VIS 2016

NameClarifier: A Visual Analytics System for Author Name Disambiguation

Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu and Weiwei Cui

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Microsoft Research
微软亚洲研究院

Name ambiguity

by

Wang Wei

？？？？？

王伟，王维，王威，王玮，汪卫， 汪伟， 汪威

Name ambiguity: People with different names

[WEI WANG]

VIS 2016

Small scale library

Manual Check

➡ Library in Universities(No error allowed)

Automatic approach

➡ Large public bibliography database

(A small number of errors are allowed)

- Purely author names.
- Publication attributes: titles, shared coauthors, venues, self-citation. Etc.
- Additional web information.

Major challenge 1/2:

The name ambiguity problem are **case by case**.

- Limited collaborators or wide range of collaborators
- One research interest or multiple research interests

Major challenge 2/2:

Uncertainties of every attribute:

- Venue cover different size of scopes (IJCV for Vision V.S. TVCG for computer graphics + visualization)
- Shared coauthors: suffer from name ambiguity themselves!
- Etc.

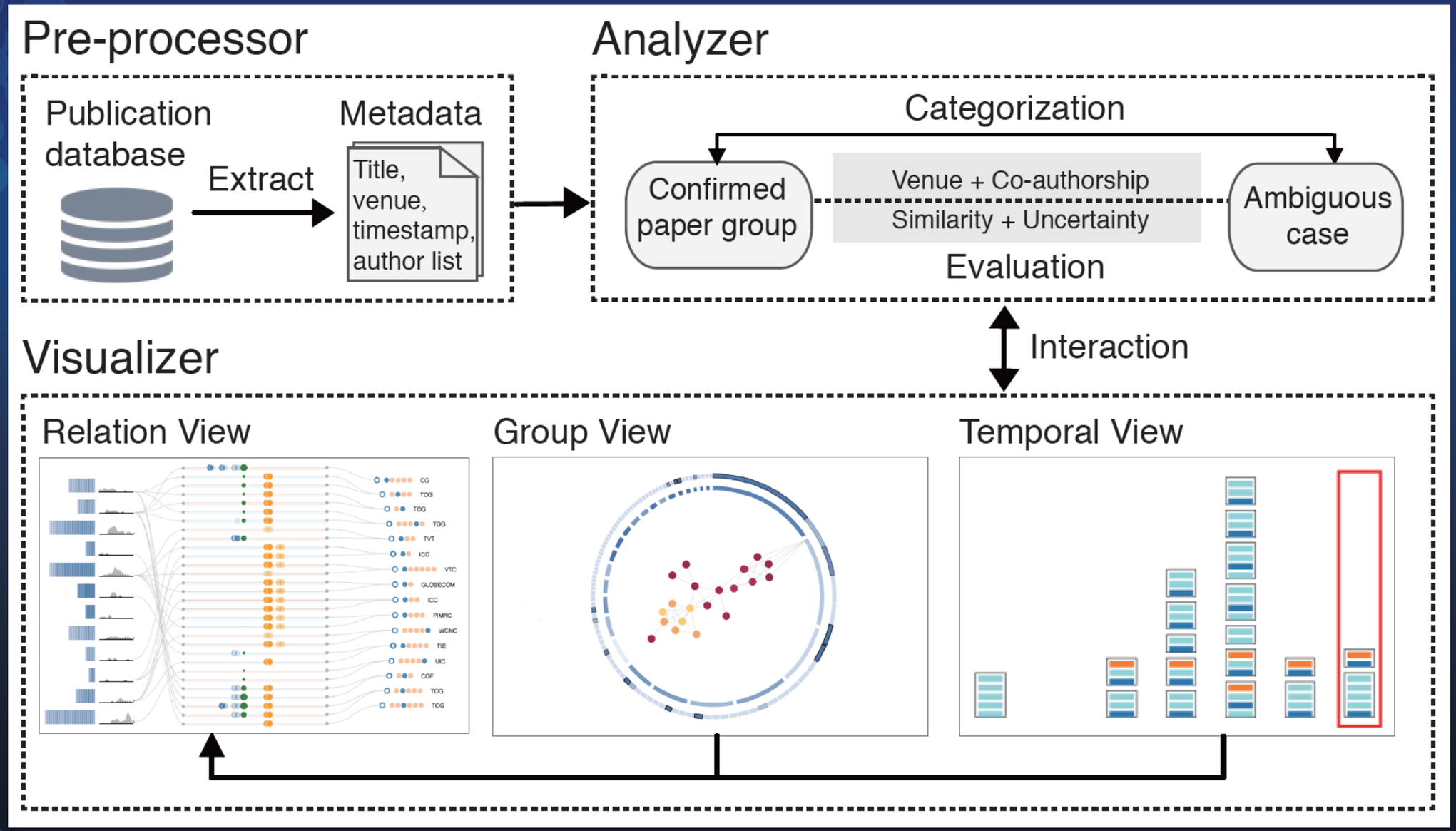
➡ No universal model

Get people involved

Our solutions:

- Customize the disambiguation on a case-by-case basis
- Mining metrics + visualization
- Traditional black box solution -> white box procedure

System framework


Pre-processor
Publication database
Extract
Metadata
Title, venue, timestamp, author list
Analyzer
Categorization
Confirmed paper group
Venue + Co-authorship
Similarity + Uncertainty
Ambiguous case
Evaluation
Interaction
Visualizer
Relation View
Group View
Temporal View

Preprocess and Data Analysis

- Confirmed authors and confirmed papers
  - Indexed authors who have been **identified**.
  - Each conformed author will associated **with multiple papers(confirmed papers group)**.

Search “Rui Wang” form dblp:

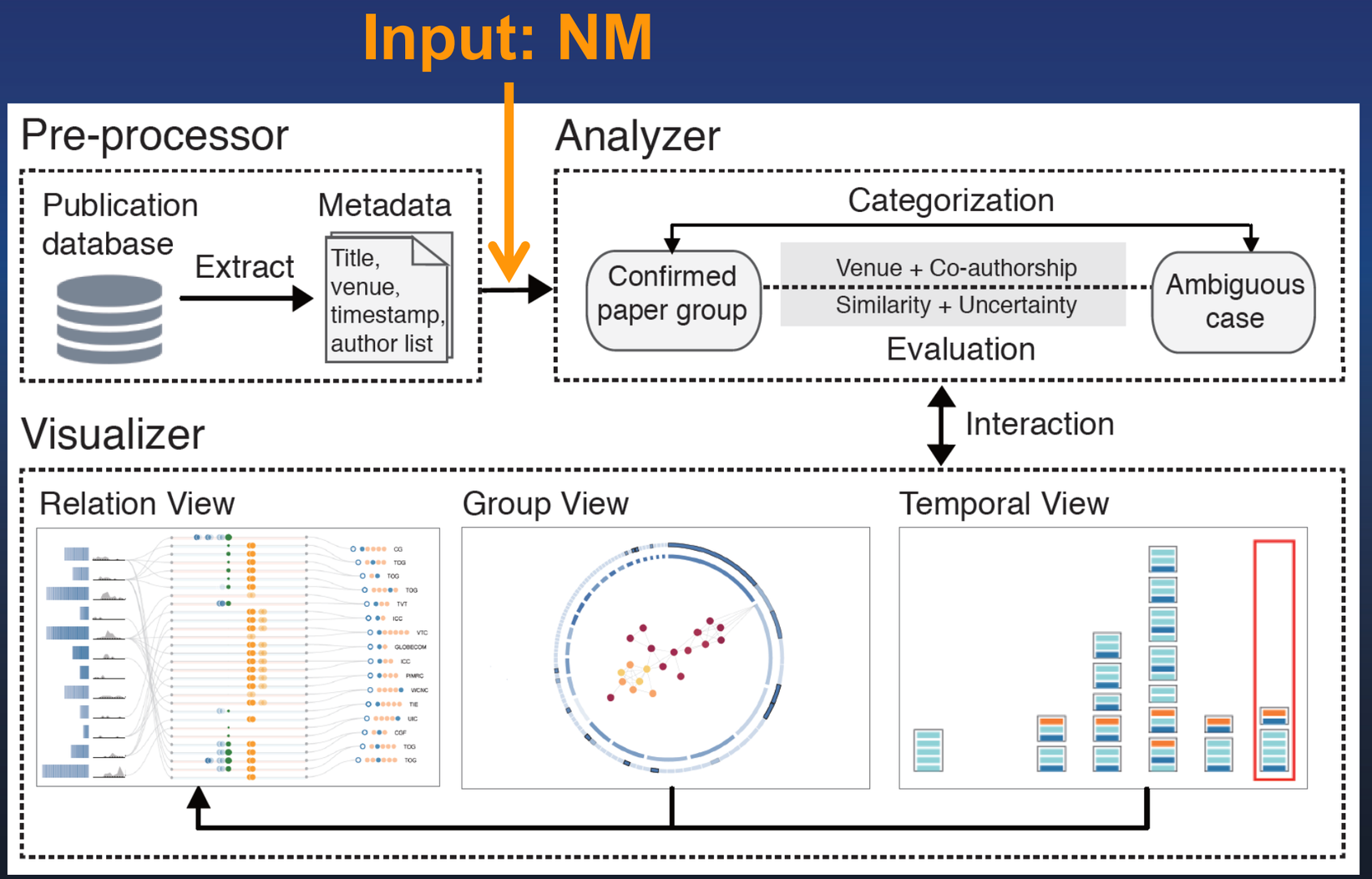
Confirmed Authors

Preprocess and Data Analysis

- Confirmed authors and confirmed papers
  - Indexed authors who have been **identified by the system**.
  - Each conformed author will associated **with multiple papers(confirmed papers group)**.

- Ambiguous names and ambiguous papers:
  - The author names which have not been identified.
  - The papers with no confirmed authors are ambiguous papers.

Preprocess and Data Analysis

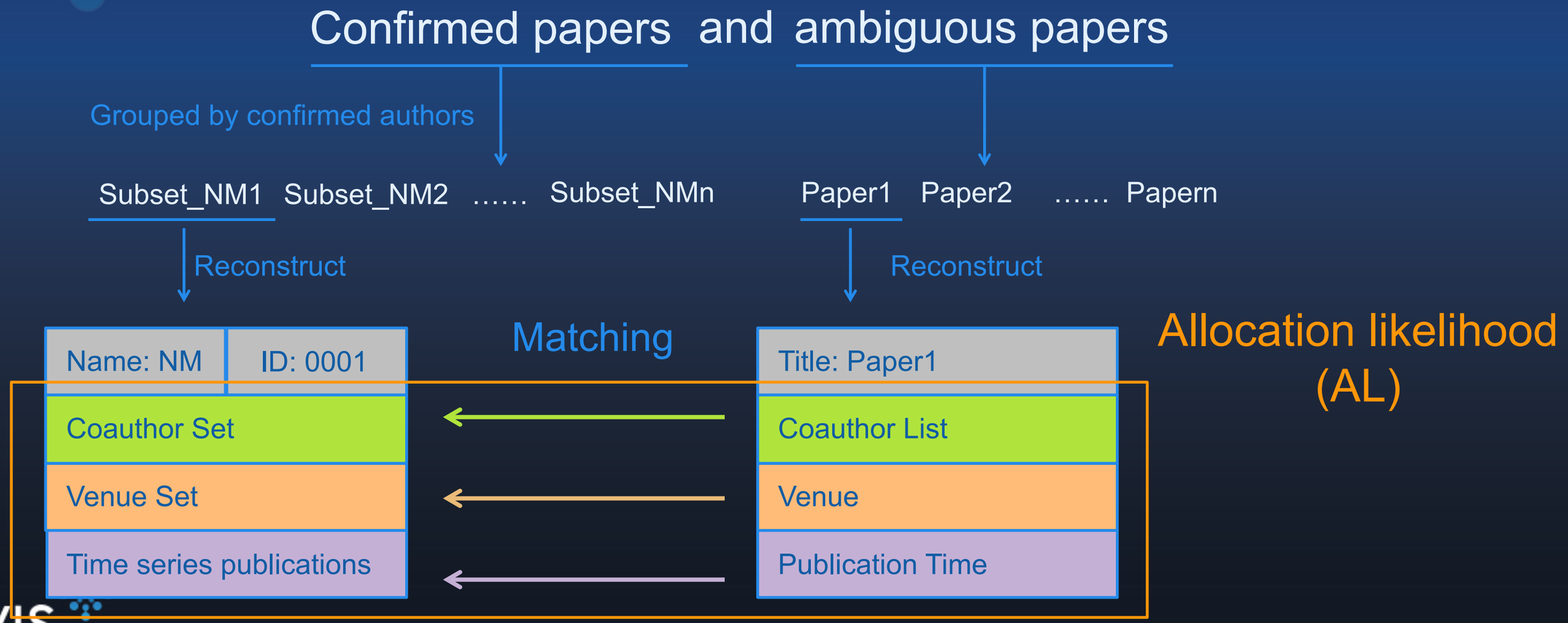


Given an author name **NM**, a collection of publications with the name NM(or approximate to NM) listed as an author will extracted from digital librariy.

VIS 2016

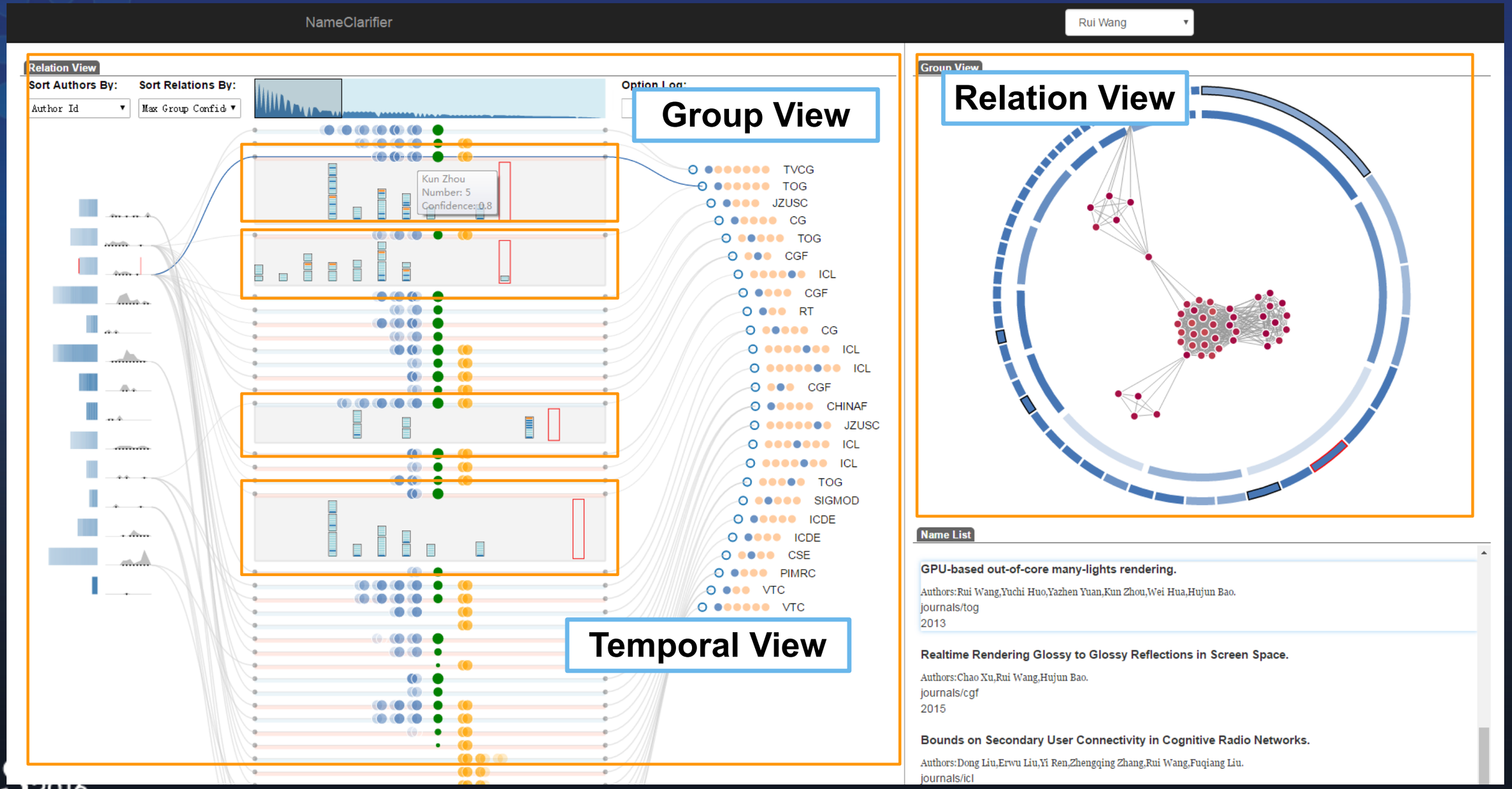
Preprocess and Data Analysis

- With the input name NM:

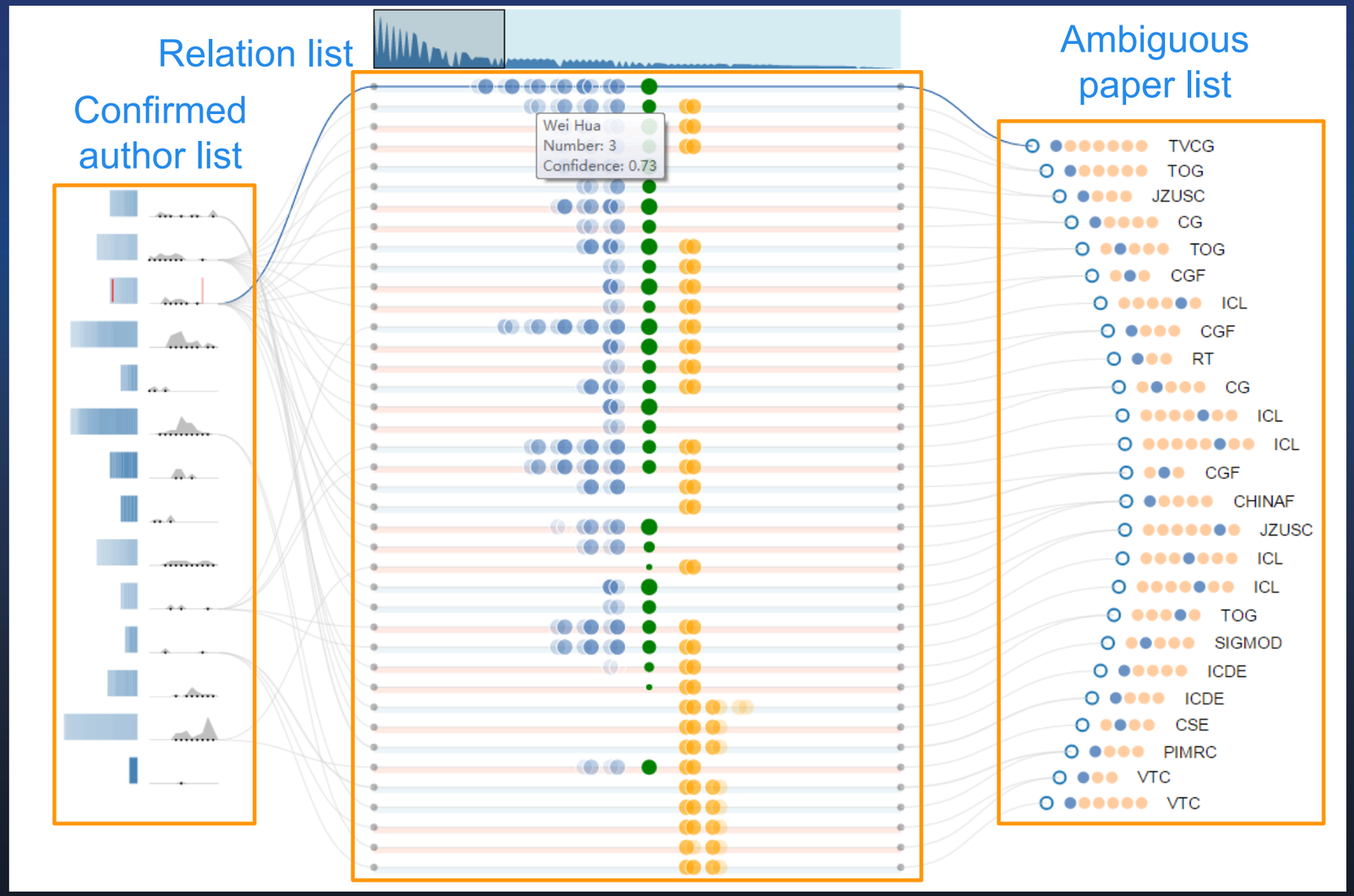


Visual Design

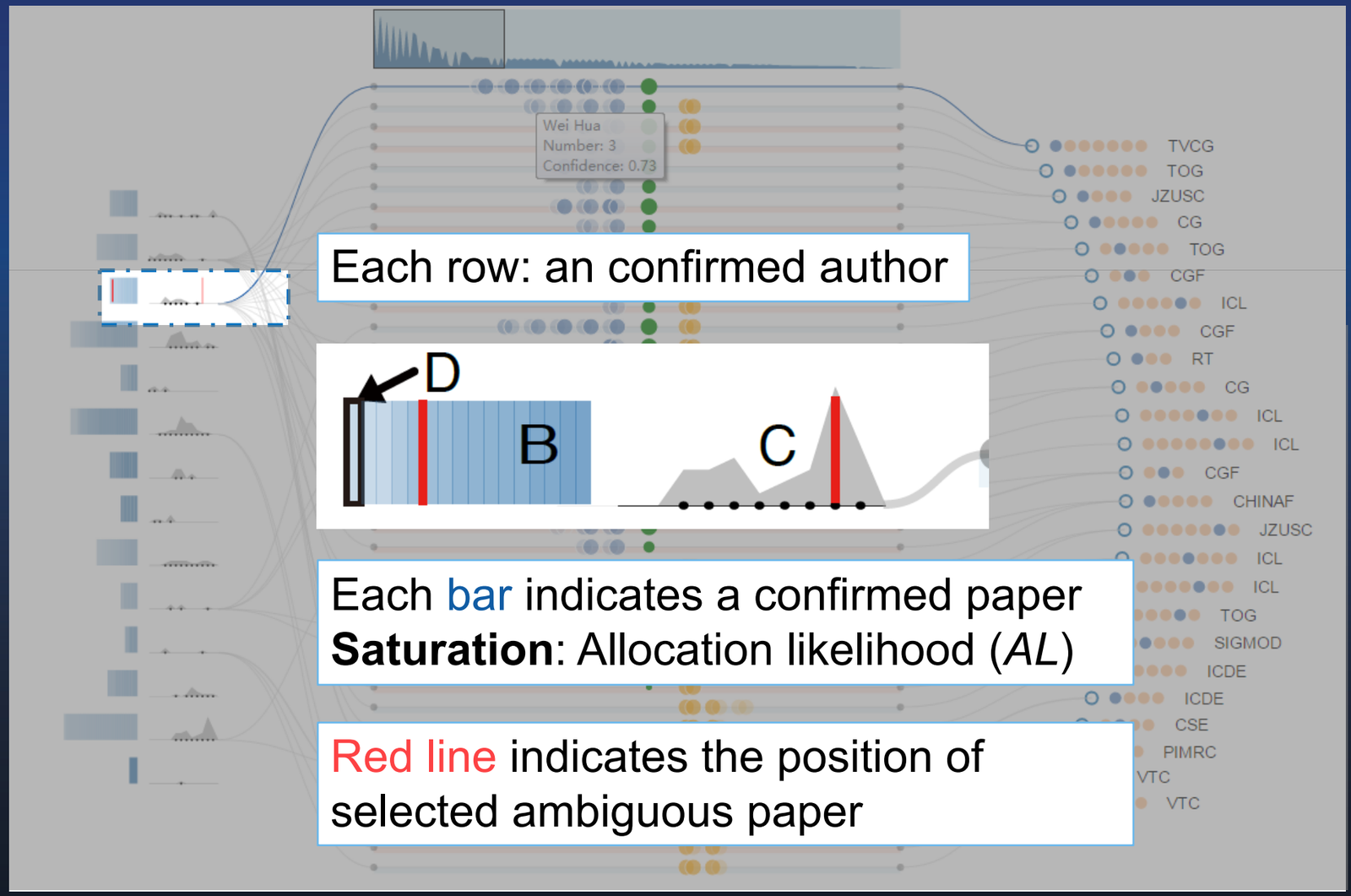
System Overview



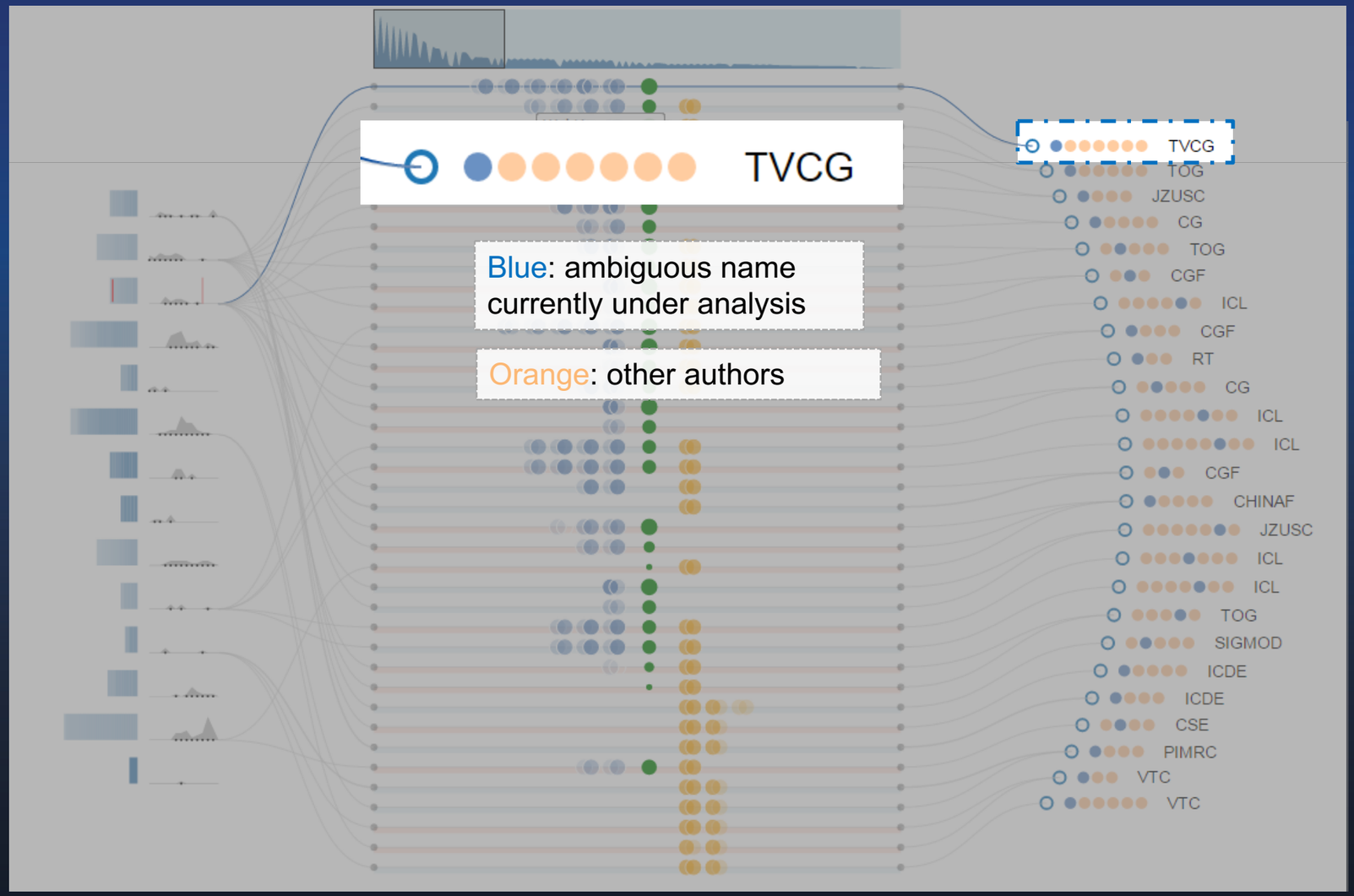
Relation View



Relation View

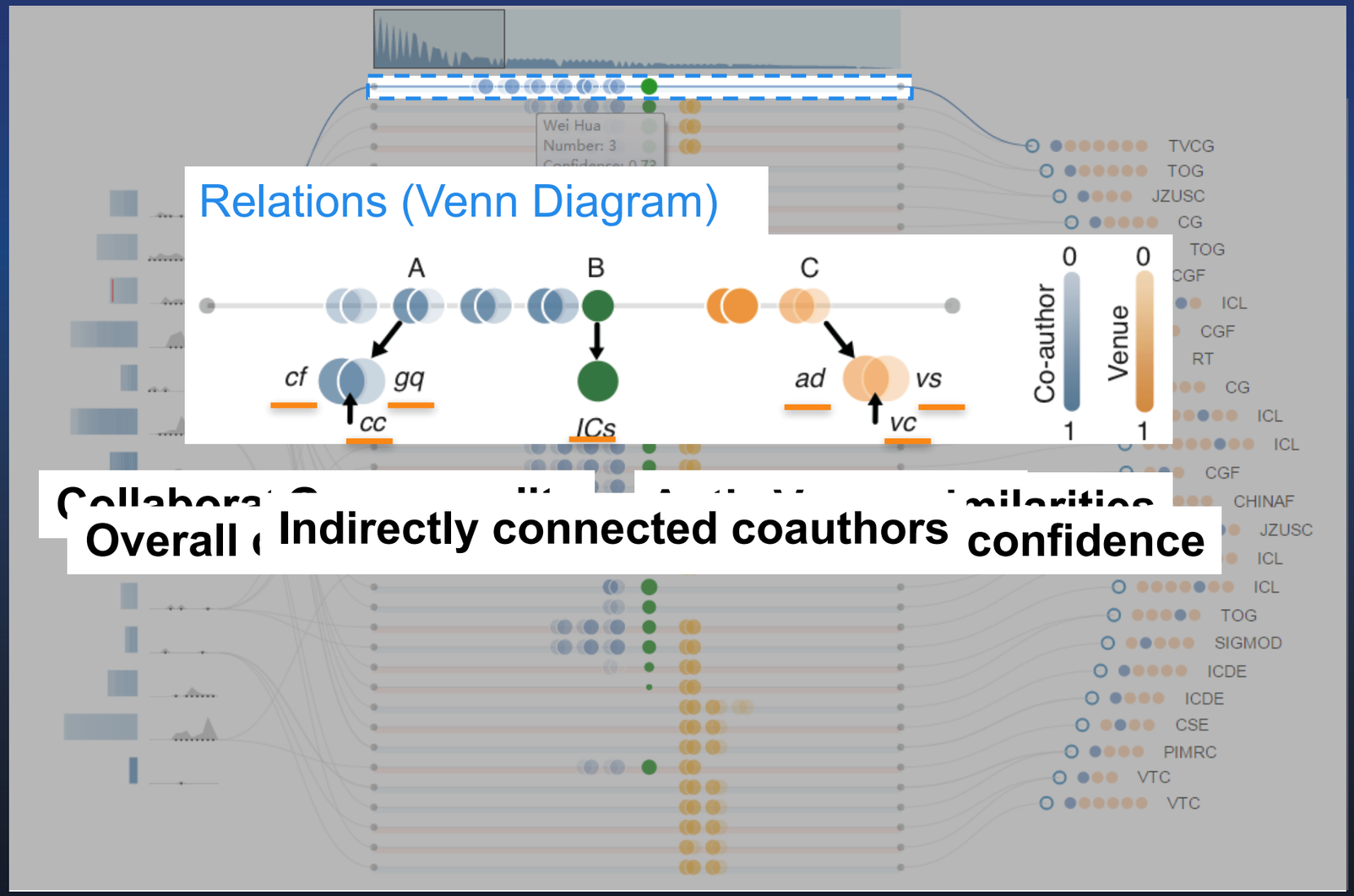


Relation View



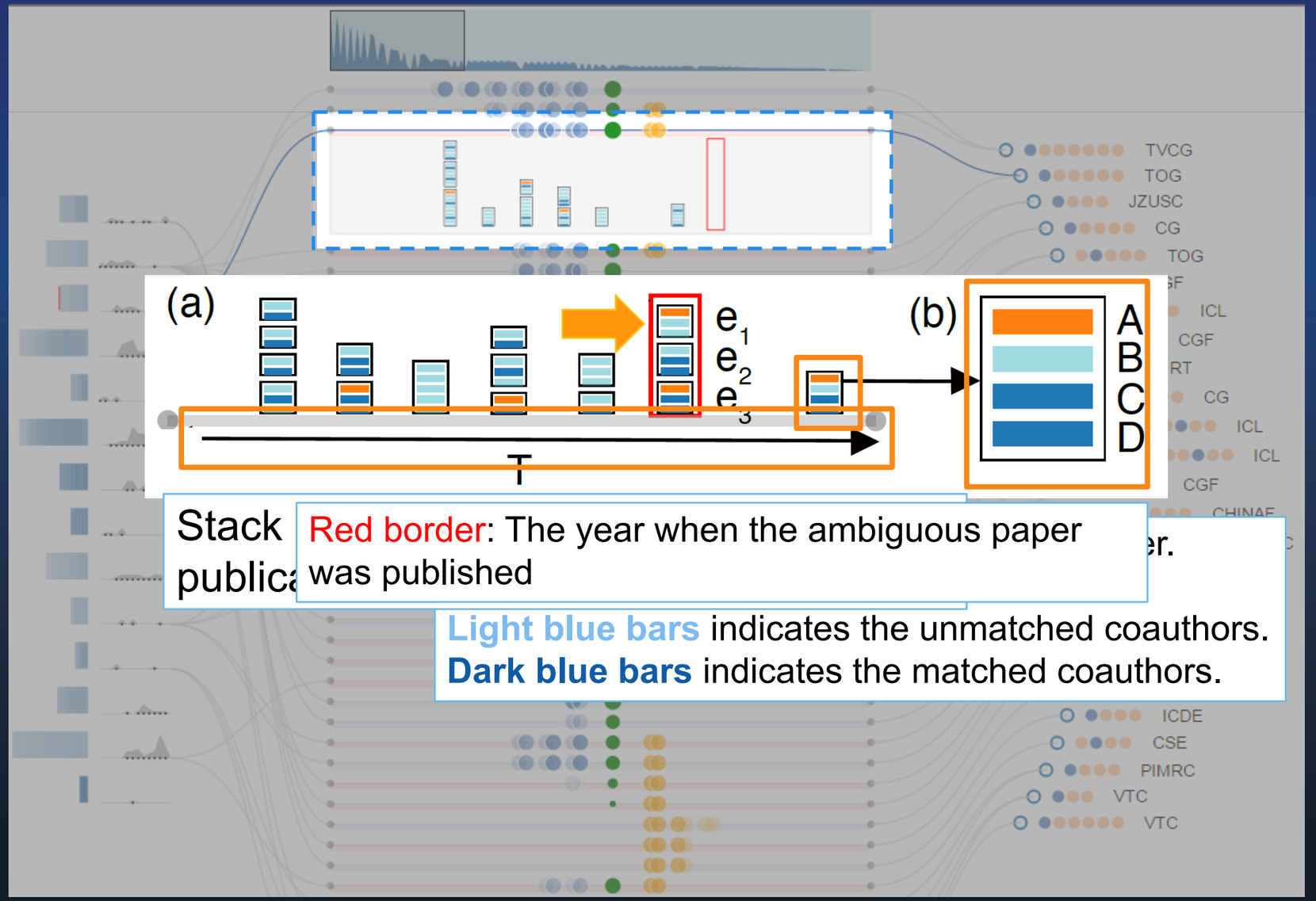
VIS 2016

Relation View



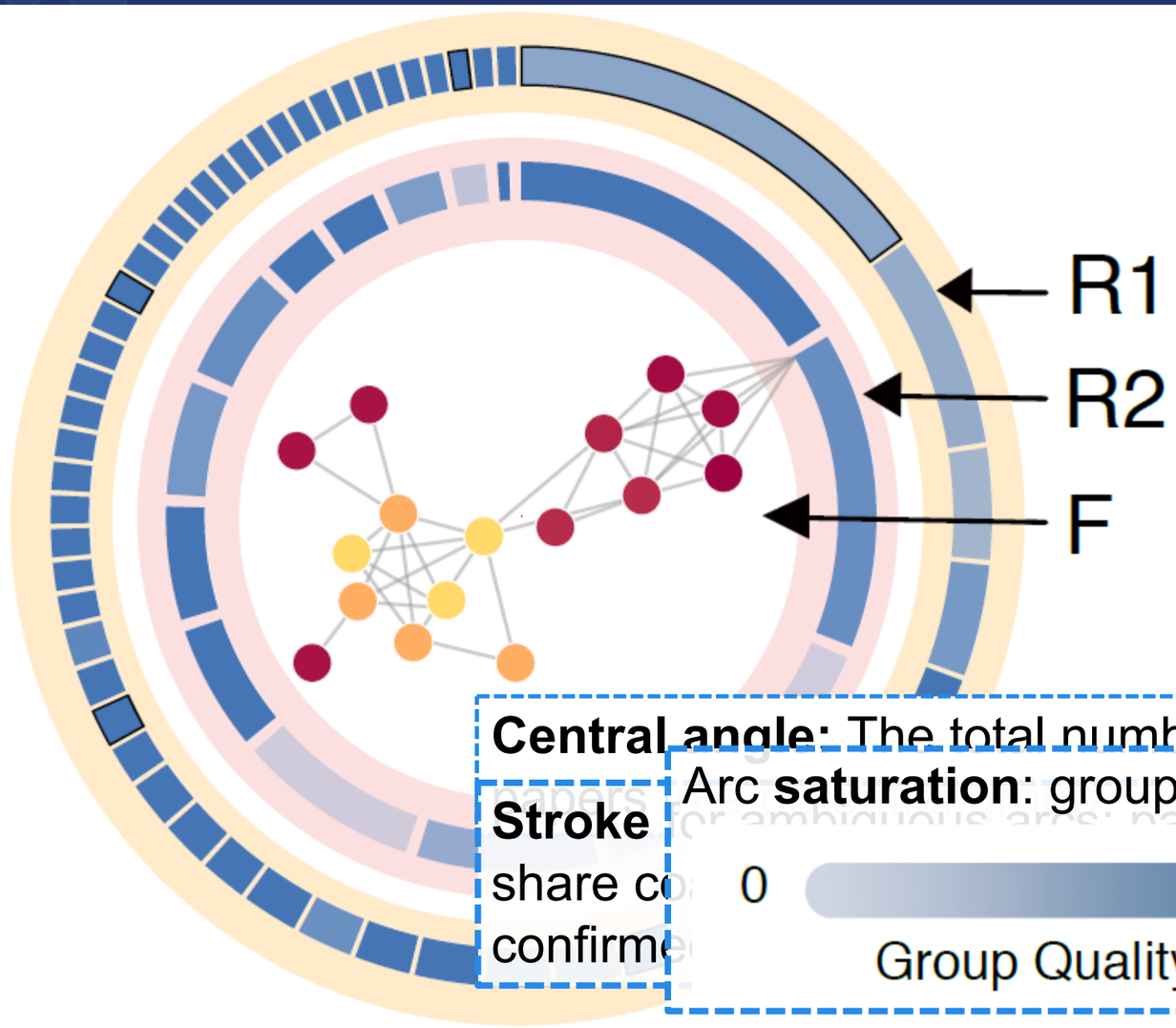
VIS 2016

Temporal View



VIS 2016

Group View



Outer ring(R1):

- Ambiguous paper group
- In each arc papers only share coauthor/venue with those in the same group

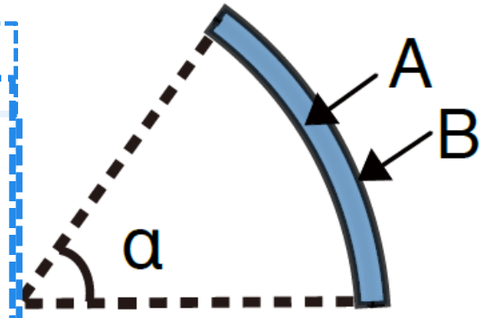
Inner ring(R2):

- Confirmed authors
- Every arc: a confirmed author

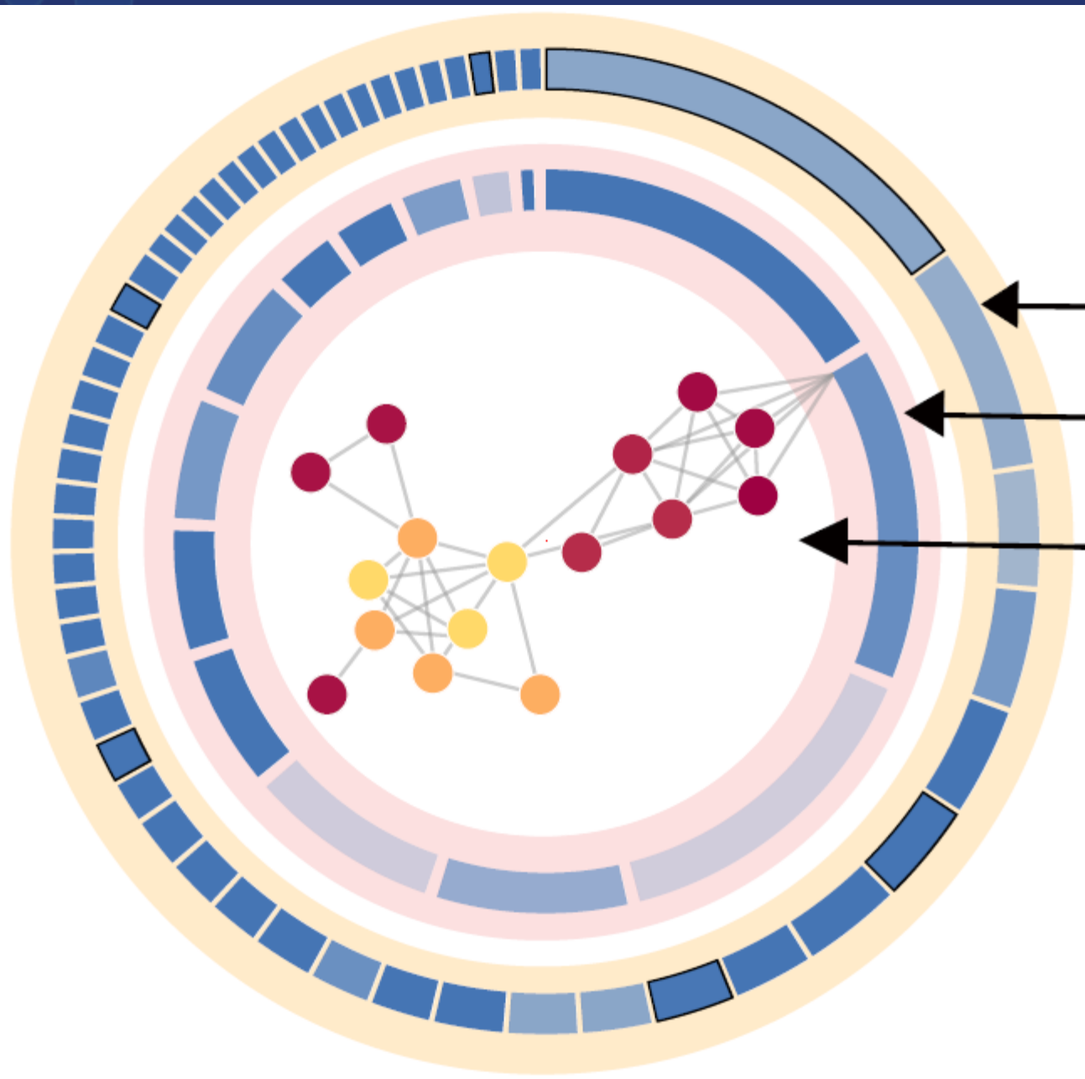
Central angle: The total number of papers for ambiguous arcs: papers

Stroke
share co
confirme

Arc **saturation**: group quality



Group View



(F) Nodes: papers in a selected ambiguous(paper) arc

Edges:

- Two ambiguous papers share coauthors
- Ambiguous papers share coauthors with confirmed authors

Node colors: publication years

Case study

Case Studies

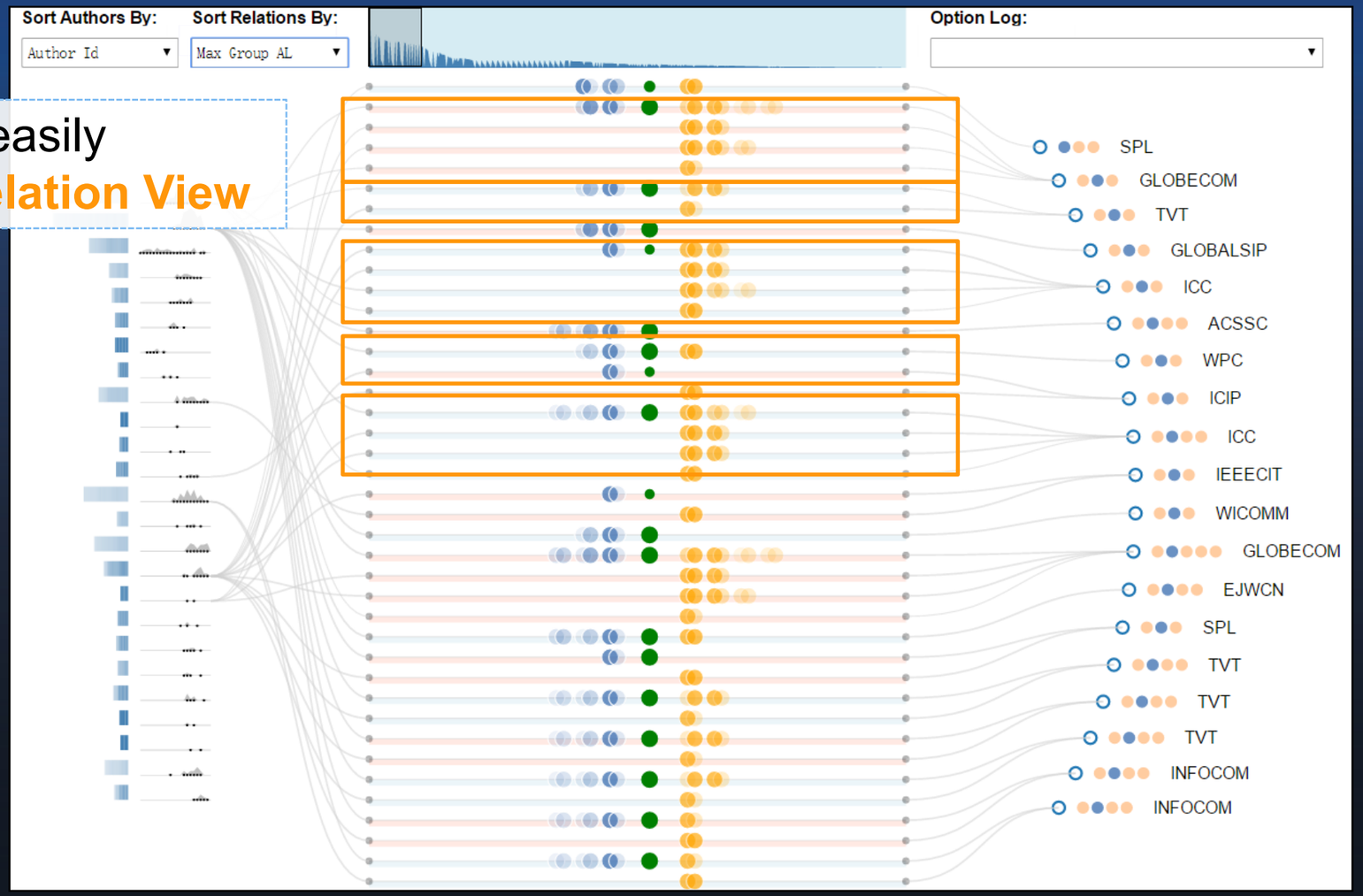
Case1: Wei Chen

Sort by Max Group Relation Allocation Likelihood

 # Total paper: 1170

- *573 ambiguous*
- *597 confirmed papers for 25 confirmed authors*

The most cases can be easily addressed directly by **Relation View**



VIS 2016

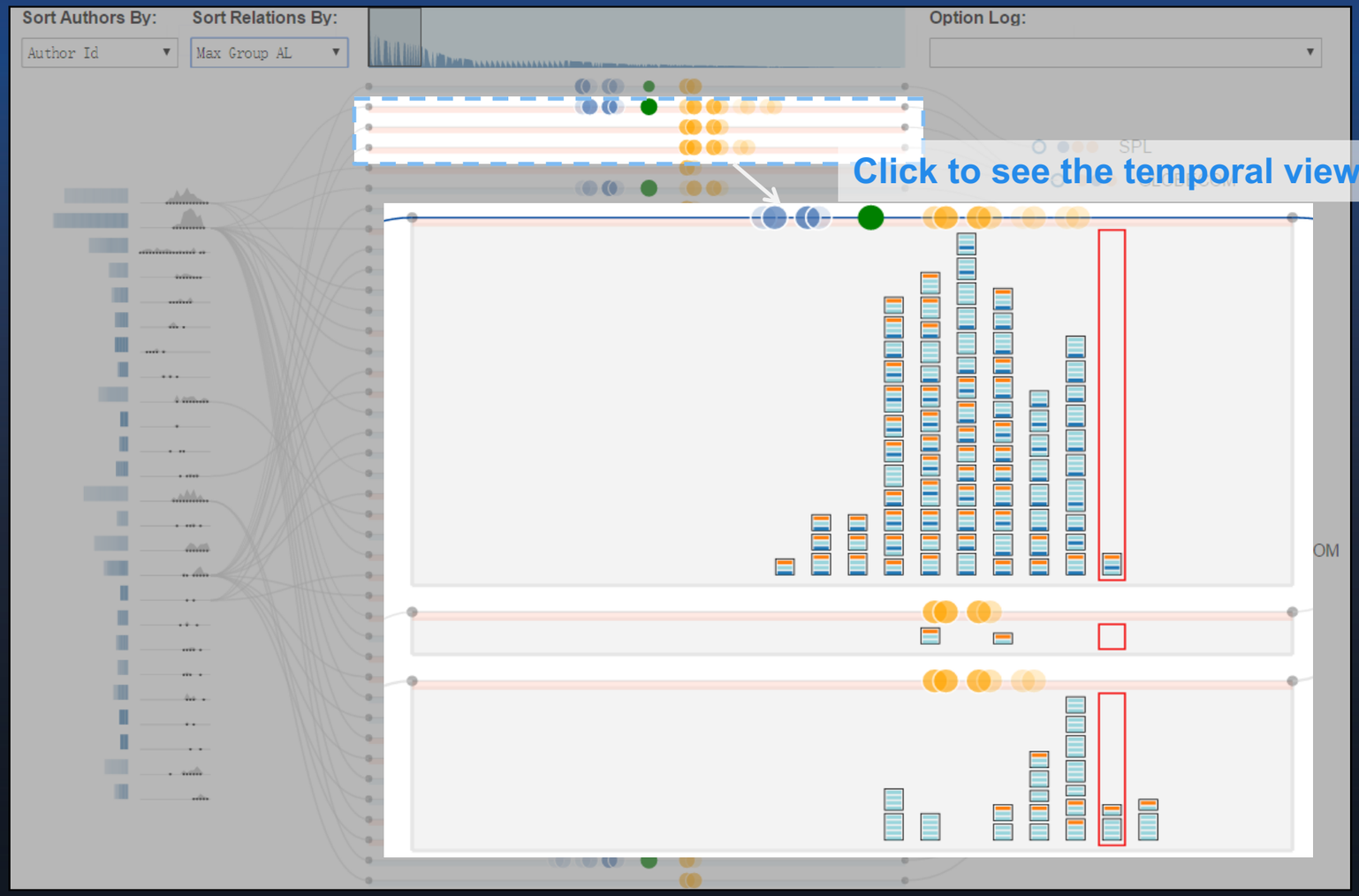
Case Studies

Case1: Wei Chen

Sort by Max Group Relation Allocation Likelihood

 # Total paper: 1170

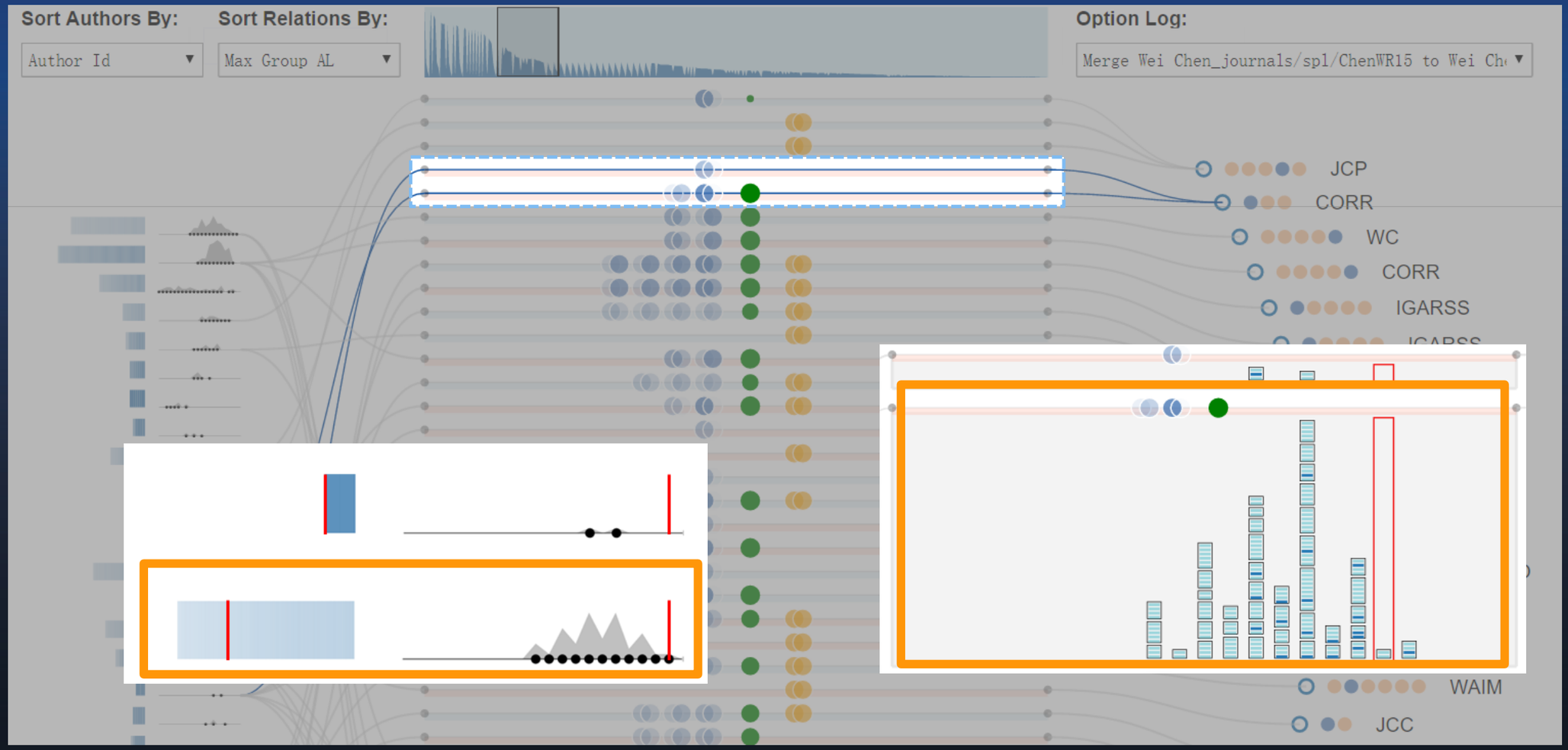
- *573 ambiguous*
- *597 confirmed papers in 25 confirmed authors*



Case Studies

Case1: Wei Chen

In some cases, the allocation likelihood is different from the visual pattern.



Case Studies

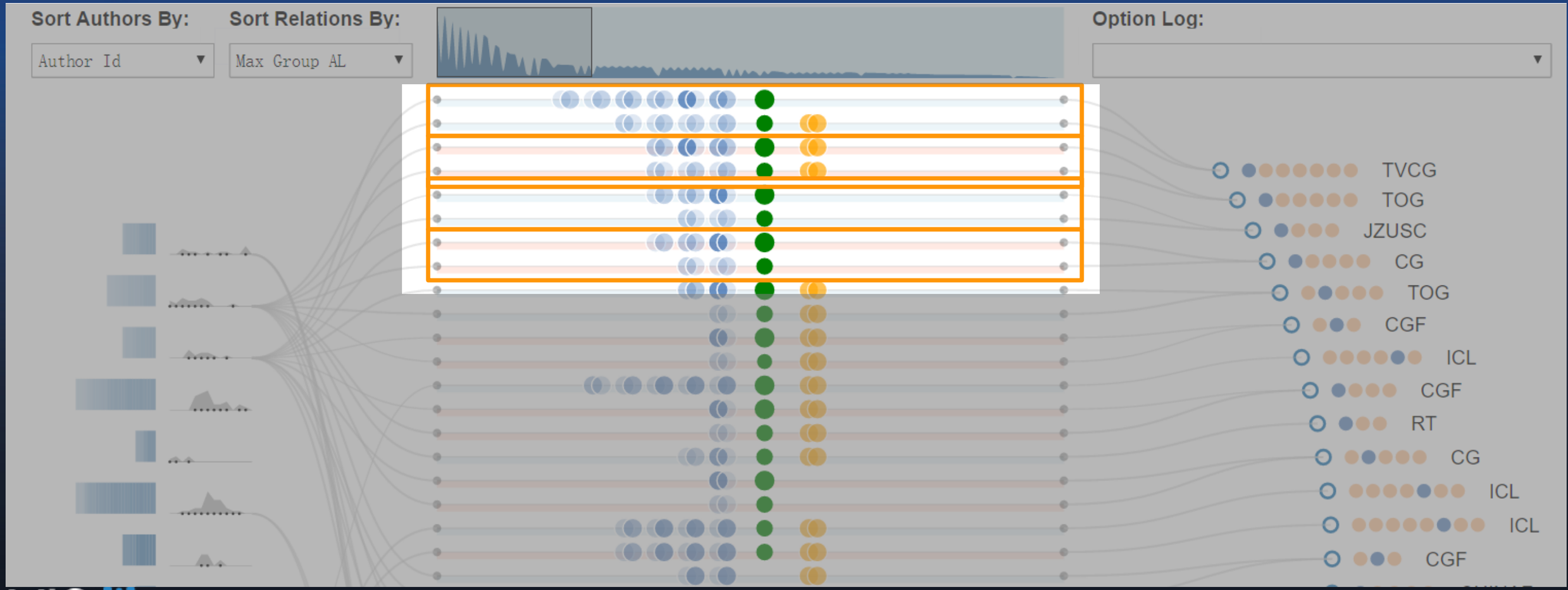
Case2: Rui Wang

 # Total paper: 560

- *179 ambiguous + 381 recognized papers*
- *15 recognized authors*

Sort by Max Group Relation Allocation Likelihood

The most tricky one: It cannot be easily distinguished through comparison link and temporal view

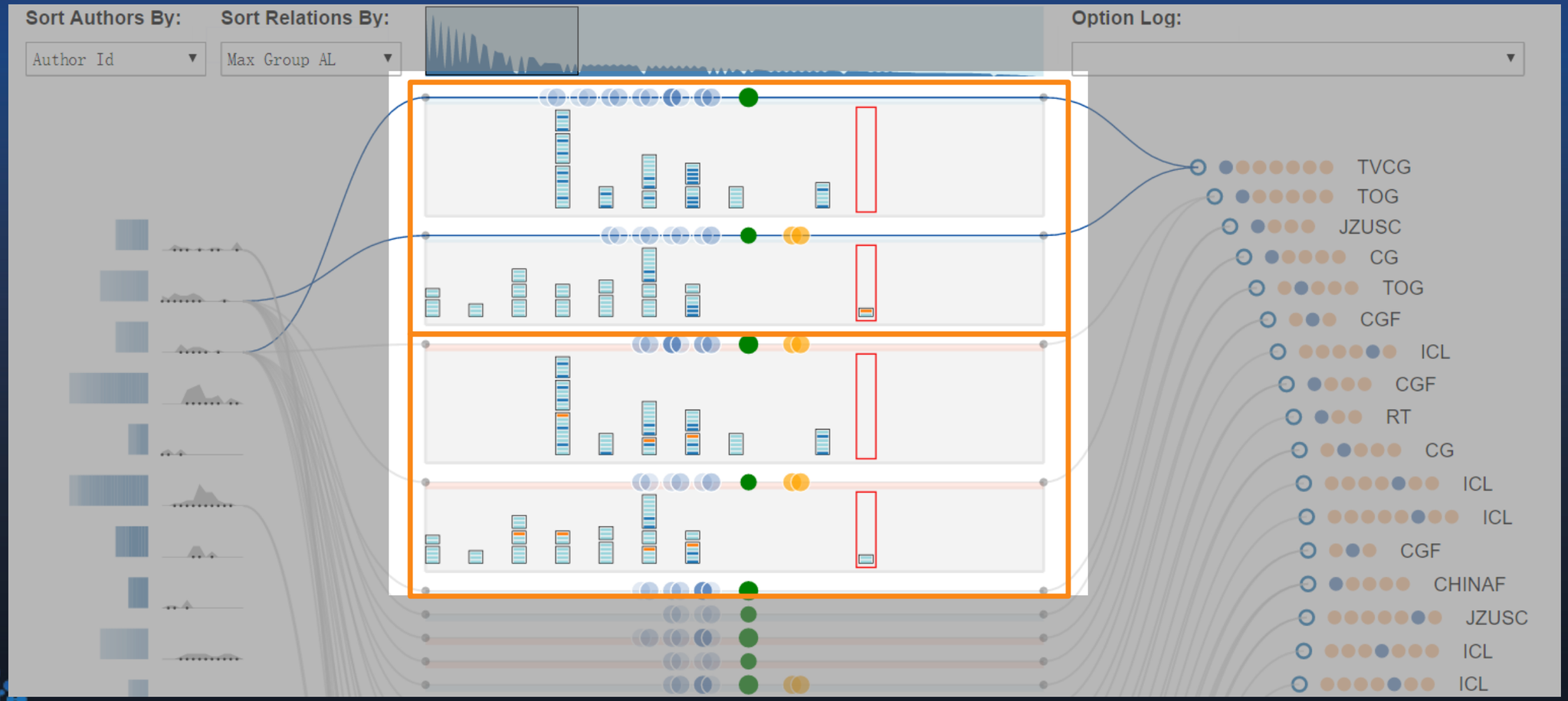


Case Studies

Case2: Rui Wang

Sort by Max Group Relation Allocation Likelihood

The most tricky one: It cannot be easily distinguished through comparison link and temporal view



Case Studies

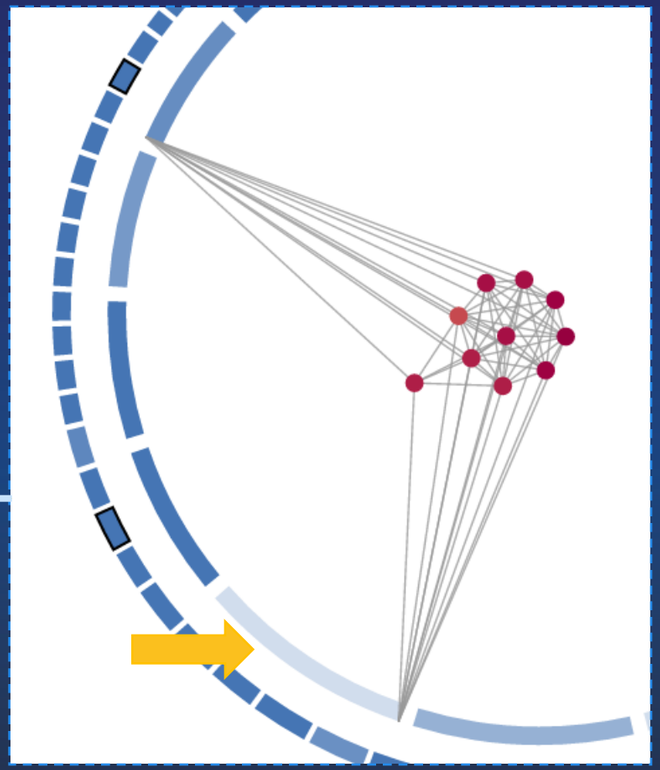
Case2: Rui Wang

Papers closely connected to both these two confirmed authors

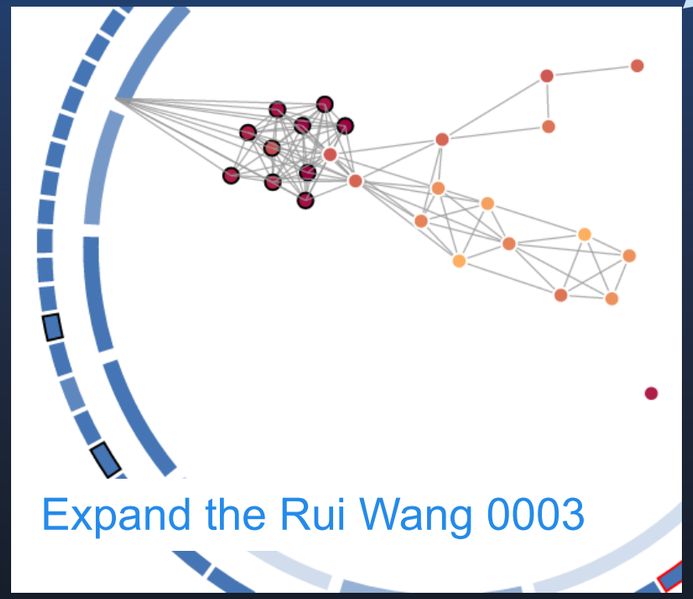
Case Studies

Case2: Rui Wang

Release papers of the Rui Wang 0003



Some nodes with the black strokes are **loosely connected** with those Rui Wang 0003's papers

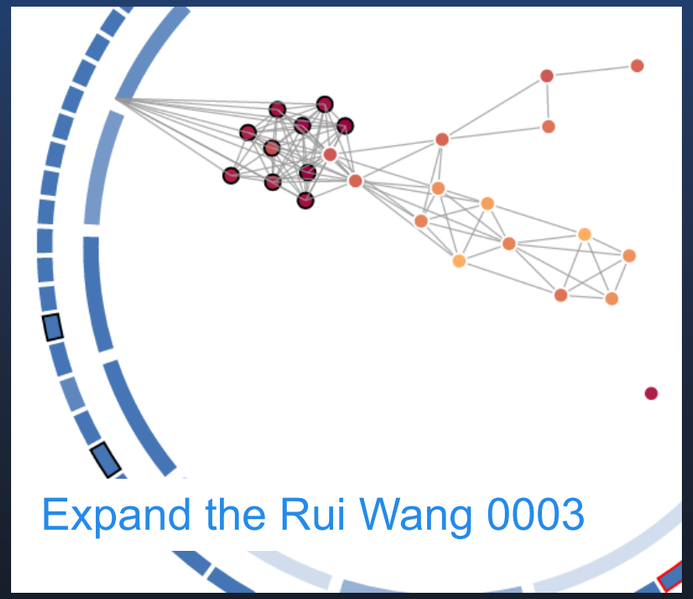


Expand the Rui Wang 0003

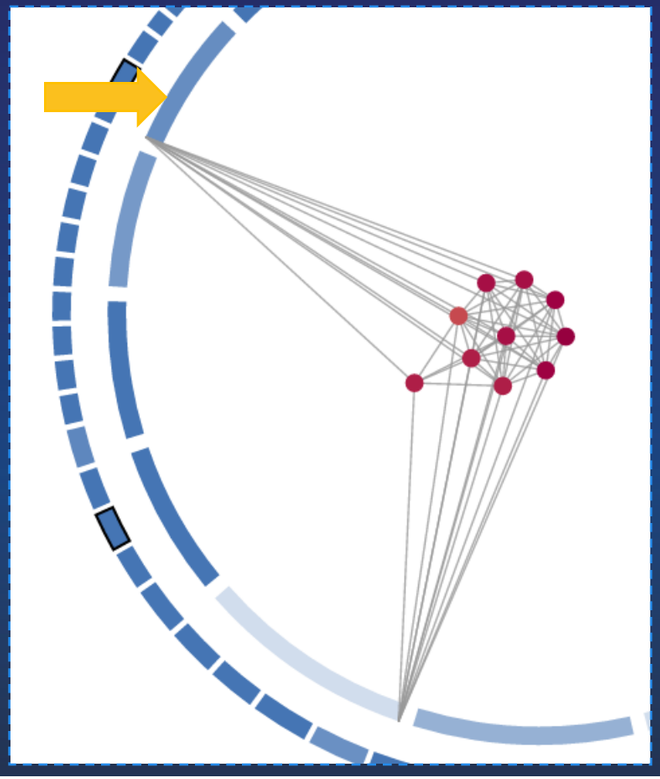
Case Studies

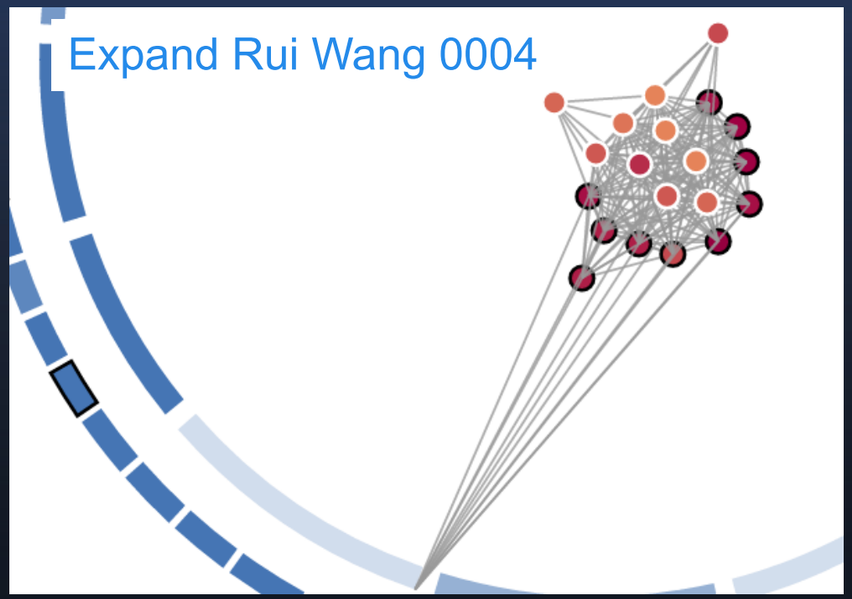
Case2: Rui Wang

Release the papers of the Rui Wang 0004



Nearly **all** the nodes with the black strokes are **tightly connected** with those Rui Wang 0004's confirmed papers.





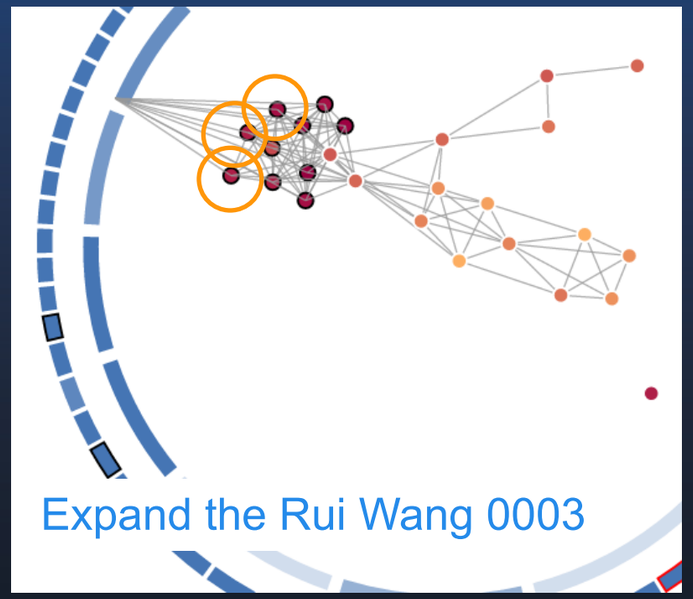
We tend to think all the ambiguous papers belong to Rui Wang 0004

VIS 2016

Case Studies

Case2: Rui Wang

Start exploration from the farthest one from 0003

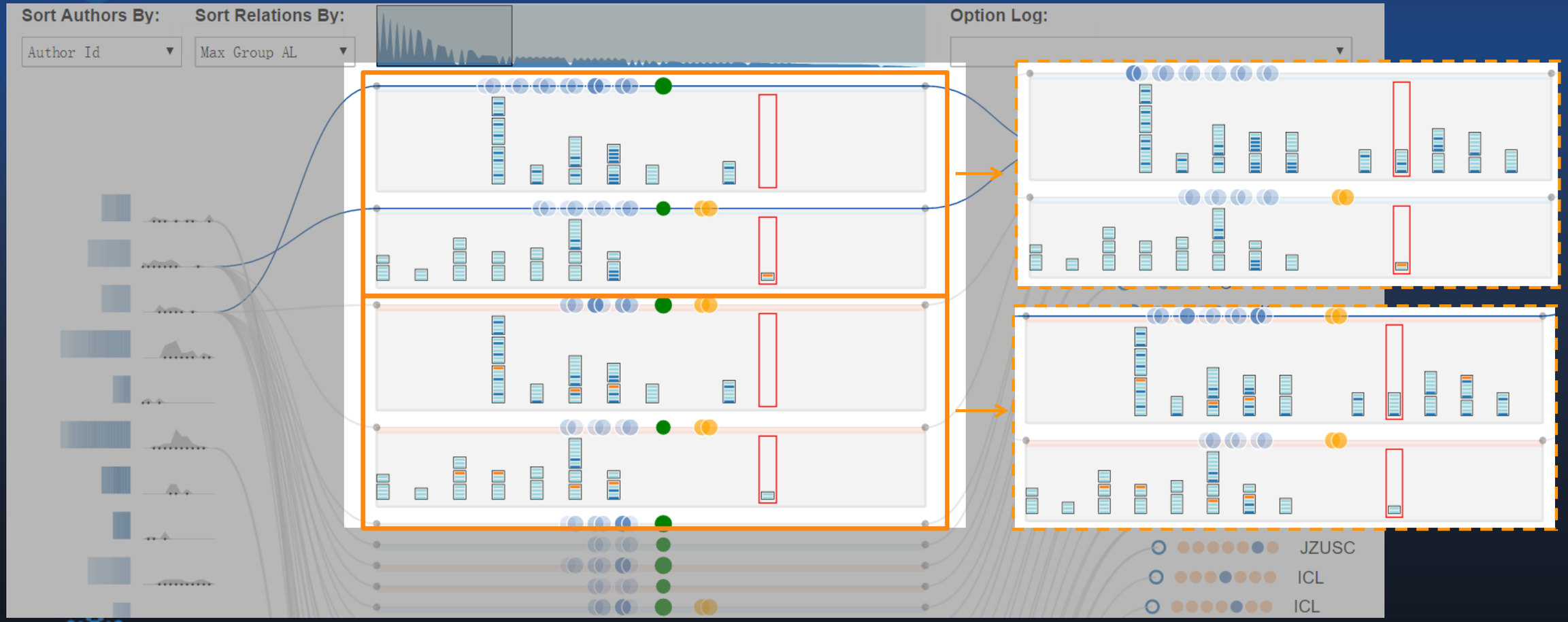


VIS 2016

Case Studies

Case2: Rui Wang

Think back to the most tricky one:



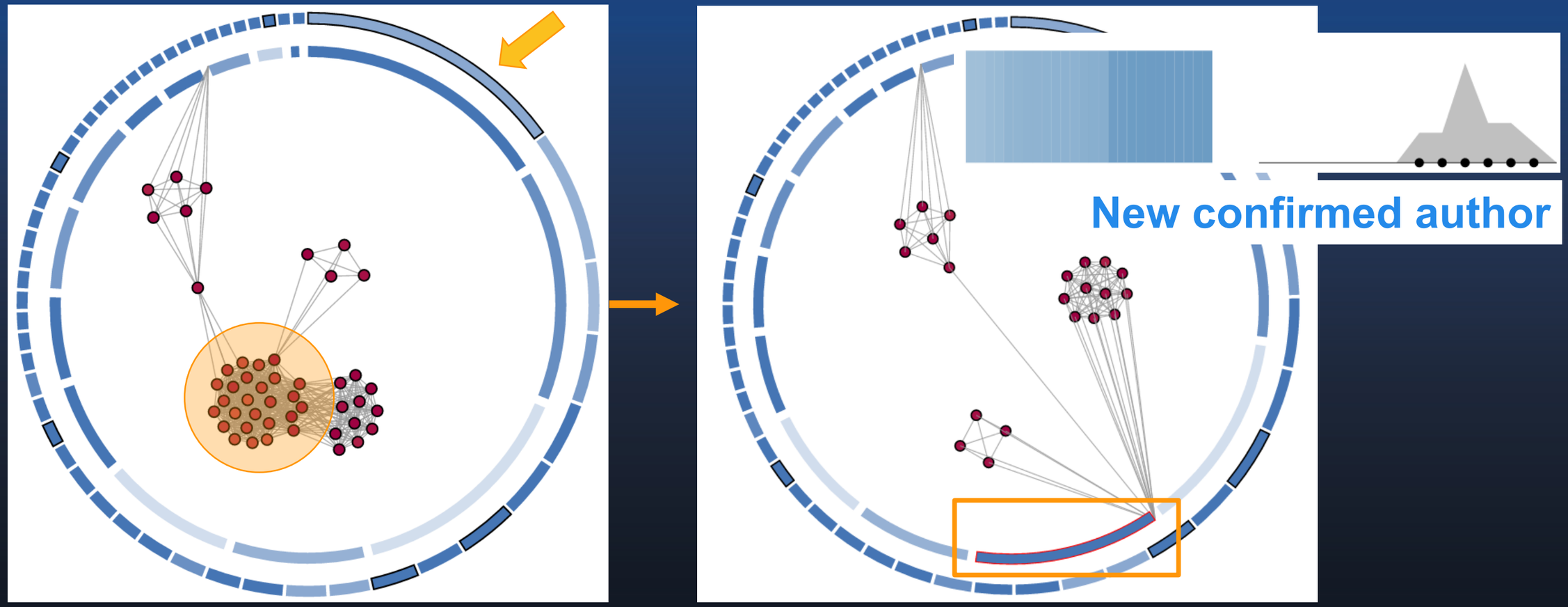
More evidence are provided to make relations **distinguishable**.

Case Studies

Case2: Rui Wang

Start from the largest ambiguous arc.

Select this part and form a new confirmed author.

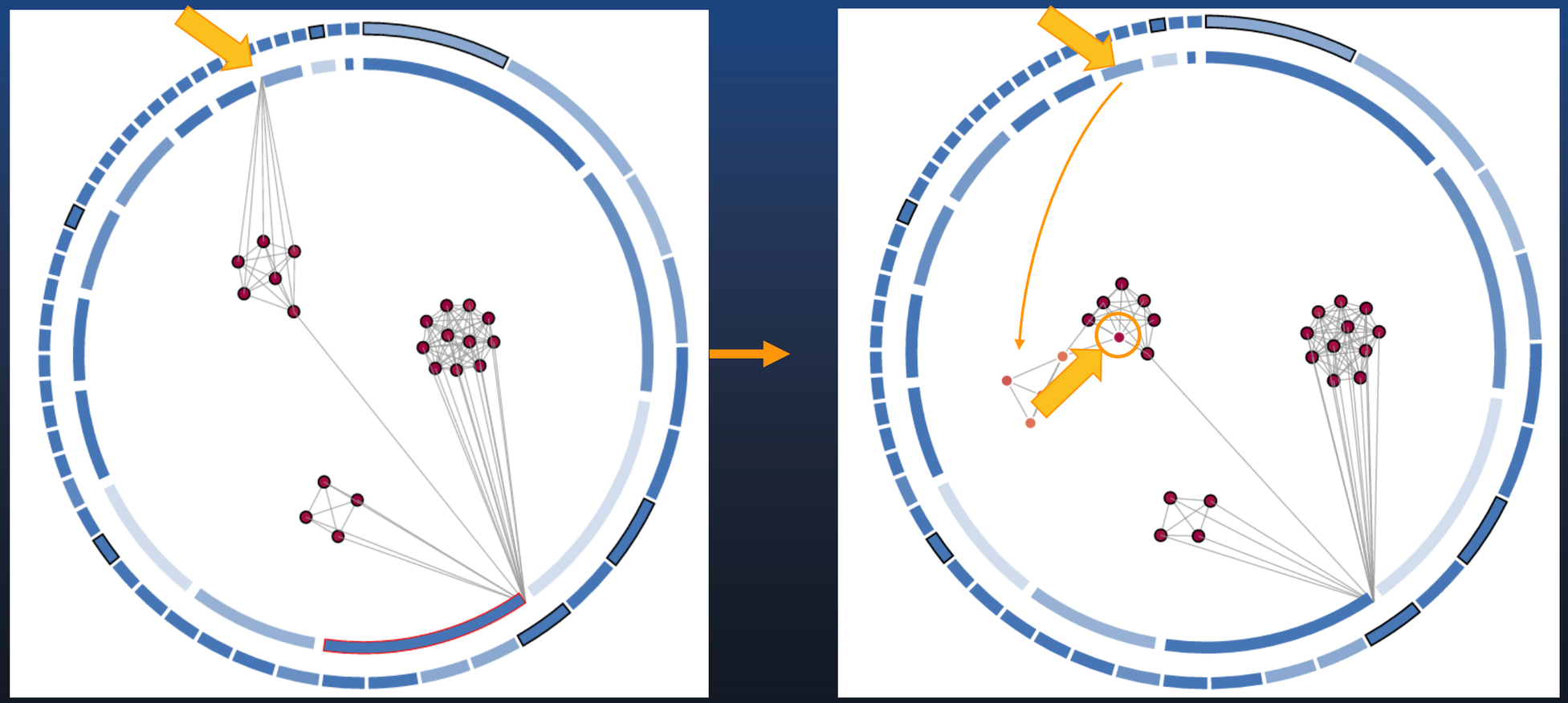


Case Studies

Case2: Rui Wang

Start from the largest ambiguous arc.

Notice that there is one connect with a confirmed authors.

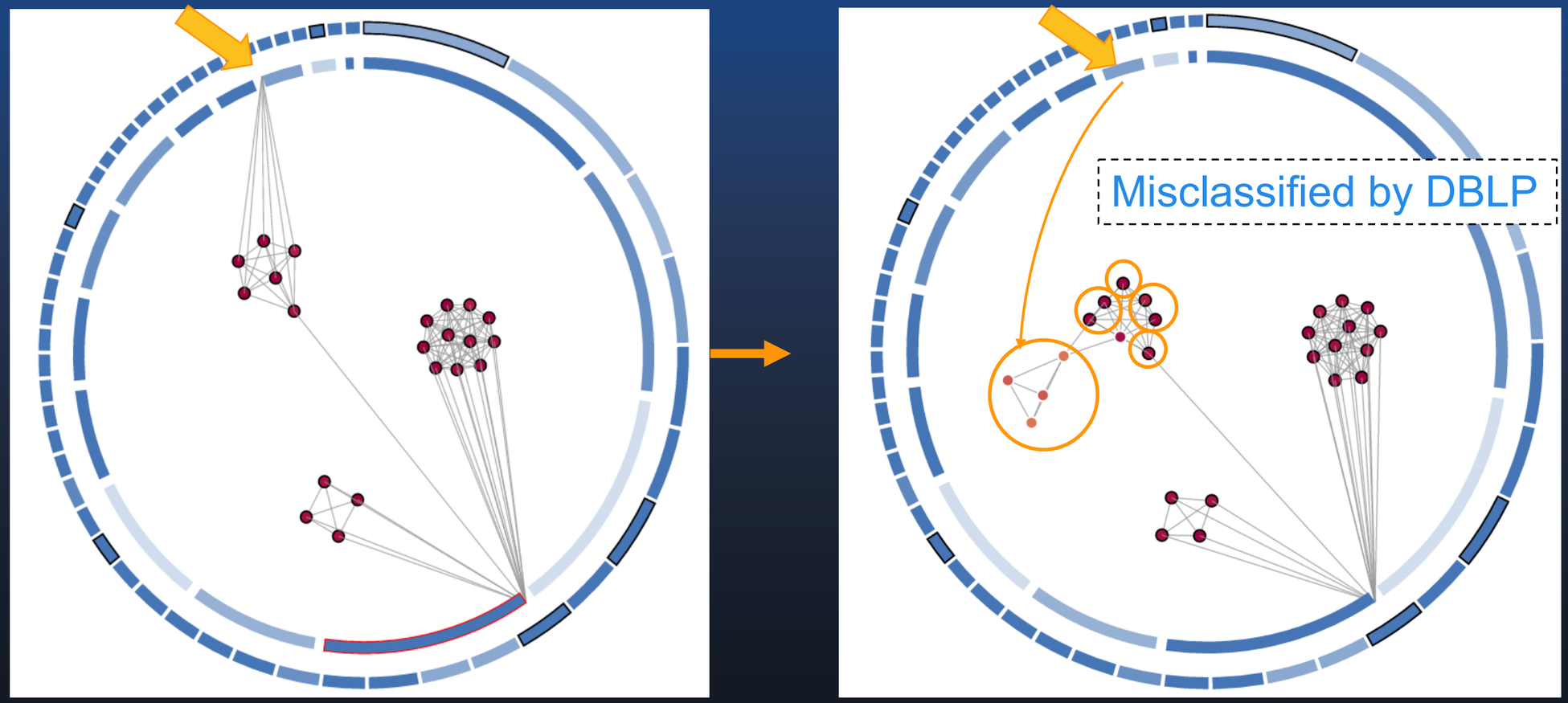


Case Studies

Case2: Rui Wang

Start from the largest ambiguous arc.

Notice that there is one connect with a confirmed authors.



Conclusion

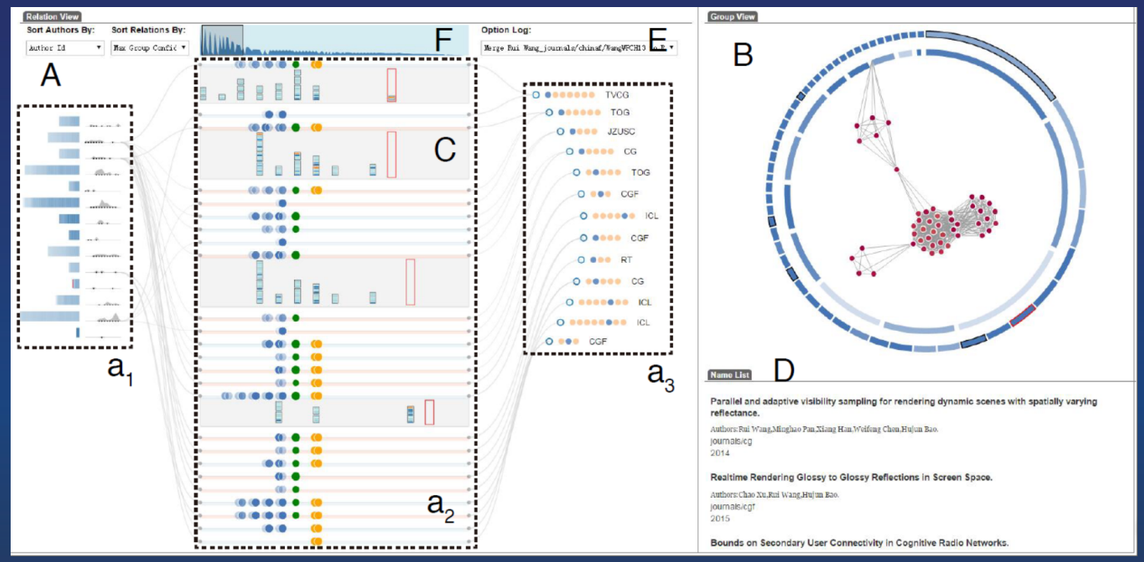
- NameClarifier, an interactive visual system for name disambiguation;
- Turn the traditional black-box solution into a white-box procedure;
- The system provides guidance instead of classification results for ambiguous cases.

Future work

- Extension to more attributes;
- Visual alarming for the improper operation;

Thank you!

Q&A



NameClarifier: A Visual Analytics System for Author Name Disambiguation

Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu and Weiwei Cui

VIS 2016

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Microsoft Research 微软亚洲研究院

Back Up

- Automatic Evaluation
  - **Allocation Likelihood**
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Back Up

- Automatic Evaluation
  - **Allocation Likelihood**
    - **Co-author Matching**
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper p_i and Ambiguous paper p_A

$$cm_{i,p} = \frac{|C(p_i) \cap C(p_A)|}{|C(p_i) \cap C(p_A)|}$$

Back Up

- Automatic Evaluation
  - **Allocation Likelihood**
    - Co-author Matching
    - **Venue Match**
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper p_i and Ambiguous paper p_A

$$vm_{i,p} = sgn(vs(v_A, v_i) - s)$$

where

$$vs(v_A, v_i) = \left| \frac{|A(v_A) \cap A(v_i)|}{|A(p_i) \cap C(p_A)|} \right|$$

Back Up

- Automatic Evaluation
  - **Allocation Likelihood**
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper p_i and Ambiguous paper p_A

$$AL(A,G) = \frac{1}{n}\sum_{i=1}^{n}(\alpha_c \cdot cm_i + \alpha_\beta \cdot vm_i)$$

Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper p_i and Ambiguous paper p_A

$$cc(c) = 1_{DC}(c) \cdot (cf(c) + gq(c))$$

Back Up

- Automatic Evaluation
  - Allocation Likelihood
    - Co-author Matching
    - Venue Match
  - Confidence Measurements
    - Co-author Confidence
    - Venue Confidence

Confirmed paper p_i and Ambiguous paper p_A

$$vc(p_i, p_a) = 1_{vr}(v_i) \cdot (ad(v_i) + vs(v_i, v_A))$$