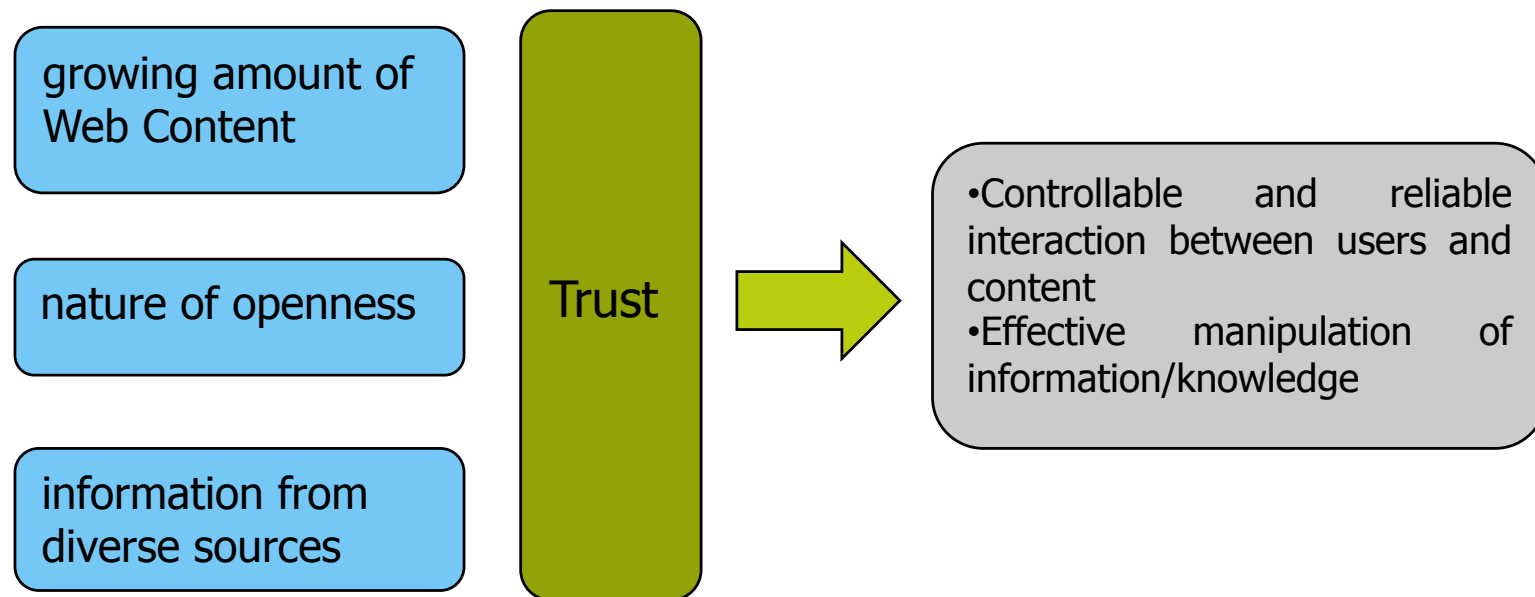




Towards Trust in Web Content Using Semantic Technologies

Qi Gao
Web Information Systems
Delft University of Technology, the Netherlands

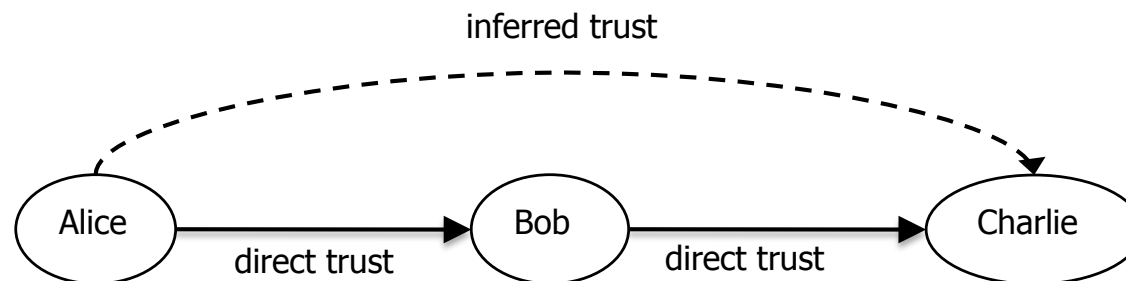
Motivation



Social trust

Social Trust

- Defined in the context of Social Networks
- Building trust within networks of people, agents or peers



Content trust

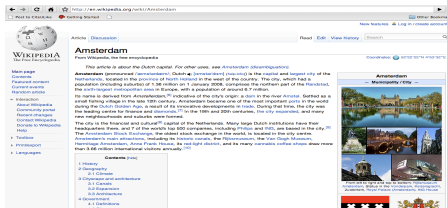
Content Trust

- Metrics to represent and assess the trustworthiness of Web content
- “A trust judgment on a particular piece of information in a given context”
- determined by many factors

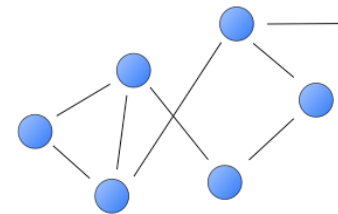
authors

```
• [cur | prev] 21:15, 22 May 2010 Mathpian93 (talk | contribs) m (104,582 bytes) (Revert to revision Mathpian93 using revision) (undo)
• [cur | prev] 21:02, 22 May 2010 94.193.135.203 (talk) (104,585 bytes) (anon-envaders-exxxxxxx)
• [cur | prev] 22:04, 21 May 2010 Mathpian93 (talk | contribs) (104,582 bytes) (←Geography: mor)
• [cur | prev] 00:41, 21 May 2010 Rvanbertum (talk | contribs) m (104,858 bytes) (←Symbol: Beth)
• [cur | prev] 03:53, 20 May 2010 18.111.107.226 (talk) (104,864 bytes) (←Government) (undo)
• [cur | prev] 23:19, 19 May 2010 75.195.32.52 (talk) (104,911 bytes) (←Government) (undo)
• [cur | prev] 22:59, 18 May 2010 67.49.119.234 (talk) (104,904 bytes) (←History) (undo)
• [cur | prev] 22:58, 18 May 2010 67.49.119.234 (talk) (104,903 bytes) (←History) (undo)
• [cur | prev] 07:32, 18 May 2010 Cydebot (talk | contribs) m (104,864 bytes) (Robot - Moving categ to Populated places established in the 19th century per CFD at Wikipedia:Categories for discussion/Log/2)
• [cur | prev] 14:24, 17 May 2010 147.162.5.50 (talk) (104,859 bytes) (Undo revision 362004959 by 14.24.17 May 2010 147.162.5.50 (talk) (104,972 bytes) (Undo revision 362004255 by 13.00.17 May 2010 208.67.34.90 (talk) (104,897 bytes) (←Expansion) (undo)
• [cur | prev] 12:49, 17 May 2010 208.67.34.90 (talk) (104,872 bytes) (←Expansion) (undo)
• [cur | prev] 12:17, 12 May 2010 The Thing That Should Not Be (talk | contribs) m (104,859 bytes) (revision by EdFast (H3) (Custom)) (undo)
```

Wikipedia articles (content)



Trust network of authors



Factors that influence the trust

- Topic
- Provenance
- preference

...

“Is a Wikipedia article trustworthy?”

“Is a paragraph or statement from an article trustworthy?”

Problem Statement– research questions

- 1. What factors can influence content trust?
- 2. How to capture information about these factors?
- 3. How to compute trust in Web content?
 - information about content itself
 - semantic similarity or relation between different pieces of content

State-of-the-art (content trust)

TRELLIS(2002)

- trust in statements based on provenance and related statement

- IWTrust(2005) <http://inference-web.org/>

- trust in answers based on sources and users

- tRDF(2008) <http://trdf.sourceforge.net/>

- Trust in RDF

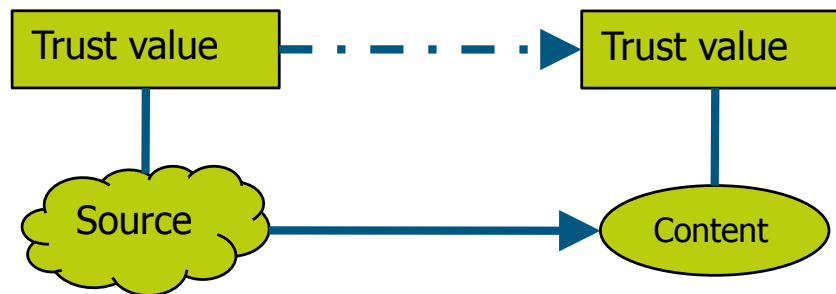
- W3C provenance incubator group(2009)

- <http://www.w3.org/2005/Incubator/prov/>

- Trust is an important usage of provenance on the Web

Key ingredient for trust - provenance

- Source-level approach



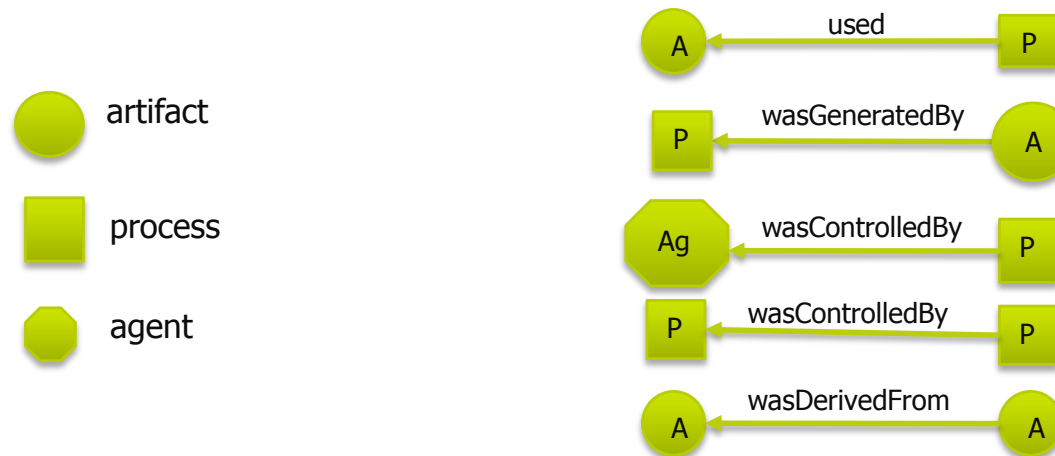
coarse-grained
and insufficient

- The provenance of Web content
 - Where did the content come from (original sources) ?
 - How was a piece of content produced?
 - Who was involved in the process?

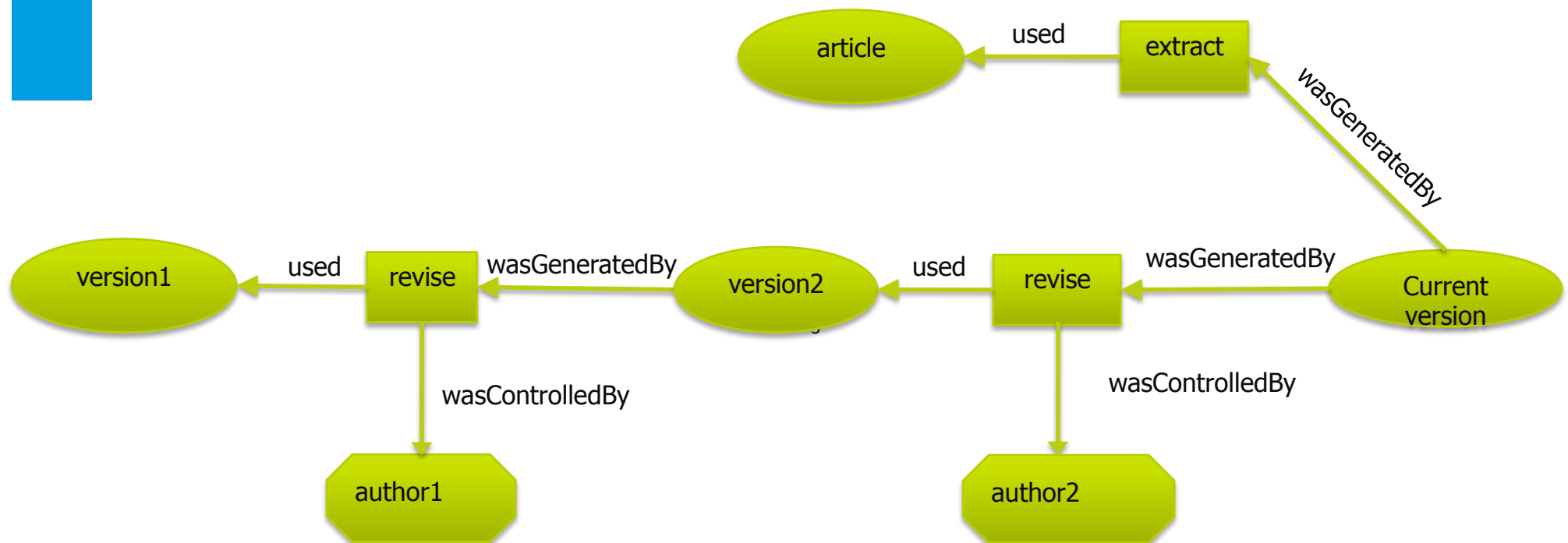
fine-grained

Provenance on the Wikipedia?

- A paragraph from an Wikipedia article as a content unit
- Provenance information
 - Which article is a paragraph from?
 - What is the revision history of this paragraph?
 - Who are the authors of this paragraph?
- Capture and represent the provenance: Open Provenance Model



Capture and represent the provenance information



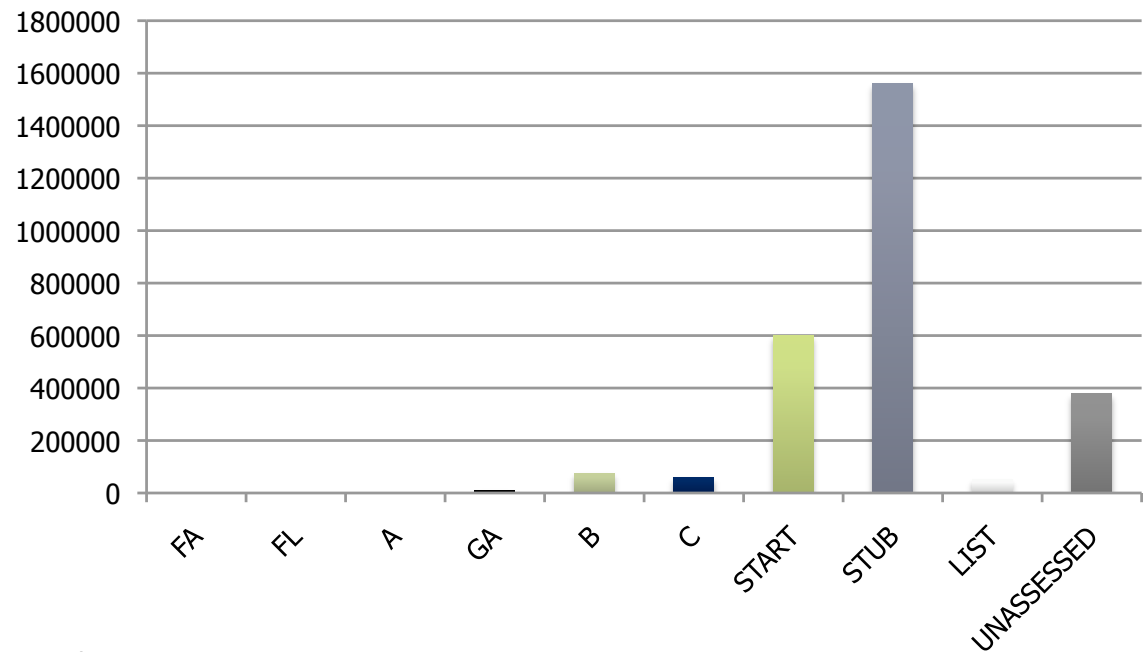
Trust Computation – Basic Principles

- Two issues for computing trust
 - (1) obtain the initial trust values
 - (2) propagate the trust values
- With the help of the semantic similarity or relation between different pieces of content
 - Hypothesis: the trust can spread among similar or related pieces of content
 - The similarity is measured based on the provenance information and content itself

A use case study- Wikipedia

- Wikipedia
 - Nature of openness and collaboration
 - Complex process (revision history) that produced the content
- Is Wikipedia trustworthy ?

Quality Class	Number*	Percentage
FA	3108	0.11%
FL	1392	0.05%
A	734	0.03%
GA	9164	0.33%
B	74785	2.73%
C	60129	2.19%
START	601925	21.96%
STUB	1560244	56.91%
LIST	51843	1.89%
UNASSESSED	378287	13.80%
Total	2741611	100.00%



*Retrieved on 17th May, 2010

**http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

Capture the provenance information

- For the experiment, we randomly select a sample set of Wikipedia articles from different quality classes.

Func1 Paragraph Identifier *ParaIden(p_j)*

Input: a wikipedia article A_i

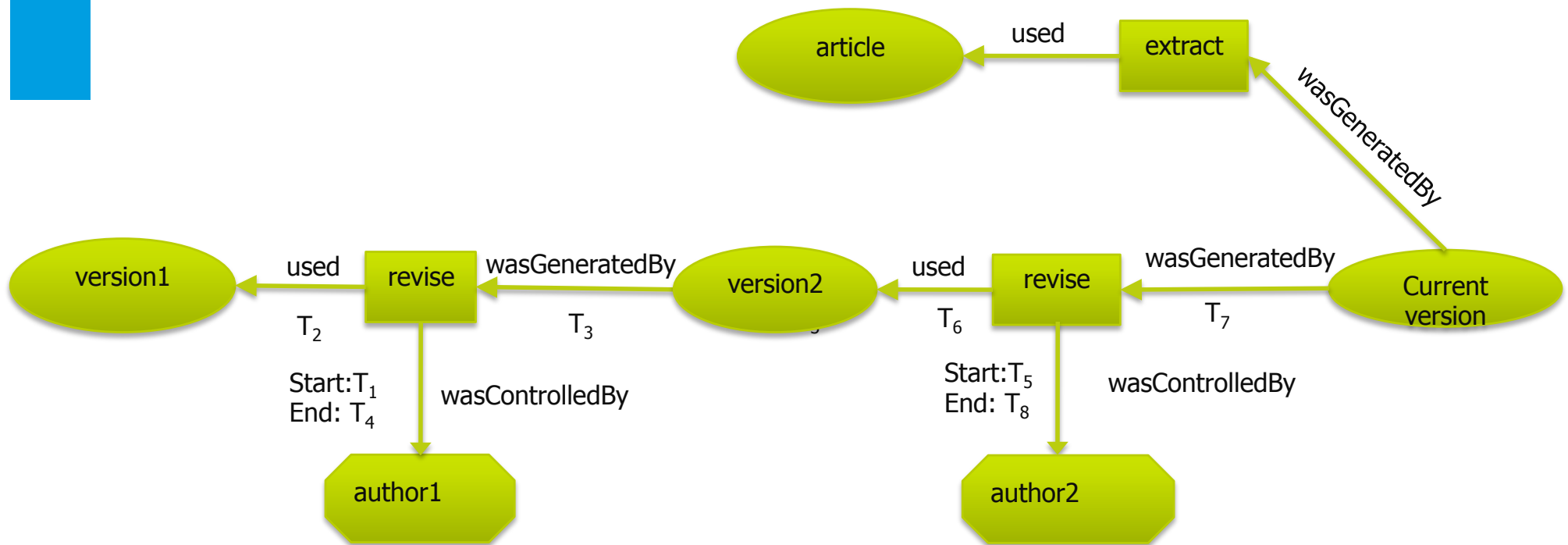
Output: a set of paragraphs in A_i , $P_{A_i}=\{p_j\}$, a set of authors who has revised the paragraph p_j and the timestamp $PF_{p_j}=\{(au_k, time_k)\}$

Func2 produce a provenance graph for each paragraph *PGProduce(p_j, AF_{fj})*

Input: a paragraph p_j , $PF_{p_j}=\{(au_k, time_k)\}$

Output: a provenance graph for p_j , PG_{p_j}

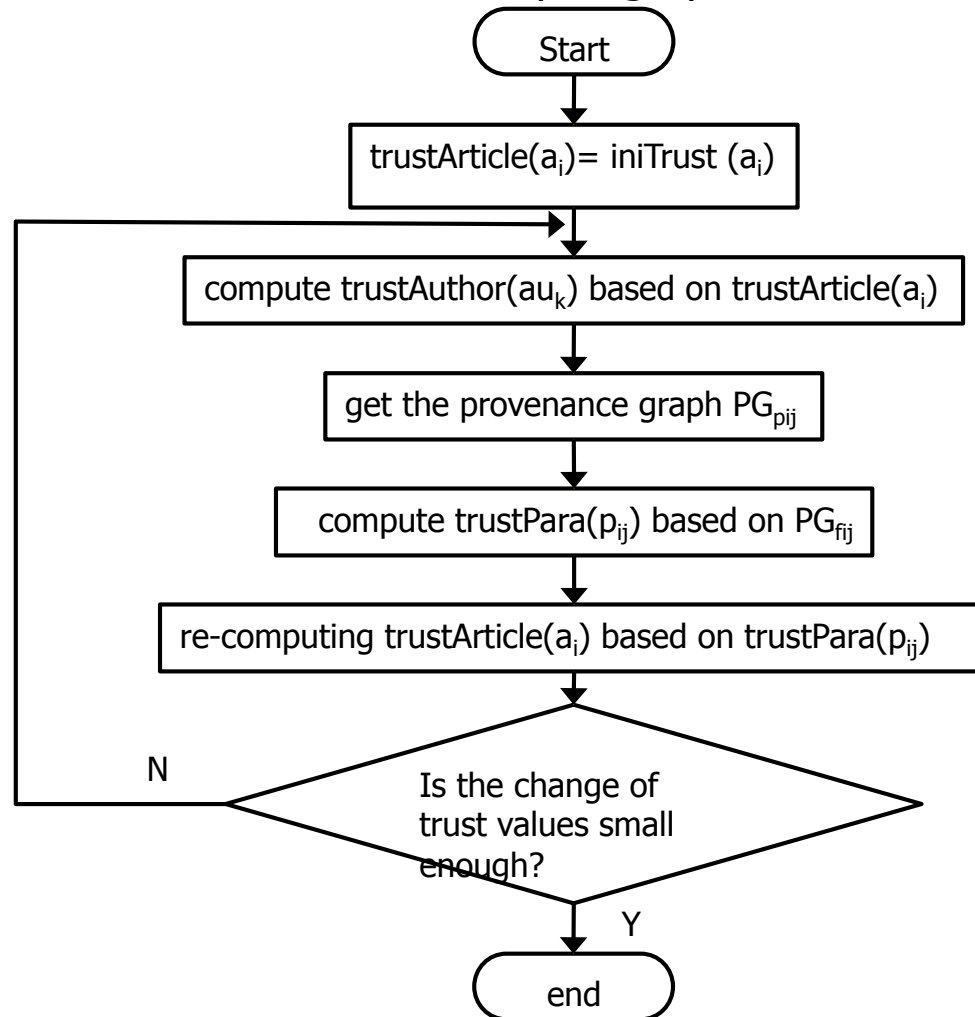
Capture the provenance information



Trust computation

Input: a set of initial trust values for all the articles $IT = \{ \text{iniTrust}(a_i) \}$, a set of paragraph in a_i $P = \{ p_{ij} \}$, a set of authors who has revised any articles $\text{Author} = \{ au_k \}$

Output: stable trust value for each paragraph, author and article,



Trust computation - semantic similarity

- Hypothesis: an author on Wikipedia can be an expert on one topic and not an expert on another topic.
- Cluster C_i is the collection of articles in a topic based on the semantic similarity

$\text{trustAuthor}(au_k) \longrightarrow \text{trustAuthor}(au_k, C_i)$

$\text{trustPara}(p_{ij}) \longrightarrow \text{trustPara}(p_{ij}, C_i)$



Evaluation

- Use an existing testbed/benchmark
 - Comparing the results between the trust computation and the Wikipedia quality assessment study
- Conduct user studies



Wrapping up

- Conclusions
 - provenance as the key ingredient for deriving trust
 - trust computation utilizing provenance information & semantic technologies
- Future work
 - other scenarios
 - other factors: context, user preference



Thank you!

Qi Gao

q.gao@tudelft.nl

[http://www.wis.ewi.tudelft.nl/
index.php/members/qi-gao](http://www.wis.ewi.tudelft.nl/index.php/members/qi-gao)