values 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

Analytical Solutions

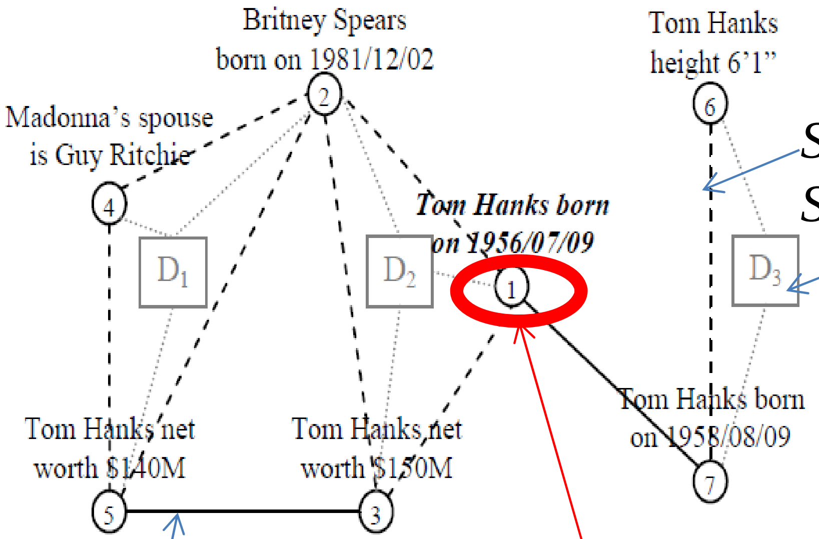
Matrix Diagonalization

Semi-Supervised Truth Discovery (SSTF)

Minimize loss funtion

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

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


Britney Spears born on 1981/12/02
Tom Hanks height 6’1”
Madonna’s spouse is Guy Ritchie
Tom Hanks born on 1956/07/09
D1
D2
D3
1
2
3
4
5
6
7
Tom Hanks net worth $140M
Tom Hanks net worth $150M
Tom Hanks born on 1958/08/09

Same Object

Same Source

Ground truth fact

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.

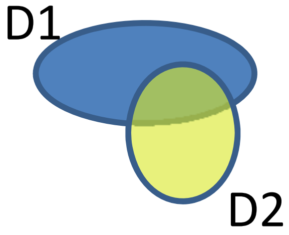
Related Work: L. Ge, J. Gao, X. Yuy, W. Fanz and A. Zhang, Estimating Local Information Trustworthiness via Multi-Source Joint Matrix Factorization, Proc. of ICDM 2012

Outline

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

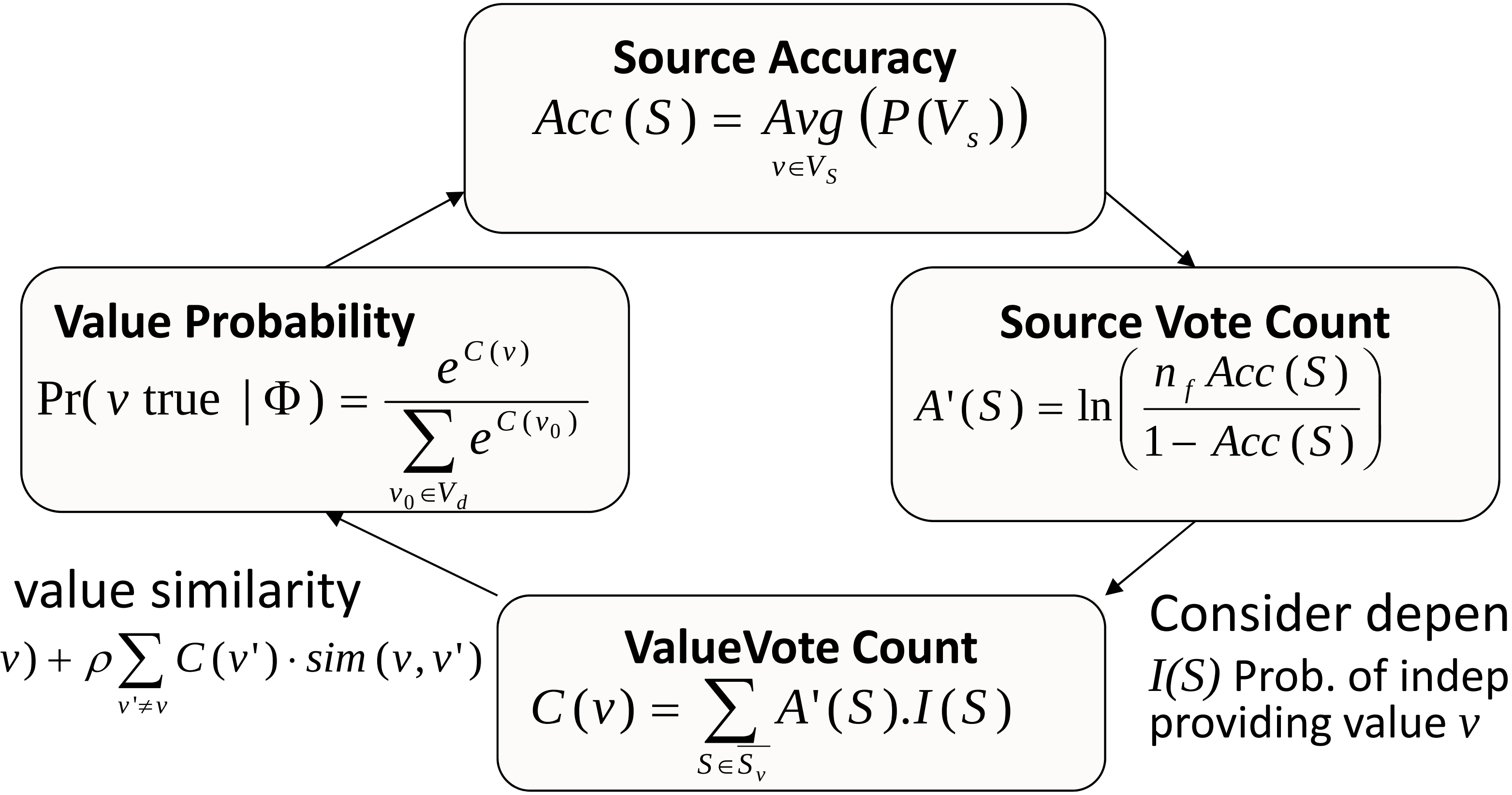
Bayesian

Source Dependence



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

  |Acc(D1 ∩ D2)-Acc(D1-D2)| > |Acc(D1 ∩ D2)-Acc(D2-D1)|⇒ S1→S2



Source Accuracy

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

Source Vote Count

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

ValueVote Count

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

Value Probability

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

Consider value similarity

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

Consider dependence

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

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

Bayesian

Depen Signature

1. Method Characteristics

❑ Initialization and parameter settings	T_s, n_f *(nb false value)*, ε *(error rate)* ,α *(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** : $S_{True} >> S_{False}$

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]

[1] [Dong et al, VLDB’09]

[2] [Pasternack Roth, WWW’13]

[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

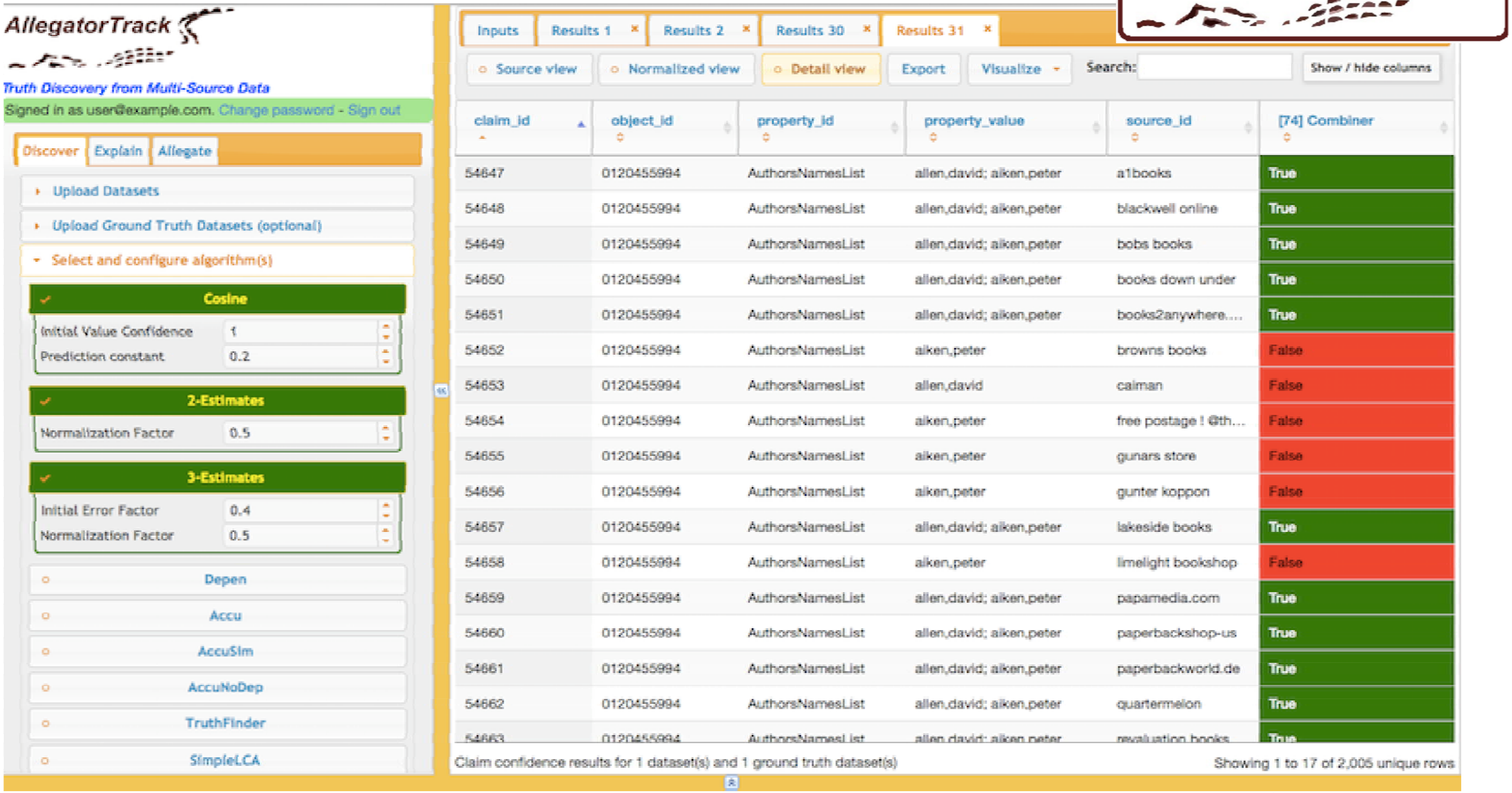
D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: An experimental evaluation. arXiv preprint arXiv:1409.6428, 2014.

Further Testing

API

AllegatorTrack

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



AllegatorTrack

Truth Discovery from Multi-Source Data

Signed in as user@example.com. Change password - Sign out

Discover
Explain
Allegate
Upload Datasets
Upload Ground Truth Datasets (optional)
Select and configure algorithm(s)

✓ Cosine

Initial Value Confidence	1
Prediction constant	0.2

✓ 2-Estimates

Normalization Factor	0.5

✓ 3-Estimates

Initial Error Factor	0.4
Normalization Factor	0.5

○ Depen
○ Accu
○ AccuSim
○ AccuNoDep
○ TruthFinder
○ SimpleLCA

claim_id	object_id	property_id	property_value	source_id	[74] Combiner
54647	0120455994	AuthorsNamesList	allen,david; aiken,peter	a1books	True
54648	0120455994	AuthorsNamesList	allen,david; aiken,peter	blackwell online	True
54649	0120455994	AuthorsNamesList	allen,david; aiken,peter	bobs books	True
54650	0120455994	AuthorsNamesList	allen,david; aiken,peter	books down under	True
54651	0120455994	AuthorsNamesList	allen,david; aiken,peter	books2anywhere....	True
54652	0120455994	AuthorsNamesList	aiken,peter	browns books	False
54653	0120455994	AuthorsNamesList	allen,david	caiman	False
54654	0120455994	AuthorsNamesList	aiken,peter	free postage ! @th...	False
54655	0120455994	AuthorsNamesList	aiken,peter	gunars store	False
54656	0120455994	AuthorsNamesList	aiken,peter	gunter koppon	False
54657	0120455994	AuthorsNamesList	allen,david; aiken,peter	lakeside books	True
54658	0120455994	AuthorsNamesList	aiken,peter	limelight bookshop	False
54659	0120455994	AuthorsNamesList	allen,david; aiken,peter	papamedia.com	True
54660	0120455994	AuthorsNamesList	allen,david; aiken,peter	paperbackshop-us	True
54661	0120455994	AuthorsNamesList	allen,david; aiken,peter	paperbackworld.de	True
54662	0120455994	AuthorsNamesList	allen,david; aiken,peter	quartermelon	True
54663	0120455994	AuthorsNamesList	allen,david; aiken,peter	revaluation books	True

Claim confidence results for 1 dataset(s) and 1 ground truth dataset(s)

Showing 1 to 17 of 2,005 unique rows

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

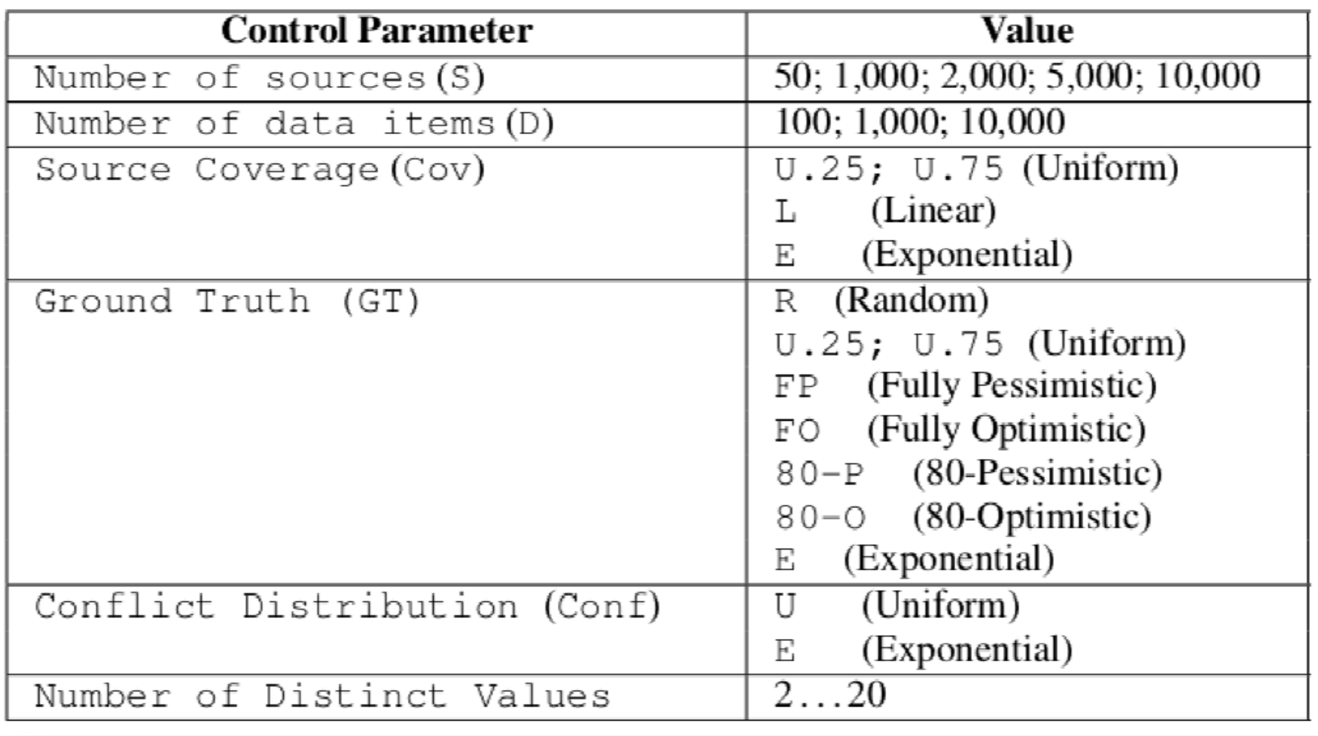
API

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

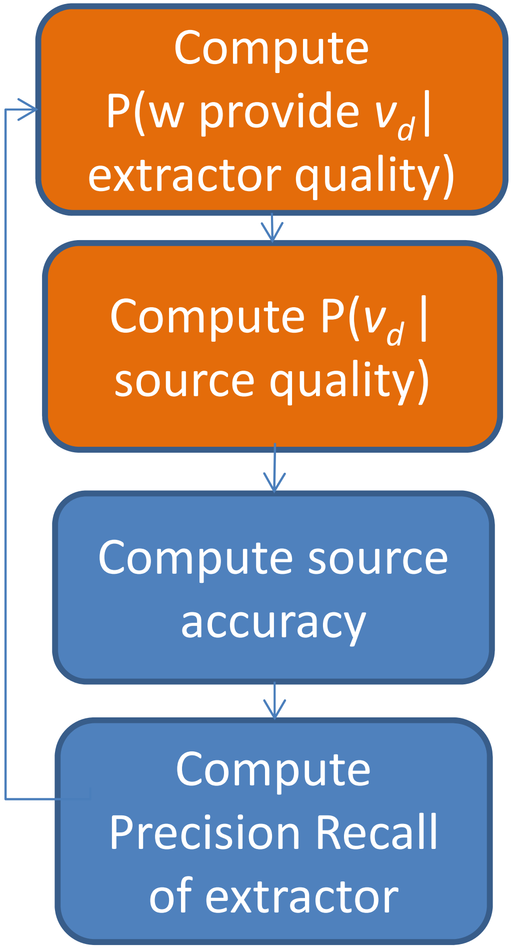
Knowledge-Based Trust

Bayesian

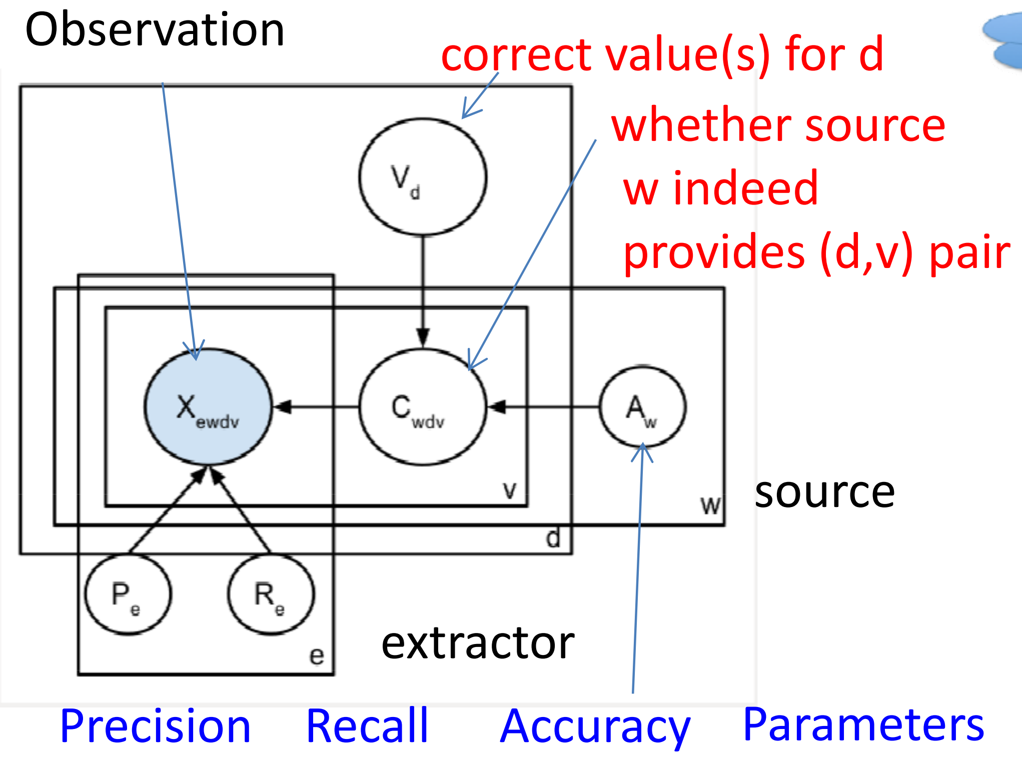
EM

Distinguish extractor errors from source errors

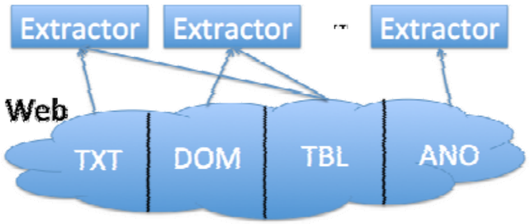
Multi-Layer Model based on EM

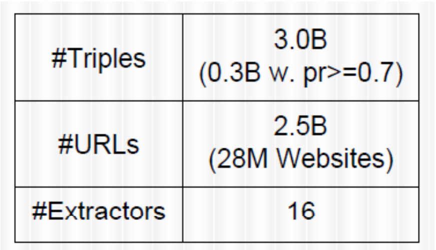


Compute P(w provide v_d | extractor quality)
Compute P(v_d | source quality)
Compute source accuracy
Compute Precision Recall of extractor



KNOWLEDGE VAULT





#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**

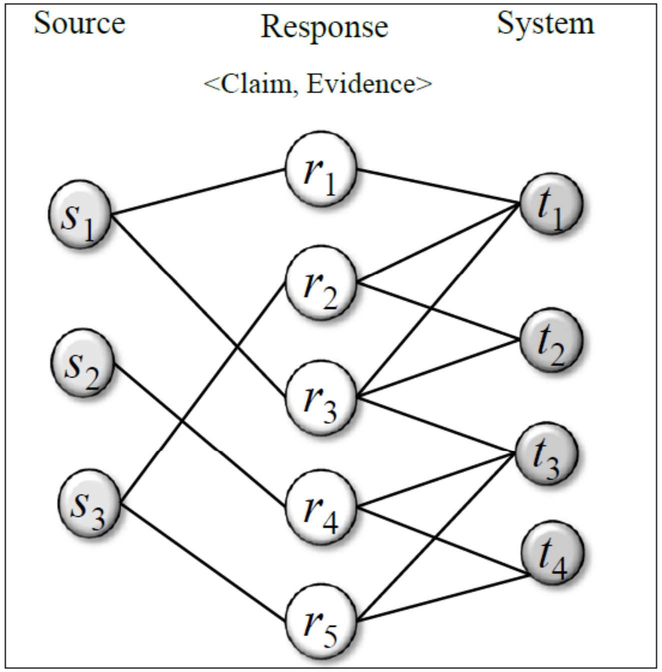
Credibility Propagation

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:

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



W^{sr} W^{rt}

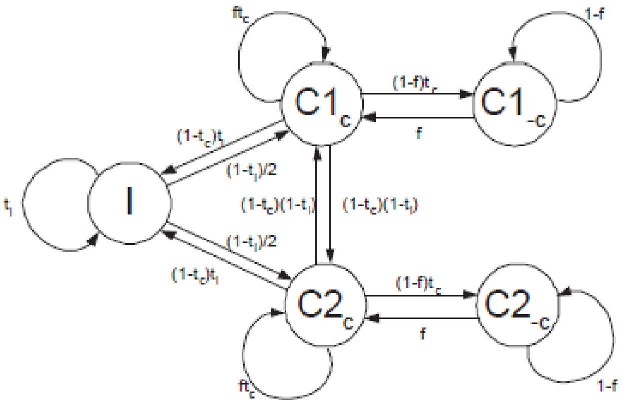
Weight matrices

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. 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

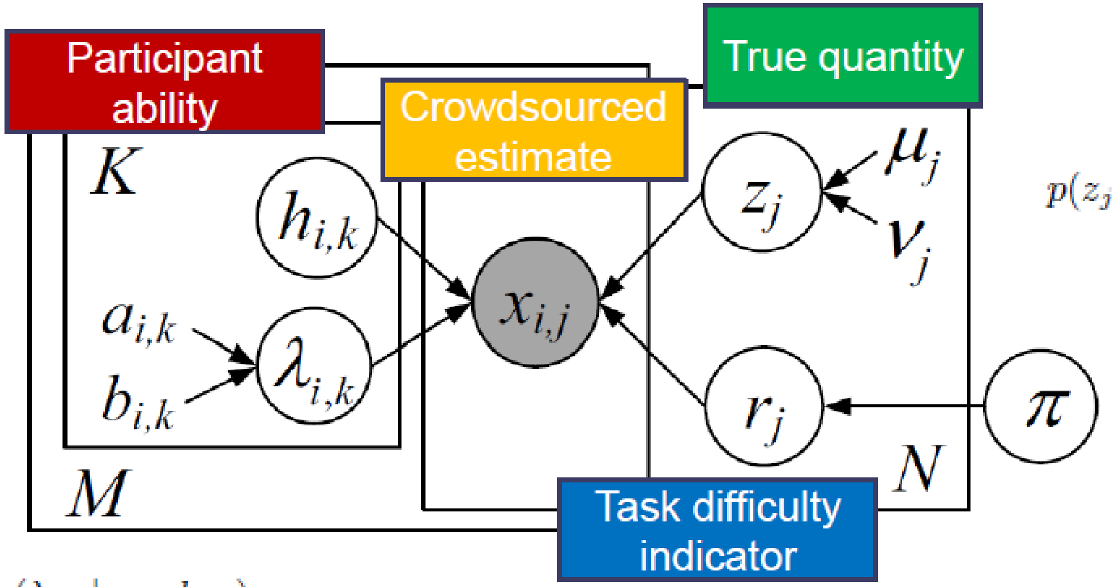
Truth discovery from crowdsourced data

Expectation Maximization

TBP (Truth Bias and Precision)

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

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



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

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

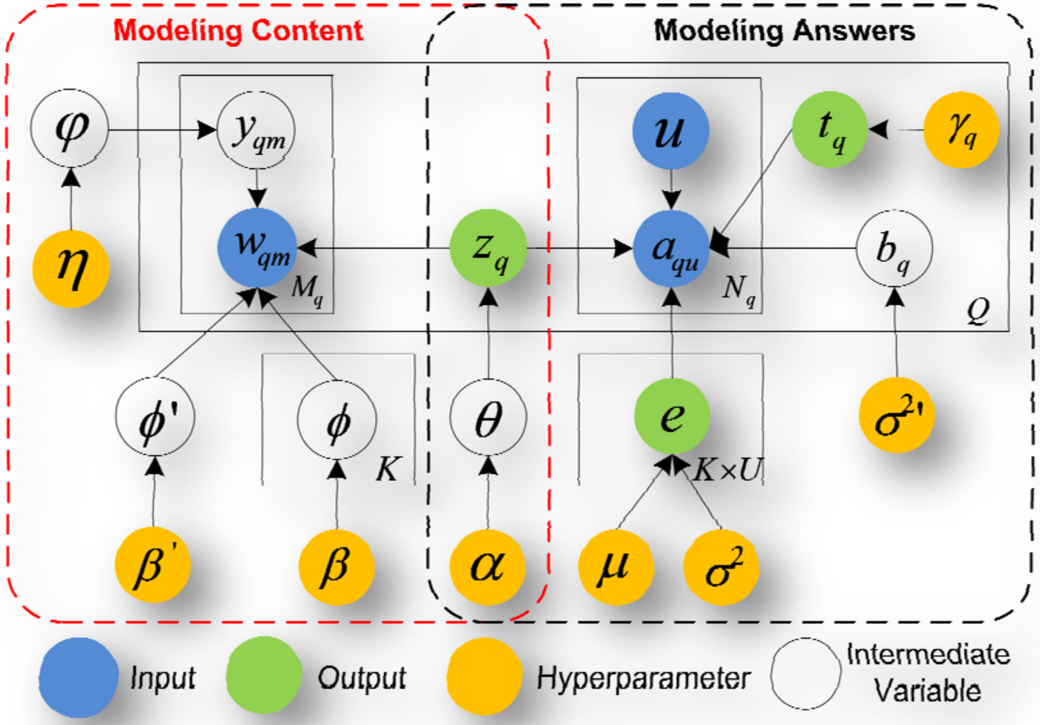
$$p(\lambda_{i,k}|a_{i,k}, b_{i,k}) = \mathsf{Gamma}(\lambda_{i,k}|a_{i,k}, b_{i,k})$$

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



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

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

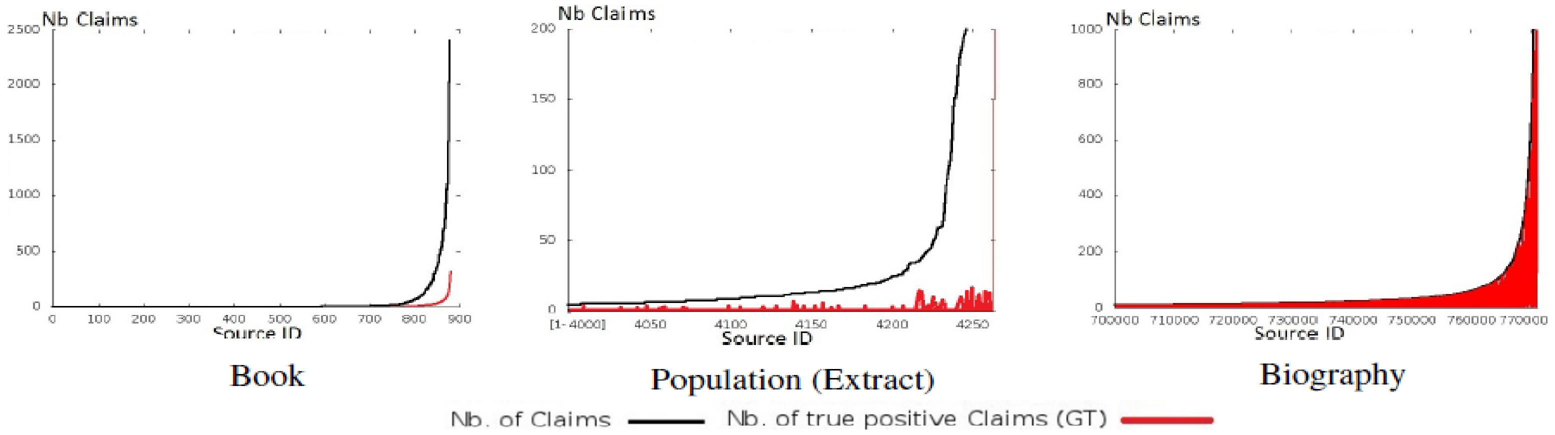
$$a_{qu} | t_q \sim logistic(e_{z_q u}, b_q)$$

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

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



Book

Population (Extract)

Biography

CADT Method for Independent and Benevolent Sources

Goal : Minimize the Variance of Source Reliability

$$\varepsilon_s \propto N(0,\sigma_s^2) \qquad \varepsilon_{combined} = \frac{\sum_{s \in S} w_s \varepsilon_s}{\sum_{s \in S} w_s}$$

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

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

Number of claims by source *s*

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

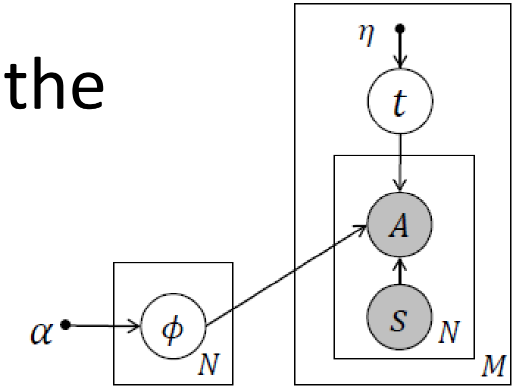
Initial value confidence for entity *n*

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 2015

- **Multi-Truth Discovery**

Tuesday, 3:55pm–5:10pm, Session 3A: Veracity

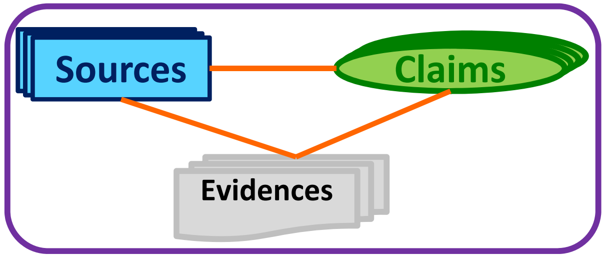
X. Wang, X. Xu, X. Li. An Integrated Bayesian Approach for Effective Multi-Truth Discovery. In CIKM 2015

- **Approximate Truth Discovery**

Tuesday, 3:55pm–5:10pm, Session 3A: Veracity

X. Wang, Q. Z. Sheng, X. S. Fang, X. Xu, X. Li, L. Yao. Approximate Truth Discovery Via Problem Scale Reduction. In CIKM 2015

Truth Discovery Challenges

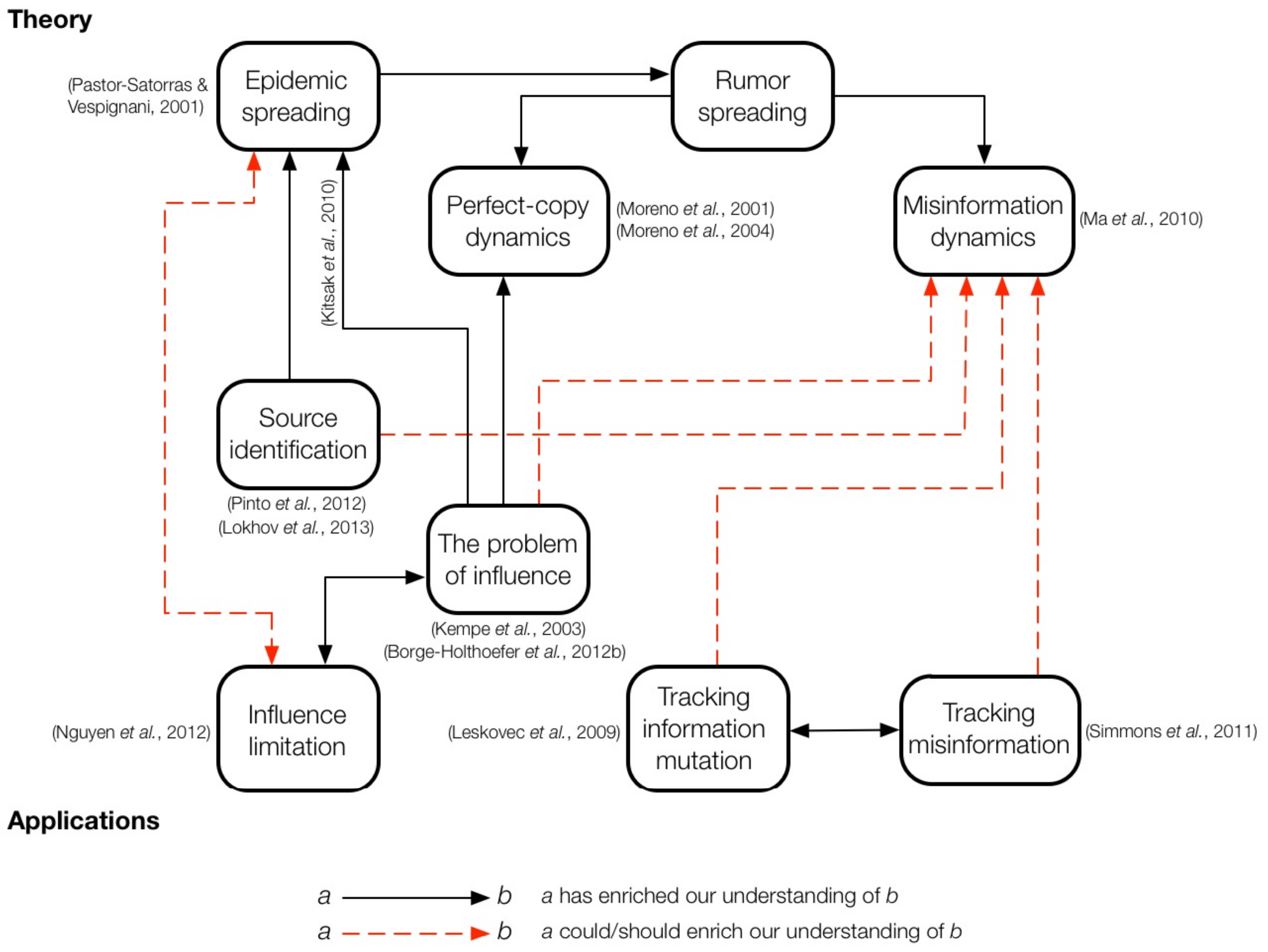


- **Multidimensional Metrics**
  - Source: Coverage, Accuracy, Exacteness, Freshness, Reputation, Dependence…
  - Claims: Popularity (i.e., supported by many or few sources) (long-tail phenomena)
  - Truth: Trivial truths (hardeness), sensitive truths, uncertain, rapidly evolving
  - Data items: Information entropy (many (or few) conflicting information)
- **Truth Discovery Modeling**
  - Voting only works with benevolent sources. What about adversarial/pessimitic scenarios?
  - Need to incorporate evidences and contextual metadata (hidden agenda of sources)
  - Need to adress 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**

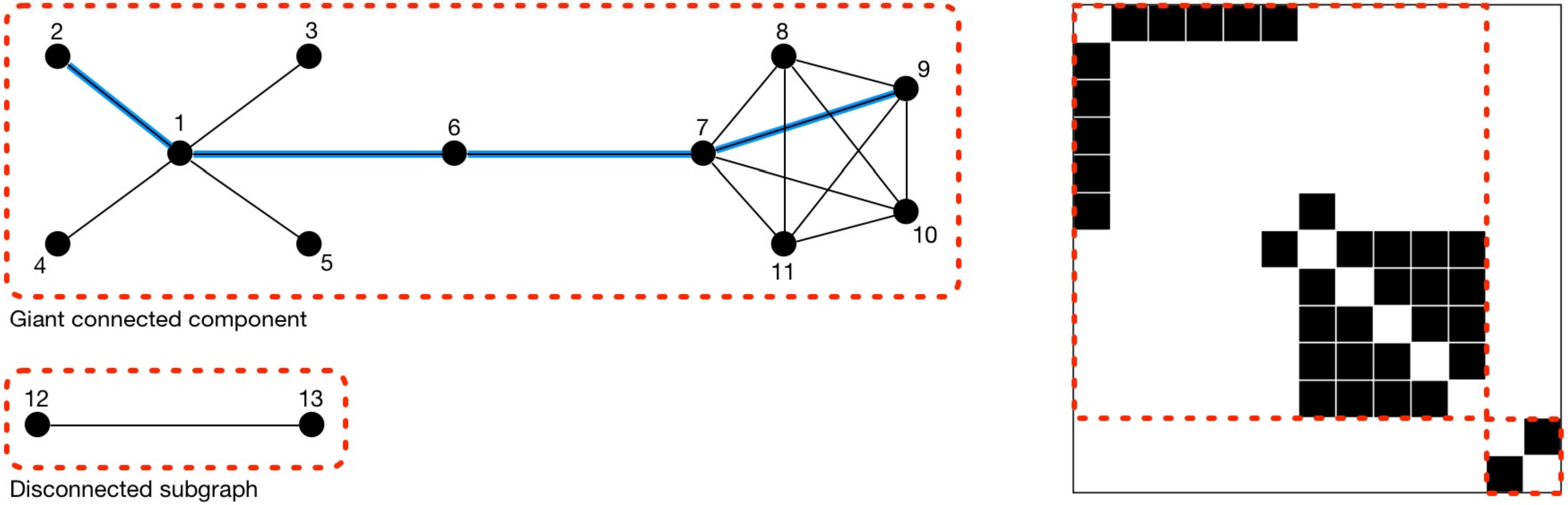
Outline

1. Motivation
2. Truth Discovery from Structured Data
3. Truth Discovery from Extracted Information
4. **Modeling Information Dynamics**
5. Challenges

Misinformation in Networked Systems

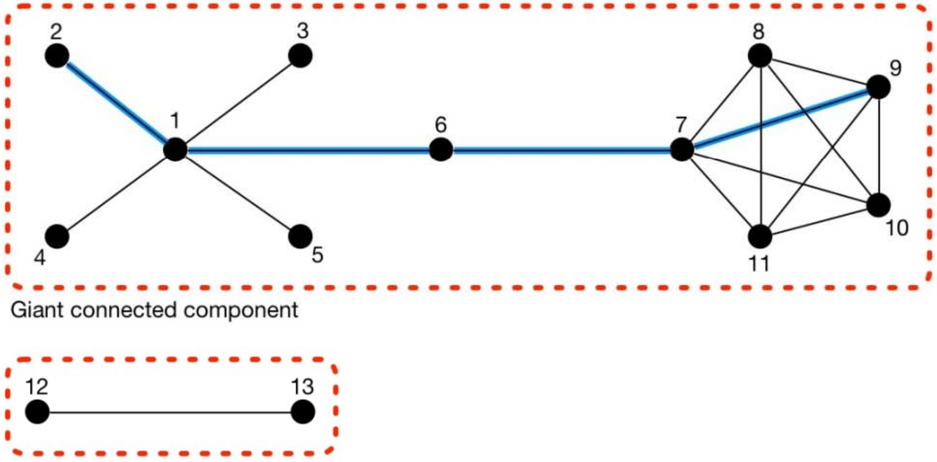


Networked Systems: Topology (I)



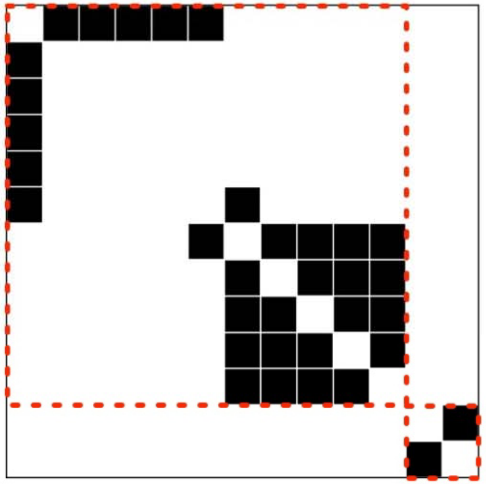
N = 13
L = 17
<*k*> = 1.3
APL = 2.4
D = 4

Networked Systems: Topology (II)



Giant connected component

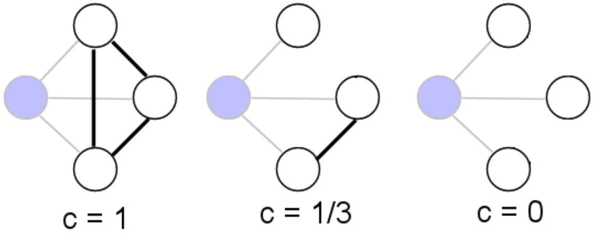
Disconnected subgraph



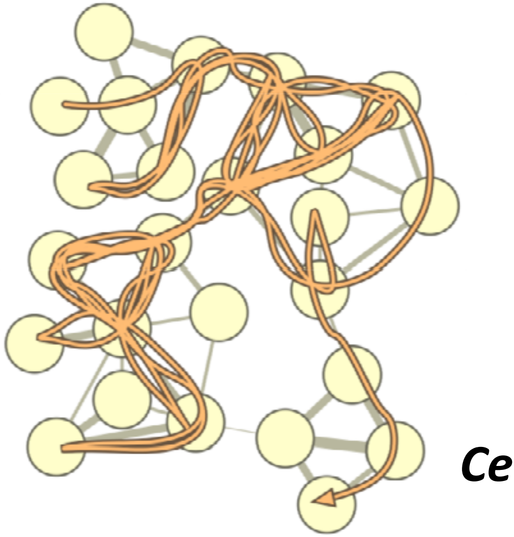
Associated adjacency matrix

N = 13
L = 17
<*k*> = 1.3
APL = 2.4
D = 4

Node →	1	2	3	4	5	6	7	8	9	10	11	12	13
Clustering	0.0	0.0	0.0	0.0	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.0	0.0
Dc	0.4	0.0	0.0	0.0	0.0	0.1	0.4	0.3	0.3	0.3	0.3	0.0	0.0
Bc	0.4	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0
Ec	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.4	0.4	0.4	0.4	0.0	0.0
Core	1	1	1	1	1	1	2	2	2	2	2	1	1



Clustering

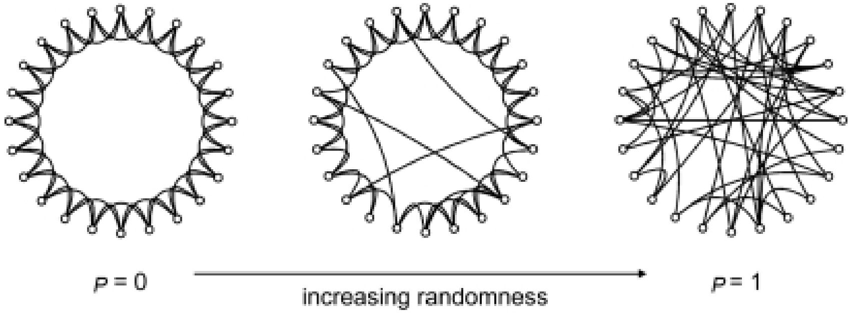


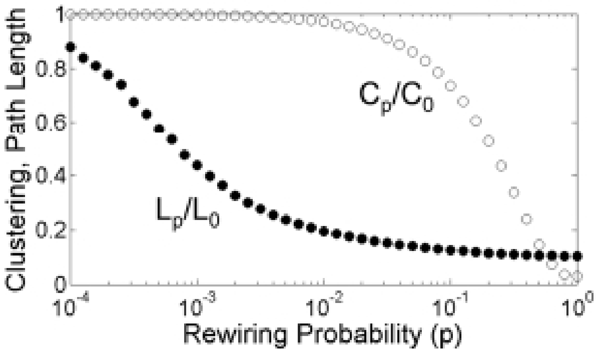
Centrality

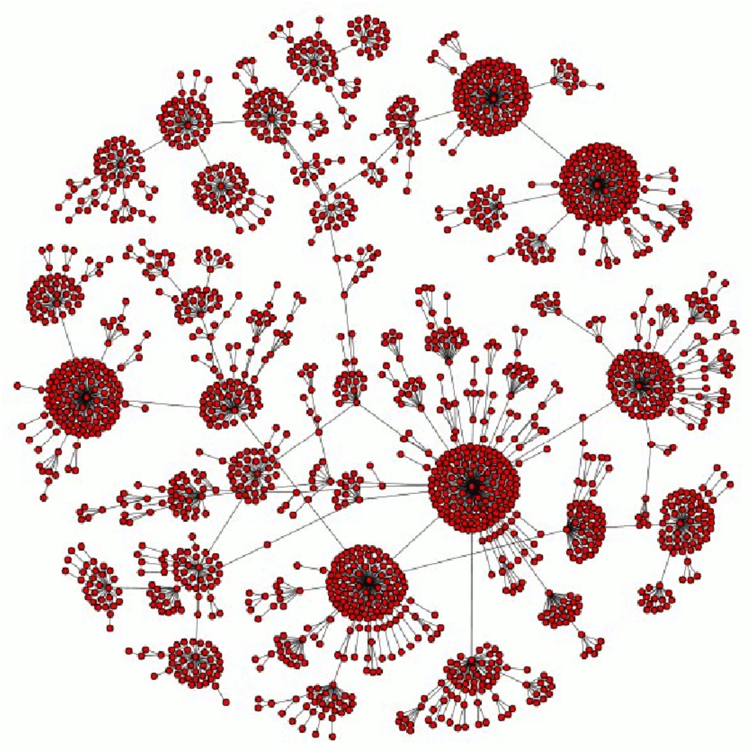
Networks: Why Topology Matters

Take a spreading process…

On which topology is it more efficient?
(faster spread, further reach)

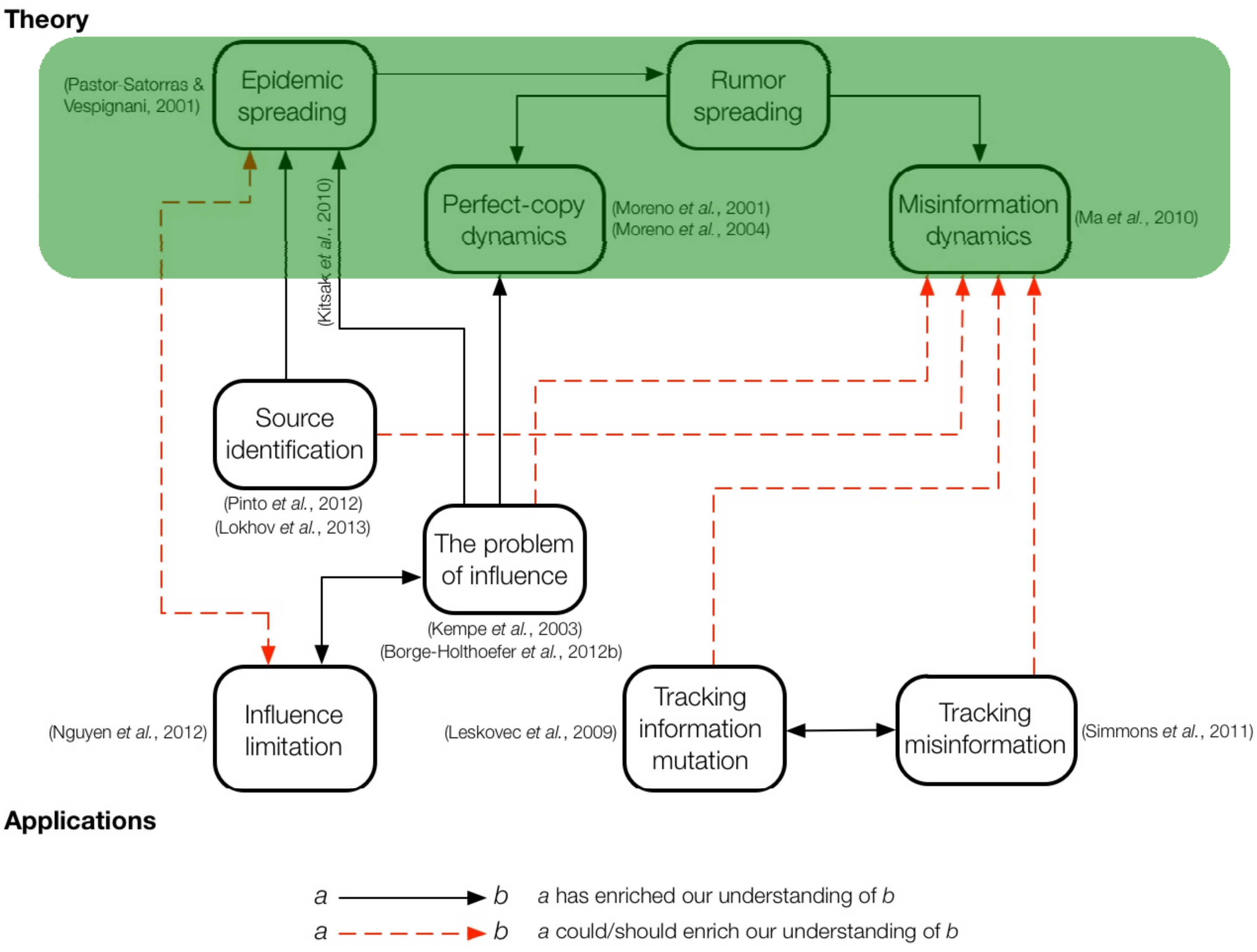




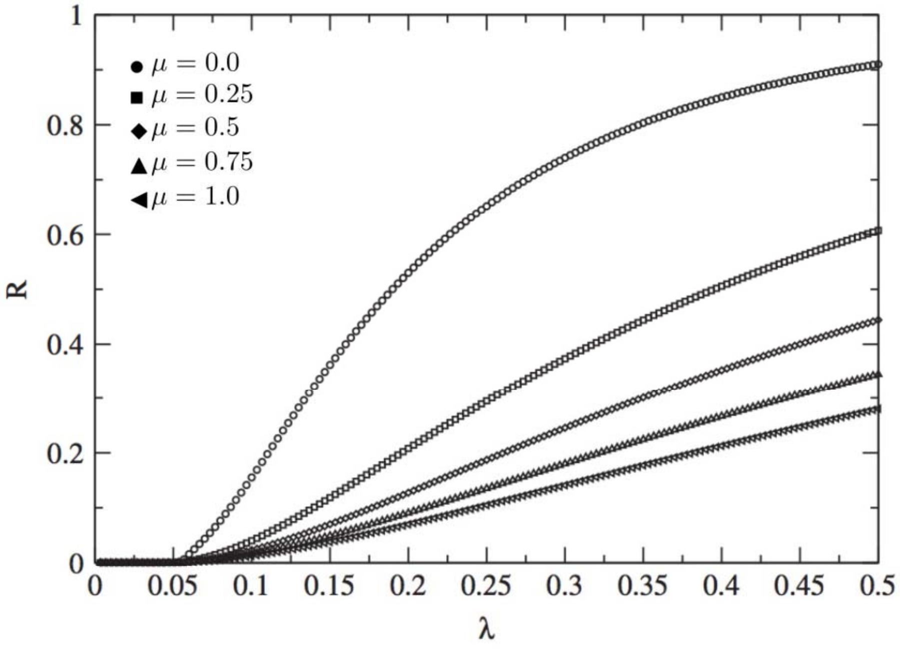


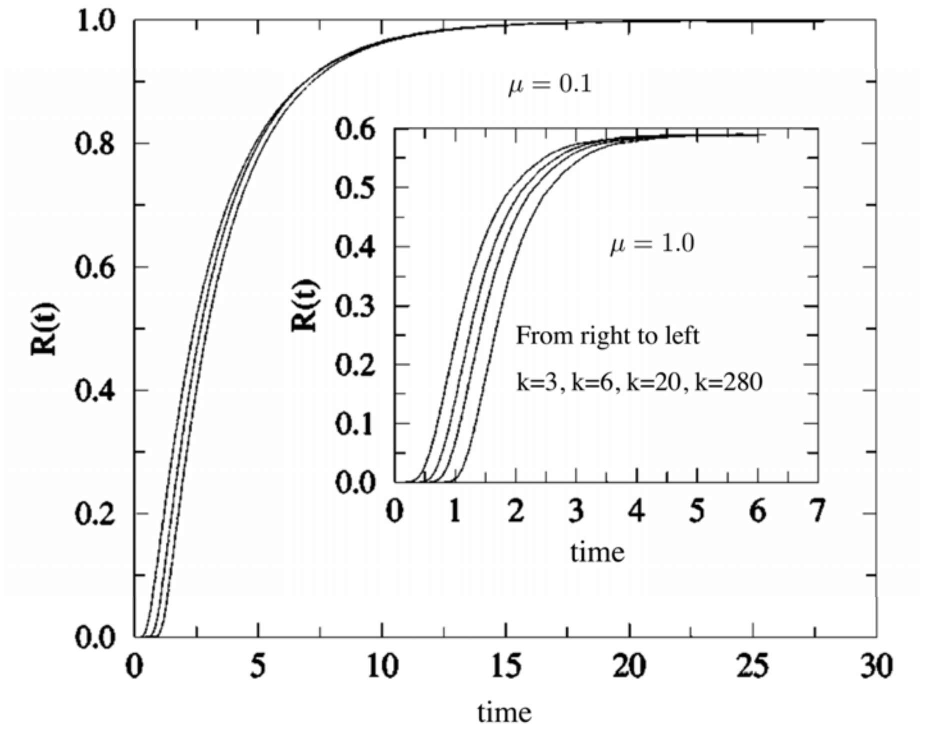
Boccaletti S. et al. (2006) *Complex networks: structure and dynamics*. Physics Reports, 424(4-5), 175–308

Misinformation in Networked Systems



Rumor spreading (I)





$$I \xrightarrow{\lambda} S$$

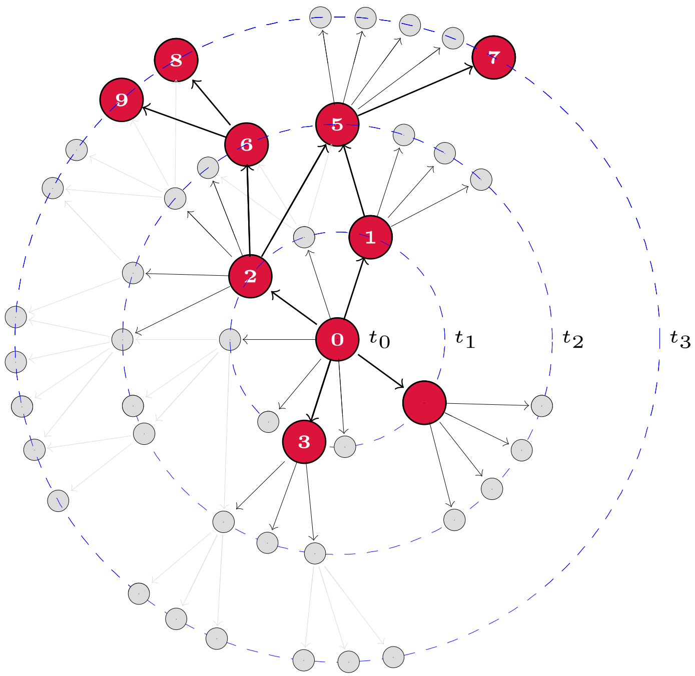
Ignorant to Spreader, with transition probability λ

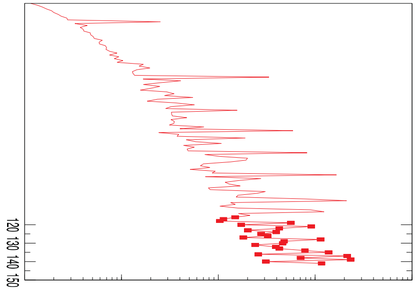
$$S \xrightarrow{\alpha} R$$

Spreader to Stifler, with transition probability α

Moreno Y., Nekovee M. & Pacheco A. (2004) *Dynamics of rumor spreading in complex networks*. Physical Review E, 69(6), 066130

Information Cascades in the Real World





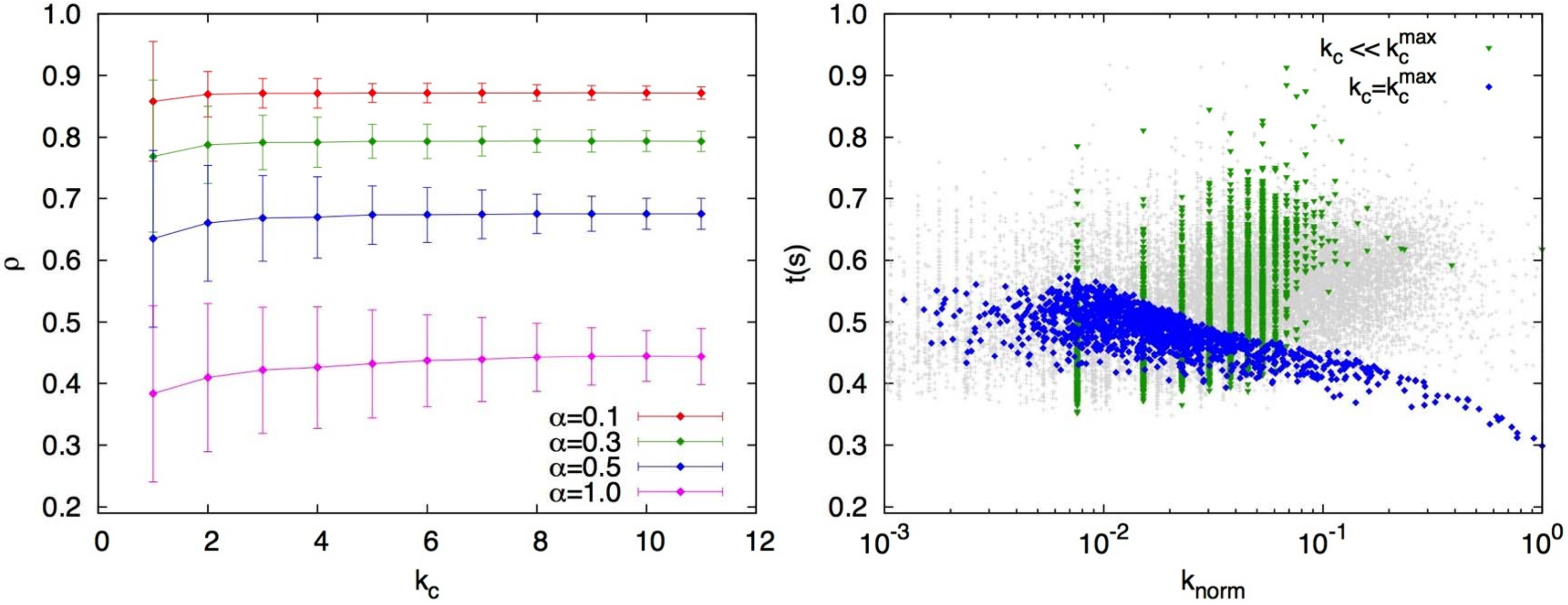
(+) It is possible to observe global cascades like models predict

(?) In real world cases global cascades are mostly achieved from central positions

Gonzalez-Bailon S., Borge-Holthoefer J., Rivero A. & Moreno Y. (2011) The Dynamics of Protest Recruitment through an Online Network. Scientific Reports, 1, 197

Borge-Holthoefer J., Rivero A. & Moreno Y. (2012) Locating privileged spreaders on an online social network. Physical Review E, 85, 066123

Rumor spreading (II)

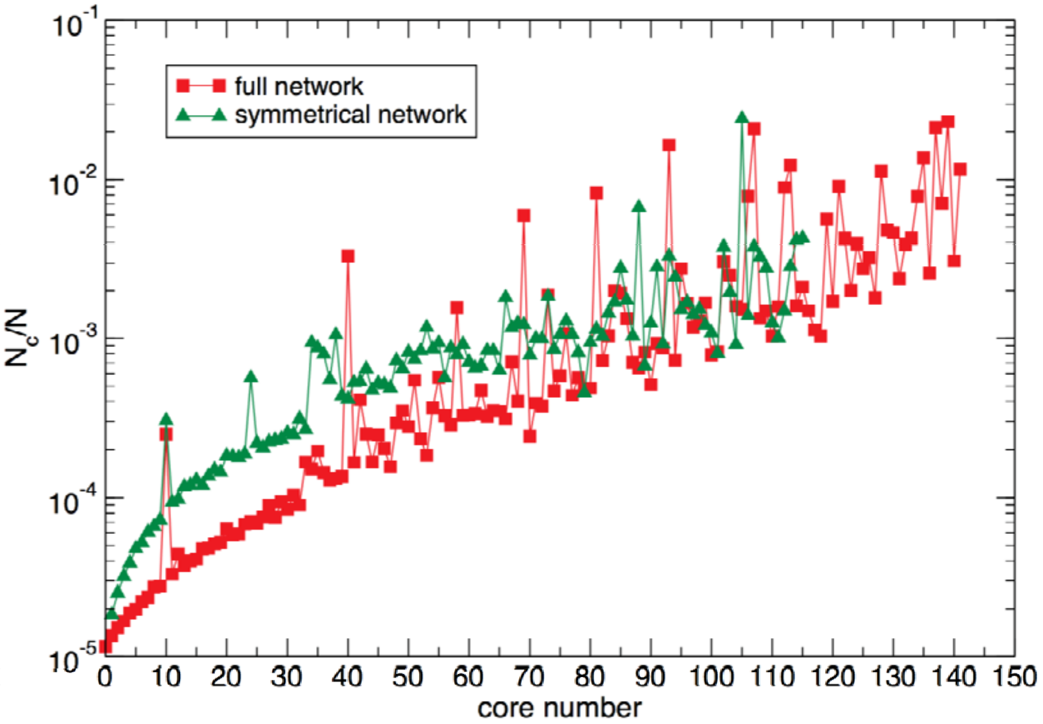


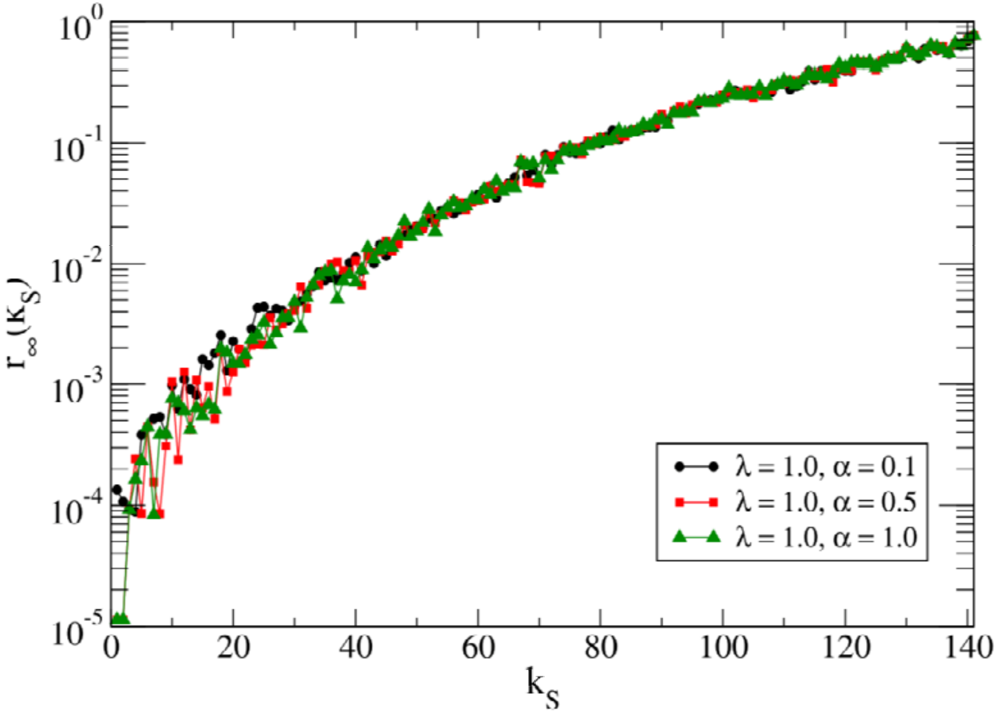
(?) In real world cases, global cascades are mostly achieved from central positions

(-) Classic rumor spreading dynamics do **not** capture the relationship between centrality and cascade success (**rather the opposite**)

Borge-Holthoefer J. & Moreno Y. (2012) *Absence of influential spreaders in rumor dynamics*. Physical Review E, 85, 026116

Evolved Rumor Dynamics: Rates





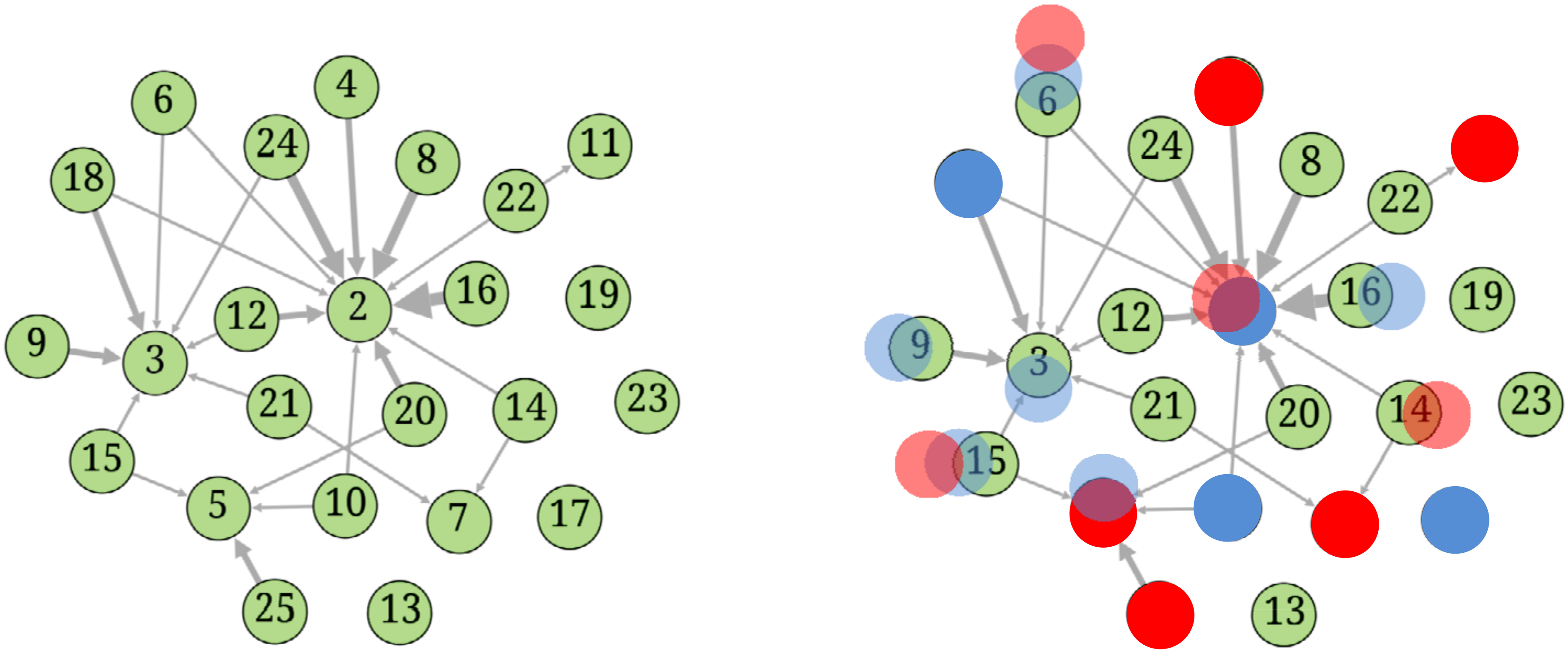
Let each node attempt to spread the rumor at a certain (individual) rate, which depends on its degree *k*

$$a_i = k_i \, / \, k_{max}$$

(-) No matter which refining features we add, information diffusion in the real world is usually **not** exact-copy dynamics

Borge-Holthoefer J., Meloni, S. Goncalves B. & Moreno Y. (2012) *Emergence of influential spreaders in modified rumor models*. Journal of Statistical Physics, 148(6), 1–11

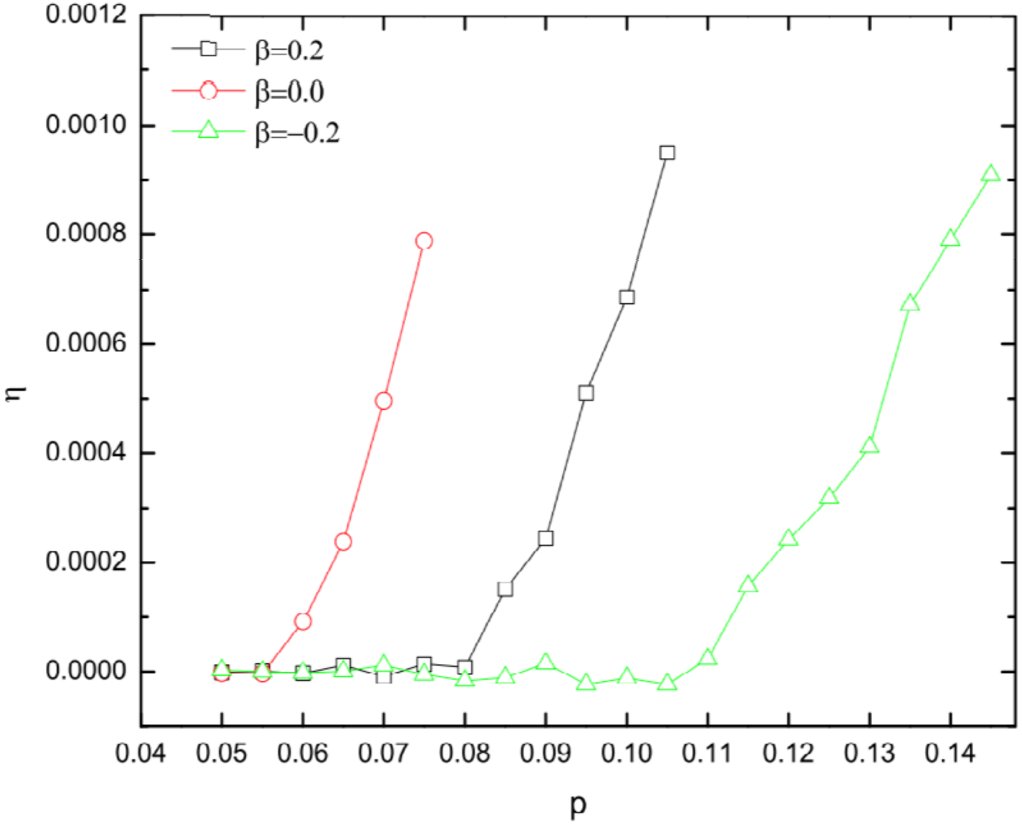
Evolved Rumor dynamics: Mutation (I)

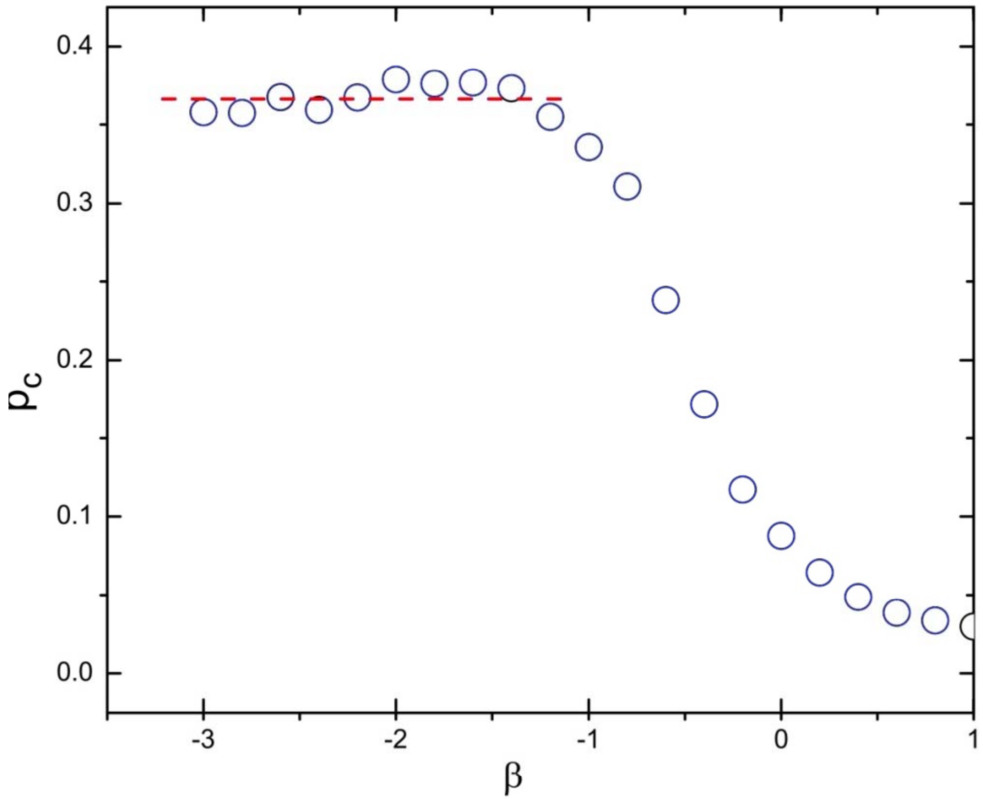


Add to the classical transition probabilities an extra one: the one determining whether information undergoes a **mutation**

Question: at which probability does information **explode**?

Evolved Rumor Dynamics: Mutation (II)



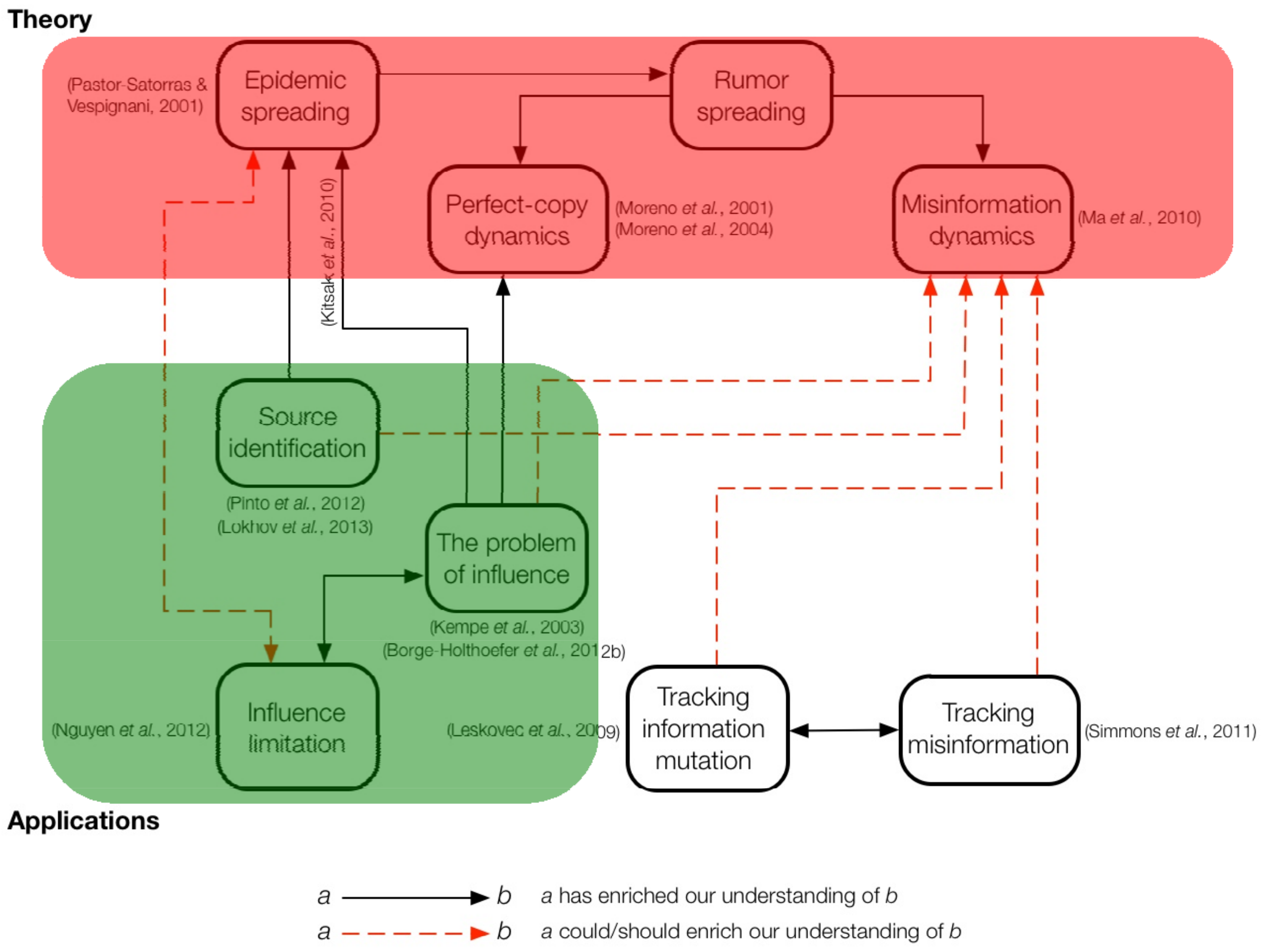


Question: at which probability does information **explode**?

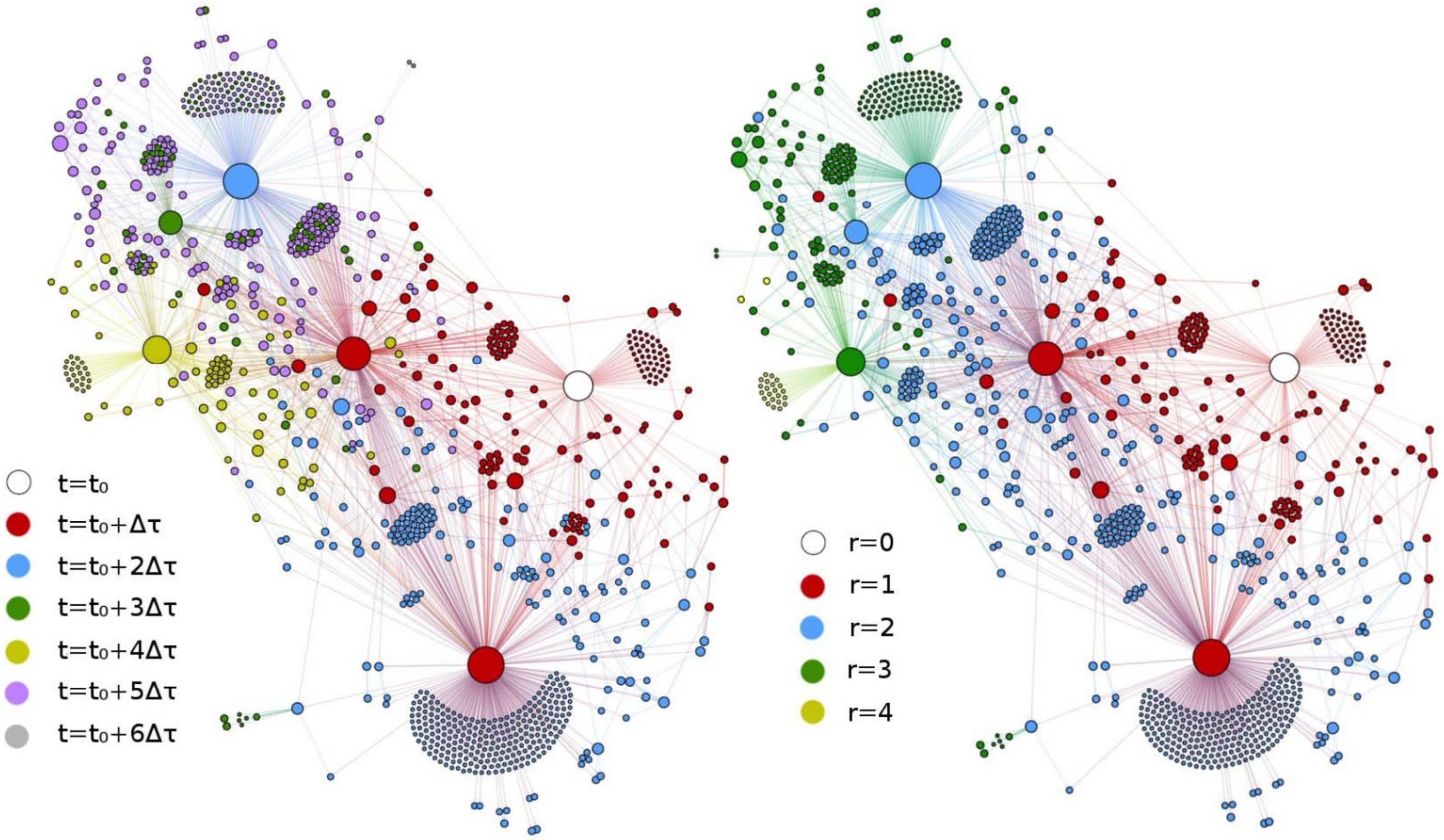
(-) Lack of connection with real world phenomena: no validation. Anyone?

Ma X.J., Wang W.-X., La Y.-C. & Zheng Z. (2010) *Information explosion on complex networks and control*. EPJB 76, 179–183

Misinformation in Networked Systems



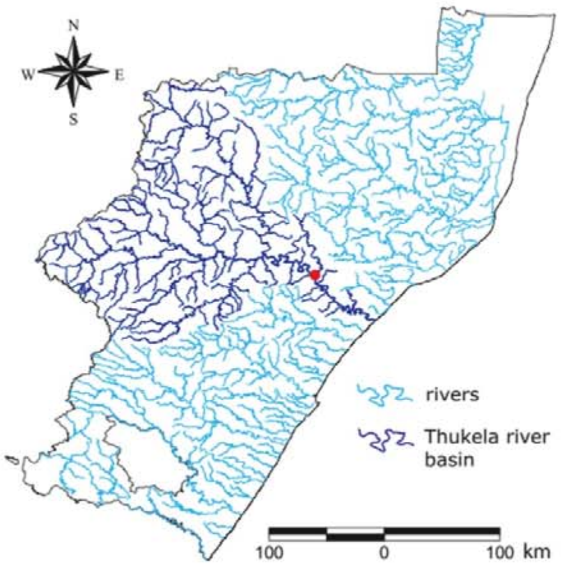
Source identification (I)



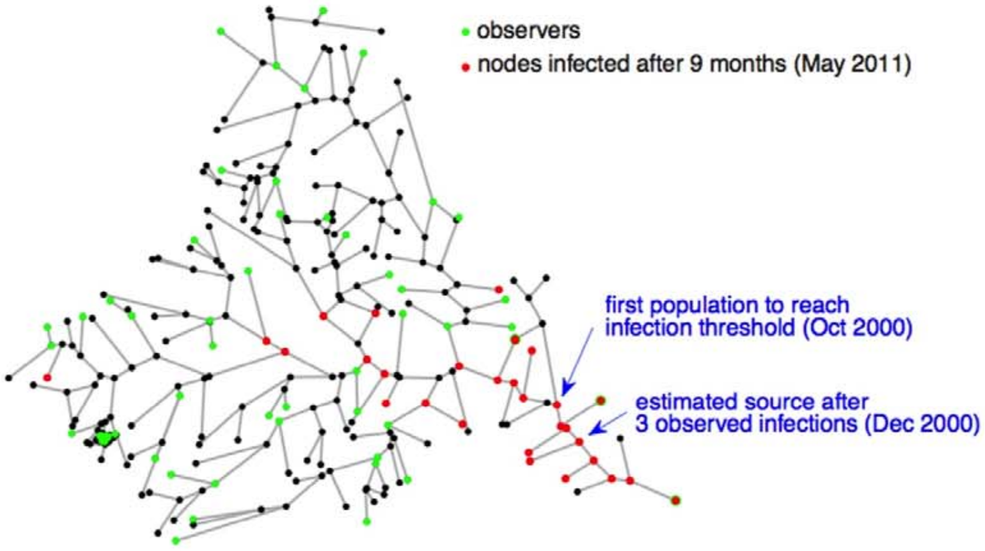
time ***distance***

Baños R.A., Borge-Holthoefer J. & Moreno Y. (2013) *The Role of Hidden Influentials in the Diffusion of Online Information Cascades*. EPJ Data Science, 2:6 doi:10.1140/epjds18

Source identification (II)



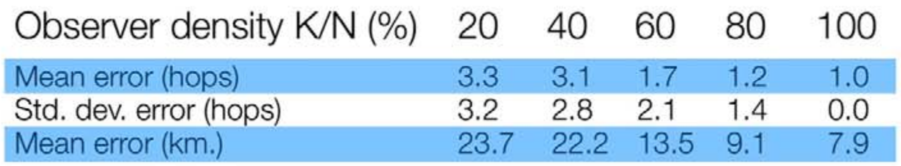
(a)



first population to reach infection threshold (Oct 2000)

estimated source after 3 observed infections (Dec 2000)

(b)

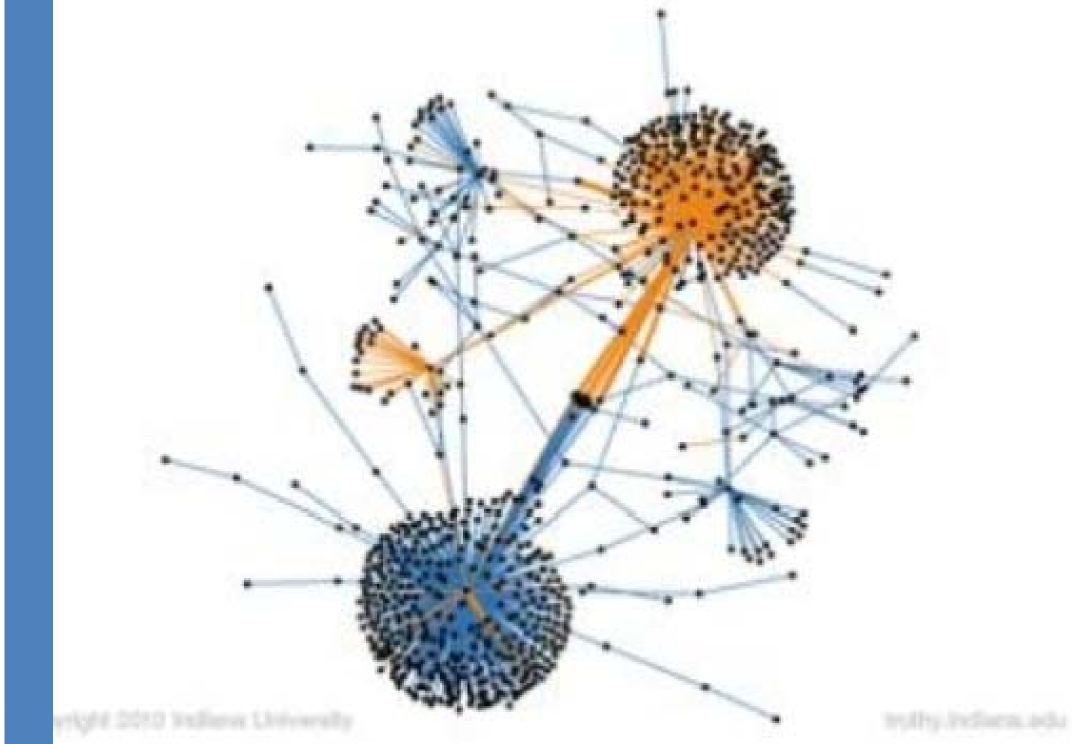


Observer density K/N (%)	20	40	60	80	100
Mean error (hops)	3.3	3.1	1.7	1.2	1.0
Std. dev. error (hops)	3.2	2.8	2.1	1.4	0.0
Mean error (km.)	23.7	22.2	13.5	9.1	7.9

(c)

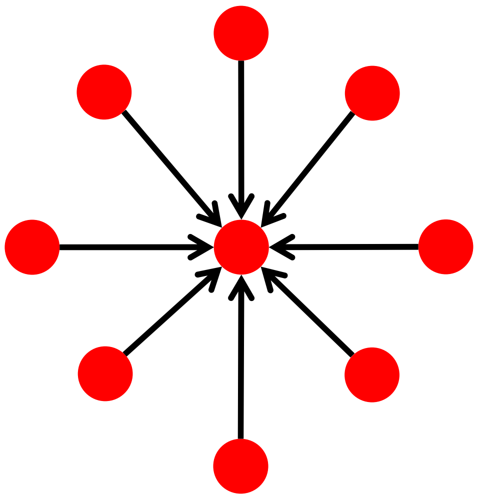
Pinto P., Thiran P. & Vetterli, M. (2012) *Locating the source of diffusion in large-scale networks*. Physical Review Letters, 6(109) 068702

Detect misinformation spreading



the “organic” look

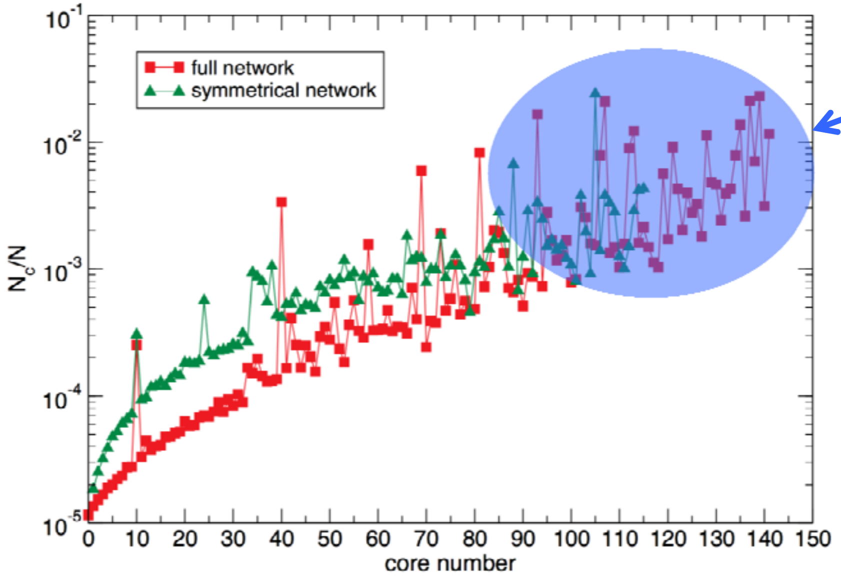
vs.



http://www.truthy.indiana.edu/

Ratkiewicz J. at al. (2011) *Truthy: Mapping the spread of astroturf in microblog streams*. Proceedings of the 20th international conference companion on World Wide Web, 249--252

Stop misinformation spreading

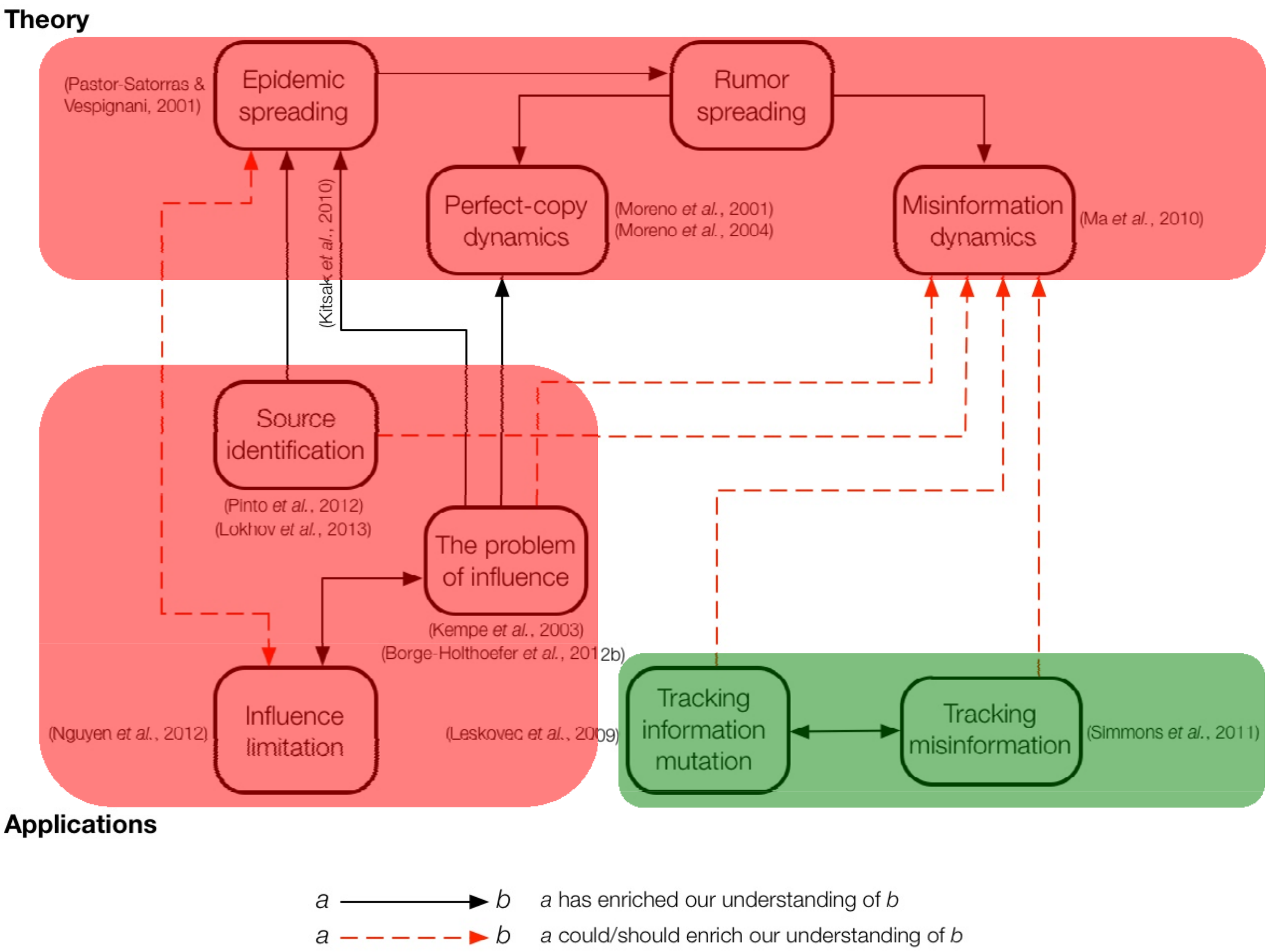


Remember: central nodes (influentials) make a better job controlling cascades

It makes sense to look for those influentials to contain misinformation spreading

Nguyen N.P., Yan G., Thai M.T. & Eidenbenz S. (2012) *Containment of misinformation spread in online social networks*. Proceedings of the 3rd Annual ACM Web Science Conference 213--222

Misinformation in Networked Systems



Meme tracking

it's my belief that this is exactly the time when the american people need to hear from the person will be the next president

↓

this is exactly the time the american people need to hear from the person who in approx 40 days will be responsible for dealing with this mess

↓

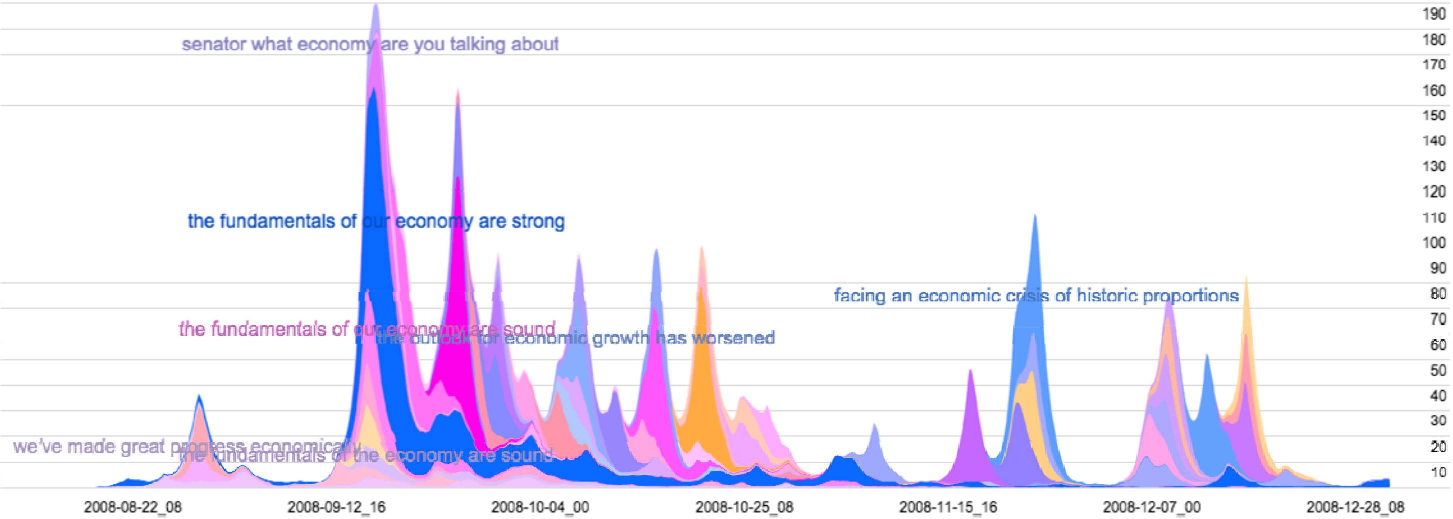
this is exactly the time when the american people need to hear from the person who in approx 40 days will be responsible for dealing with this mess

↓

it's my belief that this is exactly the time the american people need to hear from the person who in approx 40 days will be responsible with dealing with this mess

↓

it's my belief that this is exactly the time the american people need to hear from the person who in approx 40 days will be responsible with dealing with this mess it's going to be part of the president's job to deal with more than thing at once



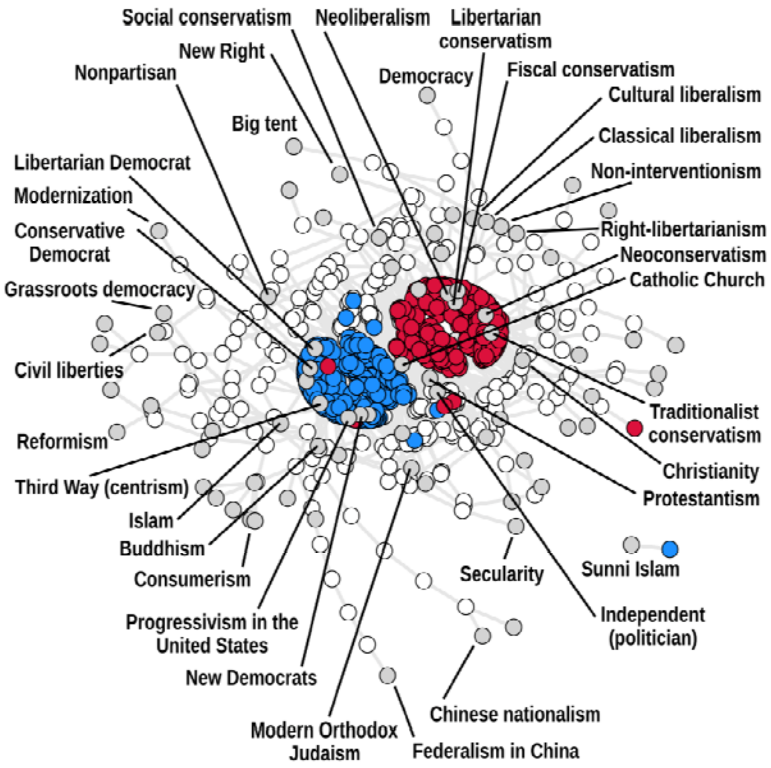
… … … … … … … … … …

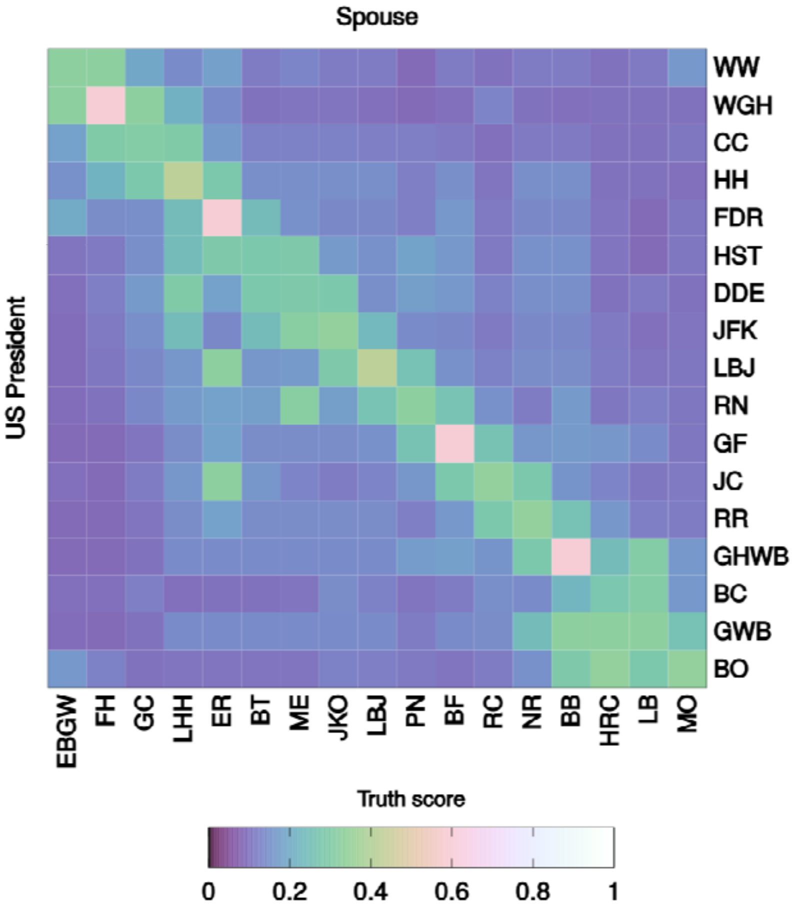
↓

part of the president's job is to deal with more than thing at once in my mind that's more important than ever

Leskovec J., Backstrom, Kleinberg J. (2009) *Meme-tracking and the Dynamics of News cycle*. Proc. 15th SIGKDD, 497-506

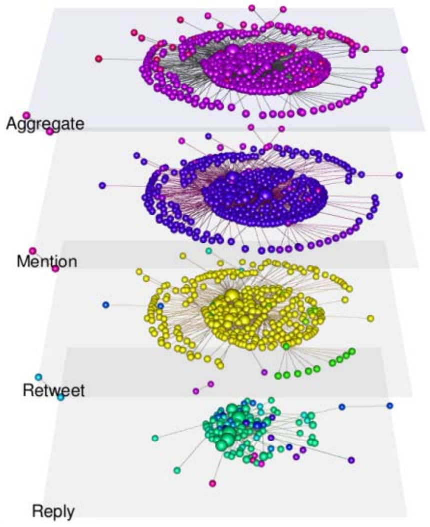
Fact-checking in Knowledge Networks

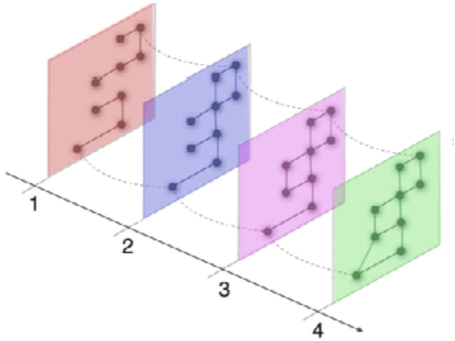


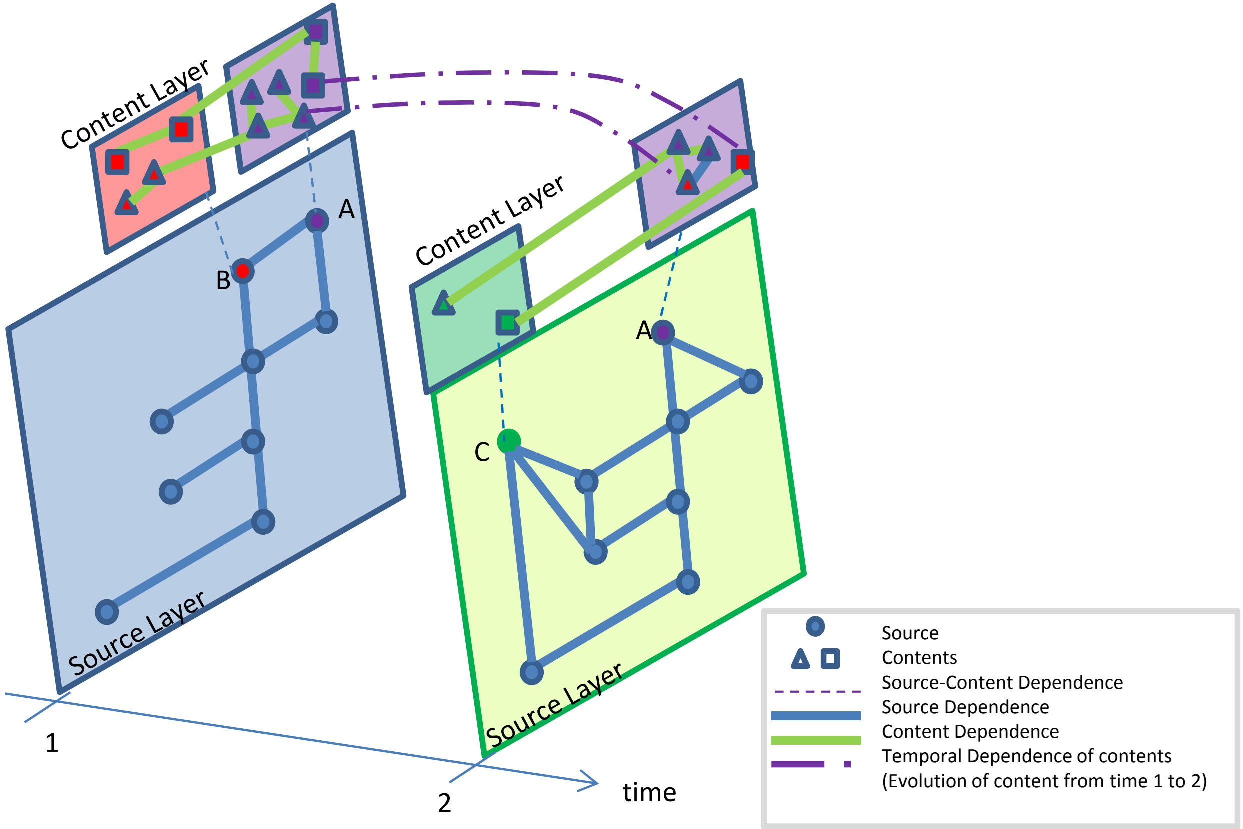


Ciampaglia G.L. et al. (2015) *Computational Fact Checking from Knowledge Networks*. PLoS ONE 10(6): e0128193. doi:10.1371/journal.pone.0128193

Future: Multi-layer Networks







Holme P. & Saramaki J. (2012) *Temporal networks*. Physics reports 519(3) 97--125

Kivela M. et al. (2014) *Multilayer networks*. Journal of Complex Networks, Vol. 2, No. 3: 203-271

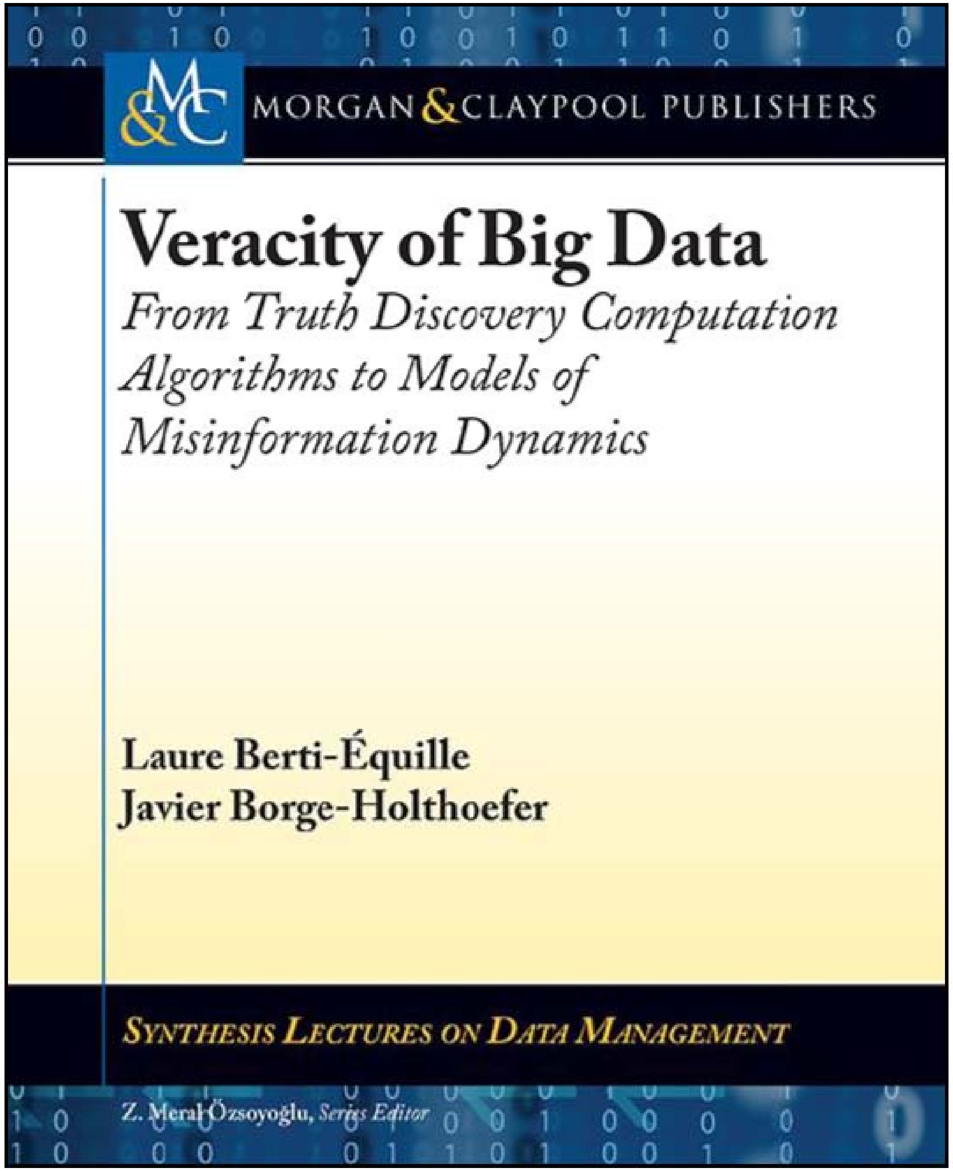
De Domenico M. et al. (2013) *Mathematical formulation of multi-layer networks*. Physical Review X, 3, 041022

De Domenico M., Porter M.A. & Arenas A. (2014) *MuxViz: a tool for multilayer analysis and visualization of networks*. Journal of Complex Networks doi: 10.1093/comnet/cnu038

Summary

- We presented an organized overview of the techniques proposed for truth discovery with recent advances from data/knowledge extraction and complex networks
- Many scientific and technological 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

Further Reading



MORGAN & CLAYPOOL PUBLISHERS

Veracity of Big Data

From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille
Javier Borge-Holthoefer

SYNTHESIS LECTURES ON DATA MANAGEMENT

Z. Meral Özsoyoglu, *Series Editor*

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!