Corroborating Information from Disagreeing Views

Alban Galland\textsuperscript{1}  Serge Abiteboul\textsuperscript{1}  Amélie Marian\textsuperscript{2}  Pierre Senellart\textsuperscript{3}

September 15, 2009, Free University of Bozen–Bolzano
Motivating Example

What are the capital cities of European countries?

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
</tr>
</tbody>
</table>
## Information: redundance

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequence</th>
<th>P. 0.67</th>
<th>R. 0.80</th>
<th>W. 0.60</th>
<th>Buch. 0.50</th>
<th>Bud. 0.43</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R. 0.33</td>
<td>F. 0.20</td>
<td>K. 0.20</td>
<td>Bud. 0.50</td>
<td>S. 0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B. 0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decision</th>
<th>Paris</th>
<th>Rome</th>
<th>Warsaw</th>
<th>?</th>
<th>Sofia</th>
</tr>
</thead>
</table>
Evaluating Trustworthiness of Sources

**Information:** redundancy, trustworthiness of sources (= average frequency of predicted correctness)

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
<th>Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.60</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.58</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.52</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.55</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.51</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.47</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Frequence P. 0.70 R. 0.82 W. 0.61 Buch. 0.53 Bud. 0.46
weighted R. 0.30 F. 0.18 K. 0.19 Bud. 0.47 S. 0.54
by trust  B 0.20

**Decision**

<table>
<thead>
<tr>
<th></th>
<th>Paris</th>
<th>Rome</th>
<th>Warsaw</th>
<th>Bucharest</th>
<th>Sofia</th>
</tr>
</thead>
</table>

P. Senellart (Télécom ParisTech)
Iterative Fixpoint Computation

Information: redundancy, trustworthiness of sources with iterative fixpoint computation

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Poland</th>
<th>Romania</th>
<th>Hungary</th>
<th>Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Paris</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.65</td>
</tr>
<tr>
<td>Bob</td>
<td>?</td>
<td>Rome</td>
<td>Warsaw</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.63</td>
</tr>
<tr>
<td>Charlie</td>
<td>Paris</td>
<td>Rome</td>
<td>Katowice</td>
<td>Bucharest</td>
<td>Budapest</td>
<td>0.57</td>
</tr>
<tr>
<td>David</td>
<td>Paris</td>
<td>Rome</td>
<td>Bratislava</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.54</td>
</tr>
<tr>
<td>Eve</td>
<td>Paris</td>
<td>Florence</td>
<td>Warsaw</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.49</td>
</tr>
<tr>
<td>Fred</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>Budapest</td>
<td>Sofia</td>
<td>0.39</td>
</tr>
<tr>
<td>George</td>
<td>Rome</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Sofia</td>
<td>0.37</td>
</tr>
</tbody>
</table>

| Frequence | P. 0.75 | R. 0.83 | W. 0.62 | Buch. 0.57 | Bud. 0.51 |
| weighted  | R. 0.25 | F. 0.17 | K. 0.20 | Bud. 0.43  | S. 0.49   |
| by trust  |         |         | B 0.19  |           |           |

Decision: Paris, Rome, Warsaw, Bucharest, Budapest
Some Complications

There might be no explicit contradictions between facts stated by different sources:

- “Paris is a city of France.”
- “Lyon is a city of France.”
- “Bolzano is a city of France.”
- ¬ “New York is a city of France.”

We want to exploit the fact that some facts are harder than other (capital of France vs capital of Vanuatu).
Context and problem

- **Context:**
  - Set of sources stating facts
  - (Possible) functional dependencies between facts
  - **Fully unsupervised setting:** we do not assume any information on the truth values of facts or the inherent trust of sources

- **Problem:** determine which facts are true and which facts are false

- **Real world applications:** query answering, source selection, data quality assessment on the web, making good use of the wisdom of crowds
Outline

1 Introduction

2 Model

3 Algorithms

4 Experiments

5 Conclusion
Outline

1 Introduction

2 Model

3 Algorithms

4 Experiments

5 Conclusion
General Model

- Set of facts $\mathcal{F} = \{f_1 \ldots f_n\}$
  - Examples: “Paris is capital of France”, “Rome is capital of France”, “Rome is capital of Italy”

- Set of views (= sources) $\mathcal{V} = \{V_1 \ldots V_m\}$, where a view is a partial mapping from $\mathcal{F}$ to $\{T, F\}$
  - Example:
    - $\neg$ “Paris is capital of France” $\land$ “Rome is capital of France”

- Objective: find the most likely real world $\mathcal{W}$ given $\mathcal{V}$ where the real world is a total mapping from $\mathcal{F}$ to $\{T, F\}$
  - Example:
    - “Paris is capital of France” $\land$ $\neg$ “Rome is capital of France” $\land$
    - “Rome is capital of Italy”
Generative Probabilistic Model

\[
V_i, f_j
\]

- \(\varphi(V_i)\varphi(f_j)\): probability that \(V_i\) “forgets” (does not state anything about) \(f_j\)
- \(\varepsilon(V_i)\varepsilon(f_j)\): probability that \(V_i\) makes an error on \(f_j\) if \(V_i\) makes a statement about \(f_j\)
- **Number of parameters:** \(n + 2(n + m)\) (\(n\) boolean parameters, \(2(n + m)\) parameters between 0 and 1).
- **Size of data:** \(\bar{\varphi}nm\) with \(\bar{\varphi}\) the average forget rate
**Obvious Approach**

- **Method:** use this generative model to find the most likely parameters given the data
  - Inverse the generative model to compute the probability of a set of parameters given the data
  - Standard machine learning technique: Expectation-Maximization

- **Not practically applicable:**
  - Equations for inverting the generative model very complex (but doable)
  - **Large number of parameters** ($n$ and $m$ can both be quite large). Any exponential technique unpractical
  - **Non-linearity** of the model ($W(f_j)$ is boolean)

- $\Rightarrow$ Heuristic fix-point algorithms
**PageRank**

- **PageRank [BP98]**: Fix-point algorithm for computing authority scores on the Web.
- Corresponds to the *equilibrium measure* of the random walk in the (slightly modified) Web graph.
- Can it be applied directly?
  - Sources-Facts: bipartite graph. Random walks (obviously) do not converge in this setting.
  - Alternative: Graph of the two-steps paths in this bipartite graph. Random walks work, but it can be shown that the equilibrium measure is *proportional to the degree* (cf. method Counting further).
  - No clear notion how to manage negative statements (negative links).
- Source of inspiration for the methods presented.
Outline

1. Introduction
2. Model
3. Algorithms
4. Experiments
5. Conclusion
Counting (does not look at negative statements, **popularity**)

\[
T \quad \text{if} \quad \frac{|\{V_i : V_i(f) = T\}|}{\max_f |\{V_i : V_i(f) = T\}|} \geq \eta \\
F \quad \text{otherwise}
\]

Voting (adapted only with negative statements)

\[
T \quad \text{if} \quad \frac{|\{V_i : V_i(f) = T\}|}{|\{V_i : V_i(f) = T \lor V_i(f) = F\}|} \geq 0.5 \\
F \quad \text{otherwise}
\]

**TruthFinder [YHY07]:** heuristic fix-point method from the literature; context slightly different (Source-Object-Fact) and method most adapted to cases with very few errors, does not deal with contradiction
Estimate the truth of facts (e.g., with voting)
Based on that, estimate the error rates of sources
Based on that, refine the estimation for the facts
Based on that, refine the estimation for the sources
... 

Iterate until a fix-point is reached (and cross your fingers it converges!).
Cosine

- The truth of a fact is what views state weighted by how error prone they are.

- The error of a view is the correlation (\(= \text{cosine similarity}\)) between its statement of facts and the predicted truth of these facts.

Precise algorithms are given in [GAMS09].
2-Estimtes

- A fact is true:
  - if a view states it is true and makes no error
  - or if a view states it is false and makes an error
- A view makes an error:
  - if it states a fact is true and the fact is false
  - if it states a fact is false and the fact is true
- Quite instable ⇒ tricky normalization
3-Estimates

- Similar in spirit to 2-Estimates but estimation of 3 parameters:
  - truth value of facts
  - error rate or trustworthiness of sources
  - hardness of facts

- Also needs tricky normalization
So far, the models and algorithms are about positive and negative statements, without correlation between facts.

How to deal with functional dependencies (e.g., capital cities)?

- **pre-filtering:** When a view states a value, all other values governed by this FD are considered stated false.
  
  *If I say that Paris is the capital of France, then I say that neither Rome nor Lyon nor . . . is the capital of France.*

- **post-filtering:** Choose the best answer for a given FD.
Outline

1 Introduction

2 Model

3 Algorithms

4 Experiments

5 Conclusion
Experiments: Generalities

What to measure?

- Quality of binary classification: percentage of error for predicting the truth
- Precision-Recall curve for top-\(k\) rated facts (classical measure for search engine results)

On what data?

- Synthetic dataset closely based upon our generative model, with all possibilities of variation
- Various real-world datasets

We assume that error rates are less than 50%!
Typical Results over Synthetic Dataset

![Graph showing recall and precision over synthetic dataset.](image)

Recall (%) vs. Precision (%) plots for different methods:
- Voting
- Counting
- Cosine
- 2-Estimates
- 3-Estimates

2009/09/15
Champions League: Maccabi Haifa - Bayern Munchen, who will win on 15 Sep?

http://www.hubdub.com/

- Prediction network (sports, politics, business, etc.)
- Bets using virtual money
- (Small) sports dataset extracted: 357 questions, 1 to 20 answers, 473 users, 3,051 statements (before pre-filtering)
<table>
<thead>
<tr>
<th>Method</th>
<th>Number of errors (no post-filtering)</th>
<th>Number of errors (with post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>278</td>
<td>292</td>
</tr>
<tr>
<td>Counting</td>
<td>340</td>
<td>327</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>458</td>
<td>274</td>
</tr>
<tr>
<td>2-Estimates</td>
<td>269</td>
<td>269</td>
</tr>
<tr>
<td>Cosine</td>
<td>357</td>
<td>357</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>272</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>Number of errors (no post-filtering)</td>
<td>Number of errors (with post-filtering)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>Voting</td>
<td>278</td>
<td>292</td>
</tr>
<tr>
<td>Counting</td>
<td>340</td>
<td>327</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>458</td>
<td>274</td>
</tr>
<tr>
<td>2-Estimates</td>
<td>269</td>
<td>269</td>
</tr>
<tr>
<td>Cosine</td>
<td>357</td>
<td>357</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>272</td>
<td>270</td>
</tr>
</tbody>
</table>

Possible to earn money on bets. Easy way to get rich!
1. Where is the city of Ushuaia located?
- Don't know
- In Italy
- In Greece
- In Argentina
- In the Ivory Coast
- In Sweden
- In Malaysia

2. What is the last word of all three parts of Dante’s *Divine Comedy* (*Hell — Purgatory — Paradise*)?
- Don’t know
- “Stars” (“Stelle”)
- “God” (“Dio”)
- “Hope” (“Speranza”)
- “Beatrice”

3. Who discovered the planet Uranus?
- Don’t know
- Sir William Herschel (in 1781)
- Urbain Le Verrier (in 1846)
- Clyde Tombaugh (in 1930)
- Percival Lowell (in 1894)

http://www.madore.org/~david/quizz/quizz1.html

- 17 questions, 4 to 14 answers, 601 participants
### General-Knowledge Quiz: 2/2

<table>
<thead>
<tr>
<th></th>
<th>Number of errors (no post-filtering)</th>
<th>Number of errors (with post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Counting</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-Estimates</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Cosine</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Possible to know the correct answer to a quiz by just looking at all answers. Automatic correction of exams is possible!
### General-Knowledge Quiz: 2/2

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of errors (no post-filtering)</th>
<th>Number of errors (with post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Counting</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2-Estimates</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Cosine</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Possible to know the correct answer to a quiz by just looking at all answers. Automatic correction of exams is possible!
It does not always work!

No magic!

- Does not take into account dependencies between sources
- Example: integration of search engine results
- Usually, when it “does not work”, 3-Estimates gives results comparable to the baseline, Cosine is not bad, 2-Estimates is very unstable
In brief

- One of the first works in truth discovery among disagreeing sources
- Collection of fix-point methods, one of them (3-Estimates) performing remarkably and regularly well
- We believe this is an important problem, we do not claim we have solved it completely
- Cool real-world applications!

All code and datasets available from http://datacorrob.gforge.inria.fr/. Details in [GAMS09].
Exploiting dependencies between sources [DBES09]

Numerical values (1.77m and 1.78m cannot be seen as two completely contradictory statements for a height)

No clear functional dependencies, but a limited number of values for a given object (e.g., phone numbers)

Pre-existing trust, e.g., in a social network

Clustering of facts, each source being trustworthy for a given field
Merci.
Sergey Brin and Lawrence Page.  
The anatomy of a large-scale hypertextual Web search engine.  

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava.  
Integrating conflicting data: The role of source dependence.  
In *Proc. VLDB*, Lyon, France, August 2009.

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart.  
Corroboration de vues discordantes fondées sur la confiance.  
In *Proc. BDA*, Namur, Belgium, October 2009.  
Conference without formal proceedings.
Xiaoxin Yin, Jiawei Han, and Philip S. Yu.  
Truth discovery with multiple conflicting information providers on the Web.  
In *Proc. KDD*, San Jose, California, USA, August 2007.