Network Analytics ER Model

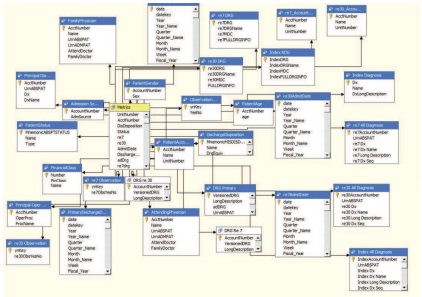
Towards a Conceptual View of Network Analytics

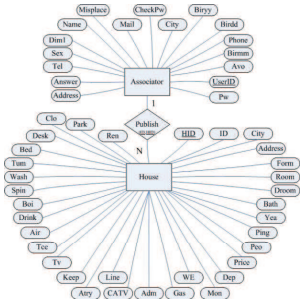
Qing Wang

Research School of Computer Science
The Australian National University
Australia
qing.wang@anu.edu.au

A Question[1]

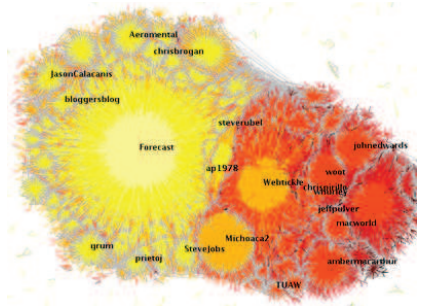
- What is the role of conceptual modelling in Big-data analytics, such as network analysis?

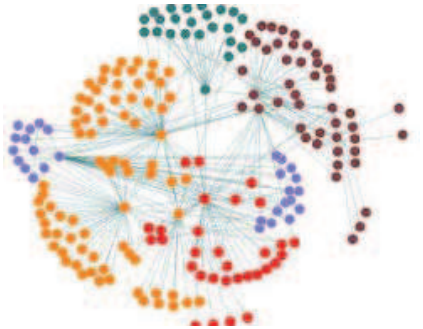




Conceptual modelling

?



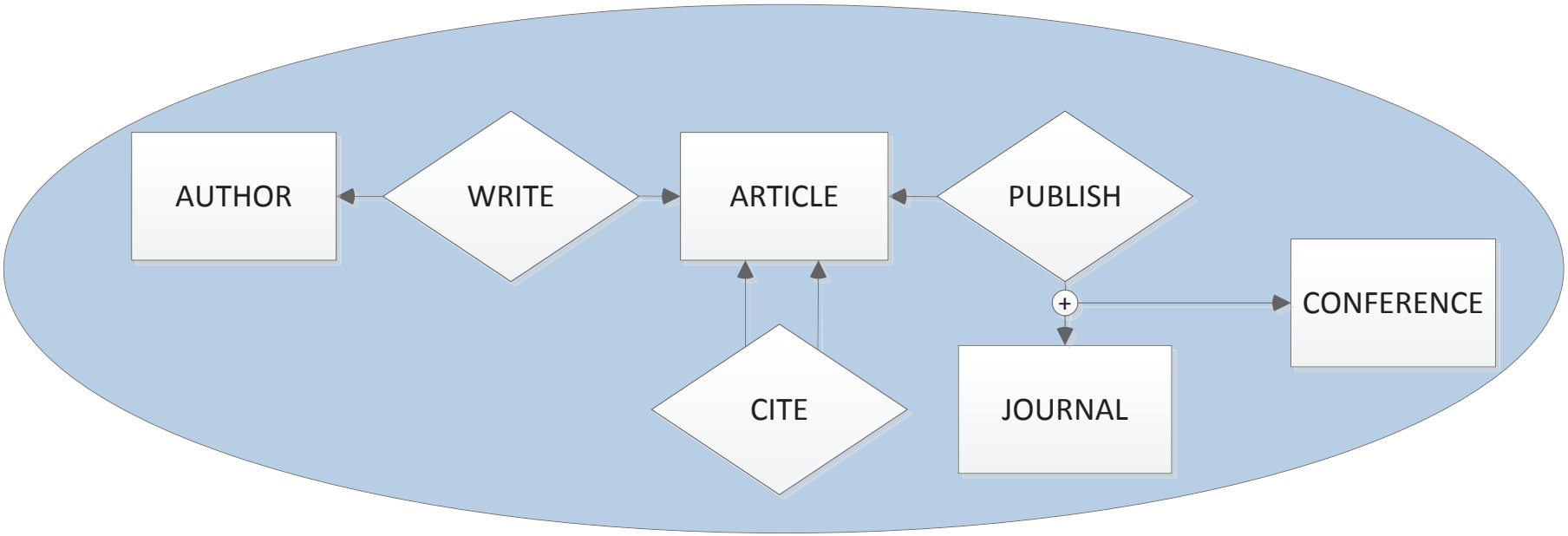


Network analysis

The images are taken from Google Image.

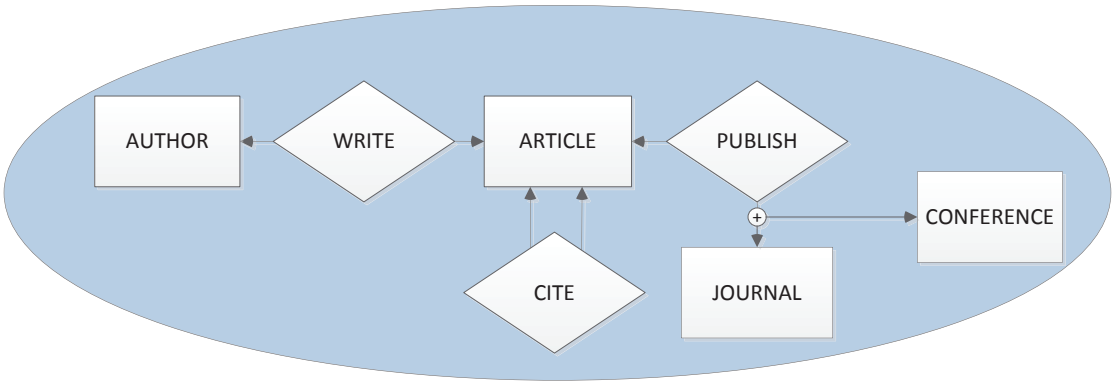
Motivating Example

- Let's start with a traditional ER model:



Motivating Example

- Queries in a bibliographical network:
  - Collaborative communities
  - Most influential articles
  - Top-k influential researchers
  - Correlation journal citation
  - ...



Motivating Example

- Queries in a bibliographical network:
  - Collaborative communities
  - Most influential articles
  - Top-k influential researchers
  - Correlation journal citation
  - ...

- Some questions:
  - **Semantic integrity**: Are they semantically relevant and consistent?
  - **Analysis efficiency**: Can the efficiency be improved by leveraging their semantics at the conceptual level?
  - **Network dynamics**: Can they be dynamically performed so as to predict trends?

Network Analytics ER Model

- We propose the Network Analytics ER Model (NAER) that extends the traditional ER models in three aspects:

  - **Structure**

    i.e., analytical types are added

  - **Manipulation**

    i.e., topological constructs are added

  - **Integrity**

    i.e., semantic constraints are extended.

The NAER Model - Structure

- Base types vs analytical types
  - **Base types**: from the data management perspective

    i.e., how to control data
  - **Analytical types**: from the data analysis perspective

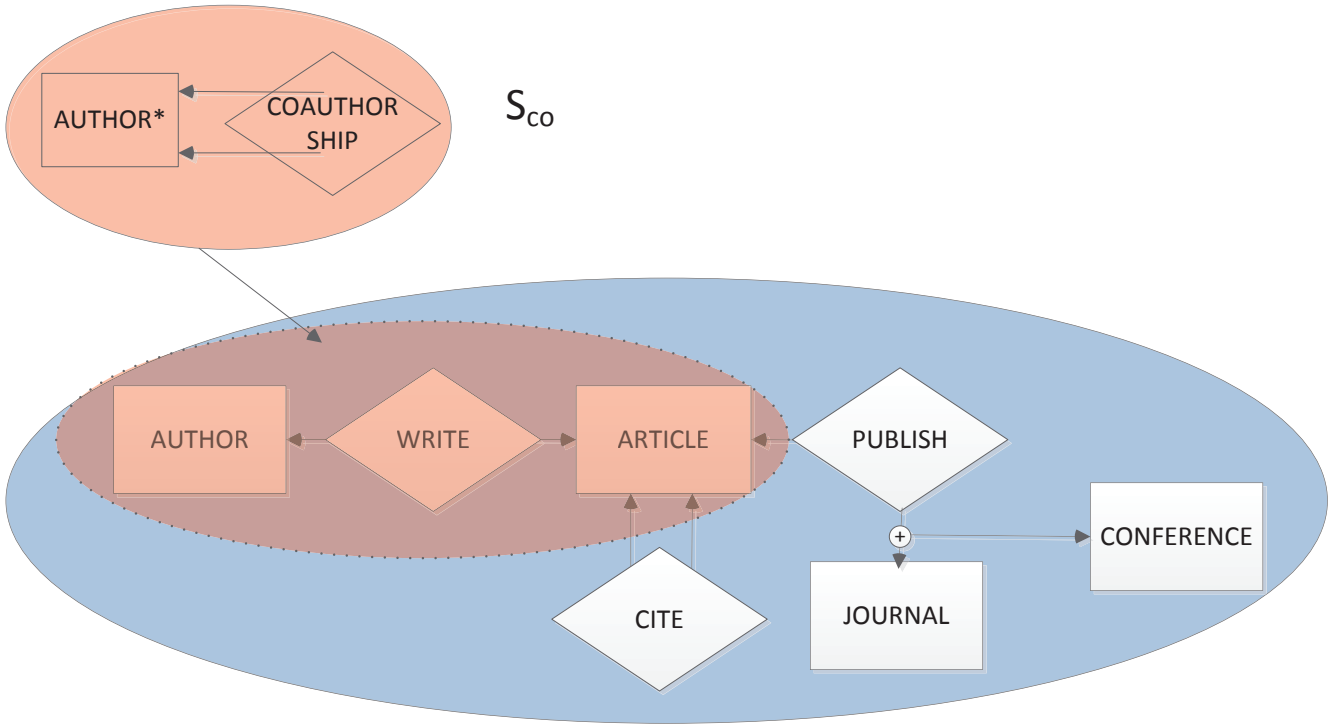
    i.e., how to use data

Base types	Analytical types
Base entity	Analytical entity
Base relationship	Analytical relationship

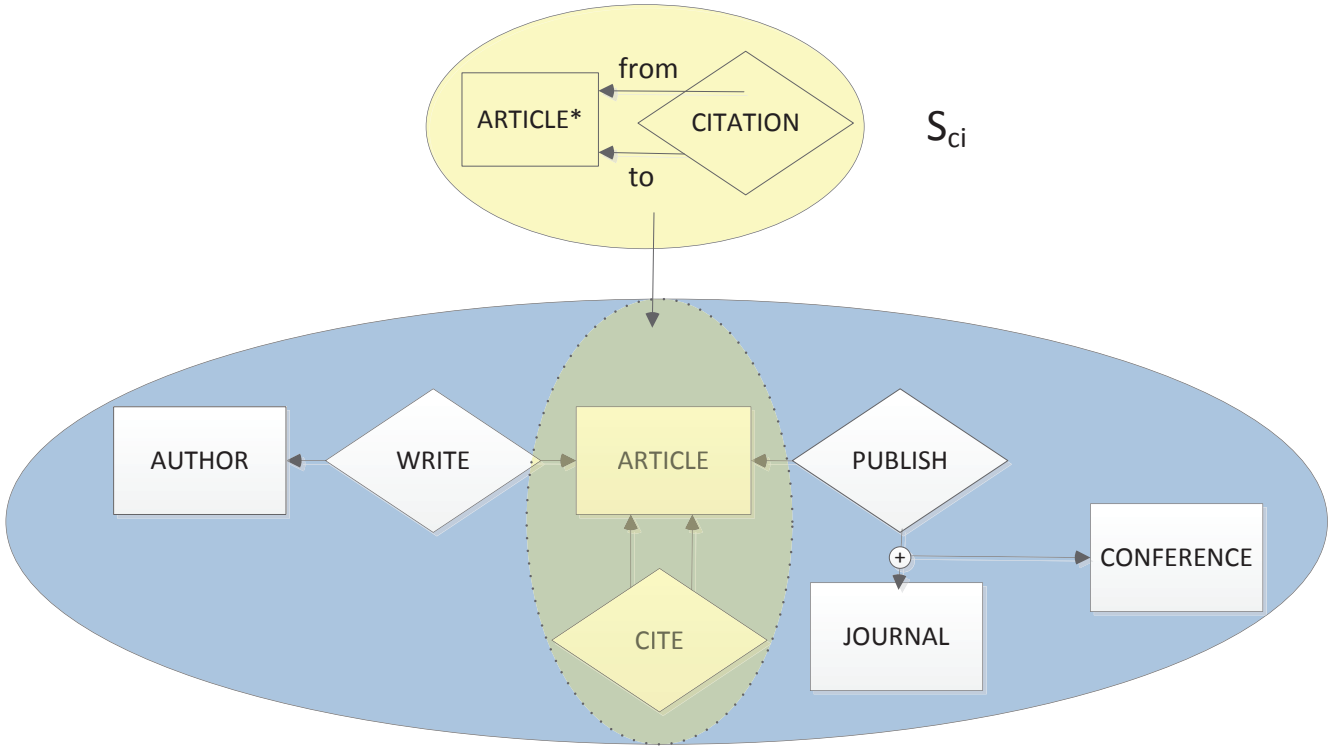
- Base types are the root from which analytical types can be derived.

The NAER Model - Example 1

- S_{co} for the query **collaborative communities**:
  - $supp(\text{AUTHOR}^*) = \{\text{AUTHOR}\}$
  - $supp(\text{COAUTHORSHIP}) = \{\text{AUTHOR}, \text{ARTICLE}, \text{WRITE}\}$.



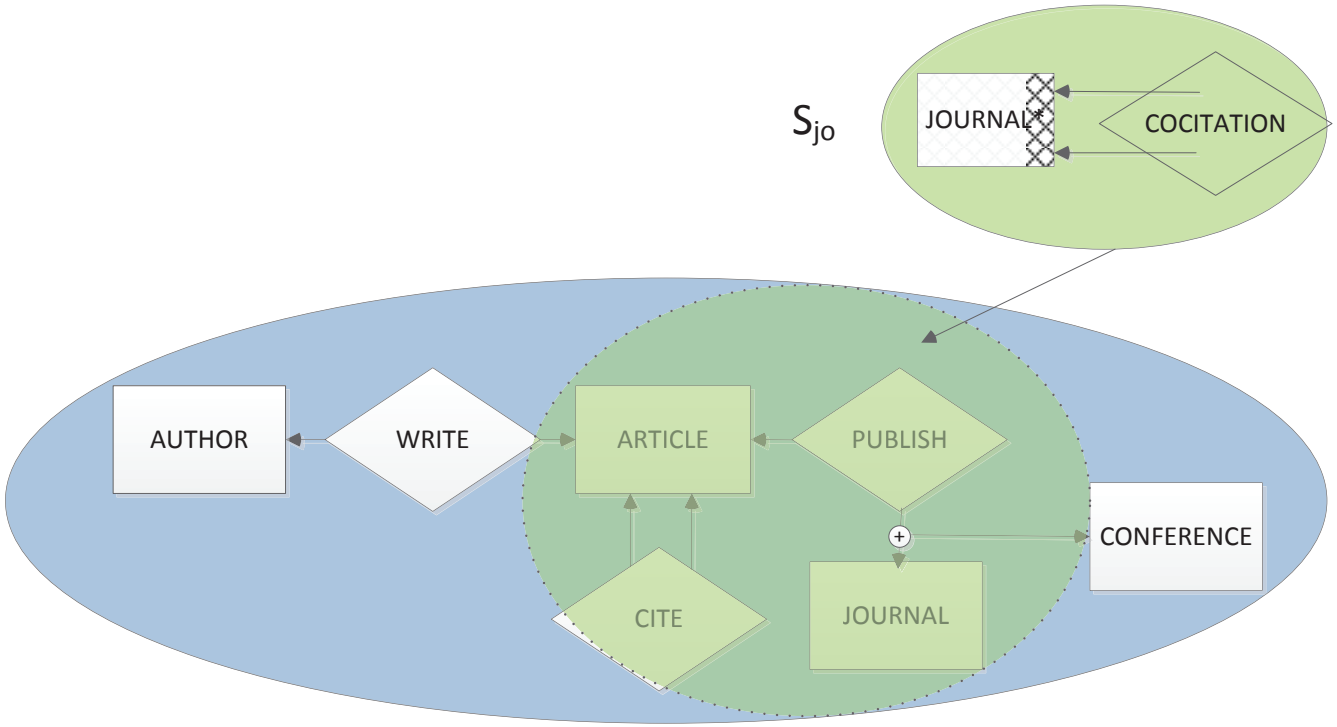
The NAER Model - Example 2



- S_{ci} for **most influential articles** and **top-k influential researchers**:
  - $supp(\text{ARTICLE}^*) = \{\text{ARTICLE}\}$
  - $supp(\text{CITATION}) = \{\text{ARTICLE}, \text{CITE}\}$

The NAER Model - Example 3

- S_{jo} for the query **correlation journal citations**:
  - $supp(\text{JOURNAL}^*) = \{\text{JOURNAL}\}$
  - $supp(\text{COCITATION}) = \{\text{ARTICLE}, \text{CITE}, \text{JOURNAL}, \text{PUBLISH}\}$



The NAER Model - Manipulation

- Using topological constructs to specify topological structures hidden underneath base entities and relationships.

  (1) CLUSTER-BY classifies elements into a set of clusters.

  (2) RANK-BY assigns rankings to elements.

- A topological measure is used in each topological construct,

  - **centrality** – CENT: $A \mapsto \mathbb{N}$ describing how central elements are in A, such as degree, betweenness and closeness centrality.

  - **similarity** – SIMI: $A \times A \mapsto \mathbb{N}$ describing the similarity between two elements in A, such as q-gram, adjacency-based and distance-based similarity.

The NAER Model - Examples

- Each **collaborative community** is a group of authors in a network over S_{co} measured by closeness centrality.

$$\text{CLUSTER-BY}(S_{co}, \text{AUTHOR}^*, \text{CENT-CLOSENESS}).$$

- The **influence of an article** is ranked, indicating its influence in terms of a network over S_{ci}, and measured by indegree centrality.

$$\text{RANK-BY}(S_{ci}, \text{ARTICLE}^*, \text{CENT-INDEGREE}).$$

- Each **correlation group** contains journals that are correlated in a network over S_{jo} and measured by betweenness centrality.

$$\text{CLUSTER-BY}(S_{jo}, \text{JOURNAL}^*, \text{CENT-BETWEENNESS}).$$

The NAER Model - Integrity

- Integrity constraints over topological constructs:
  - **disjoint** (resp. OVERLAPPING) on CLUSTER-BY
    Clusters must be disjoint (resp. can be overlapping).
  - **connected** on CLUSTER-BY
    For each cluster, there is a path between each pair of its members, running only through elements of the cluster.
  - **edge-density** on CLUSTER-BY
    For each cluster, its members have more edges inside the cluster than edges with other members who are outside the cluster.
  - **total** (resp. PARTIAL) on RANK-BY
    Every element must be (resp. may not necessarily be) ranked.

Analytical Framework

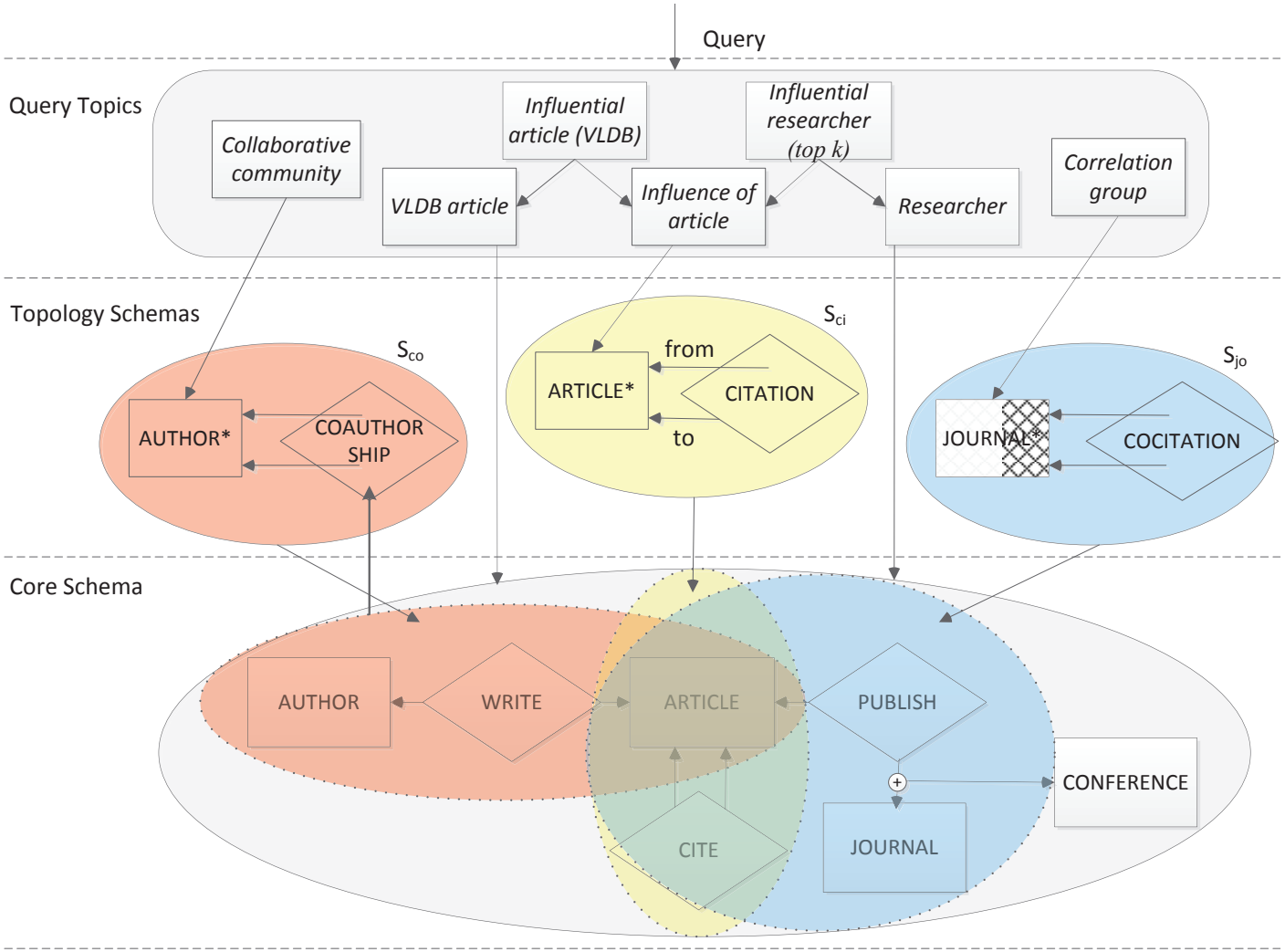
- Our analytical framework has three components:
  - A relatively large **core schema**

    i.e., base entity and relationship types
  - A number of small **topology schemas**

    i.e., analytical entity and relationship types
  - A collection of **query topics**

    i.e., trees, each representing a hierarchy of query object classes

Analytical Framework



Design Principles

- But, how should we design such an analytical framework in practice?

(1) Identify **data requirements**

(2) Design the core schema based on the data requirements

(3) Identify **query requirements**

(4) Design topology schemas based on the query requirements

(5) Identify constraints

Design Principles – Questions

Question I: What are data and query requirements?

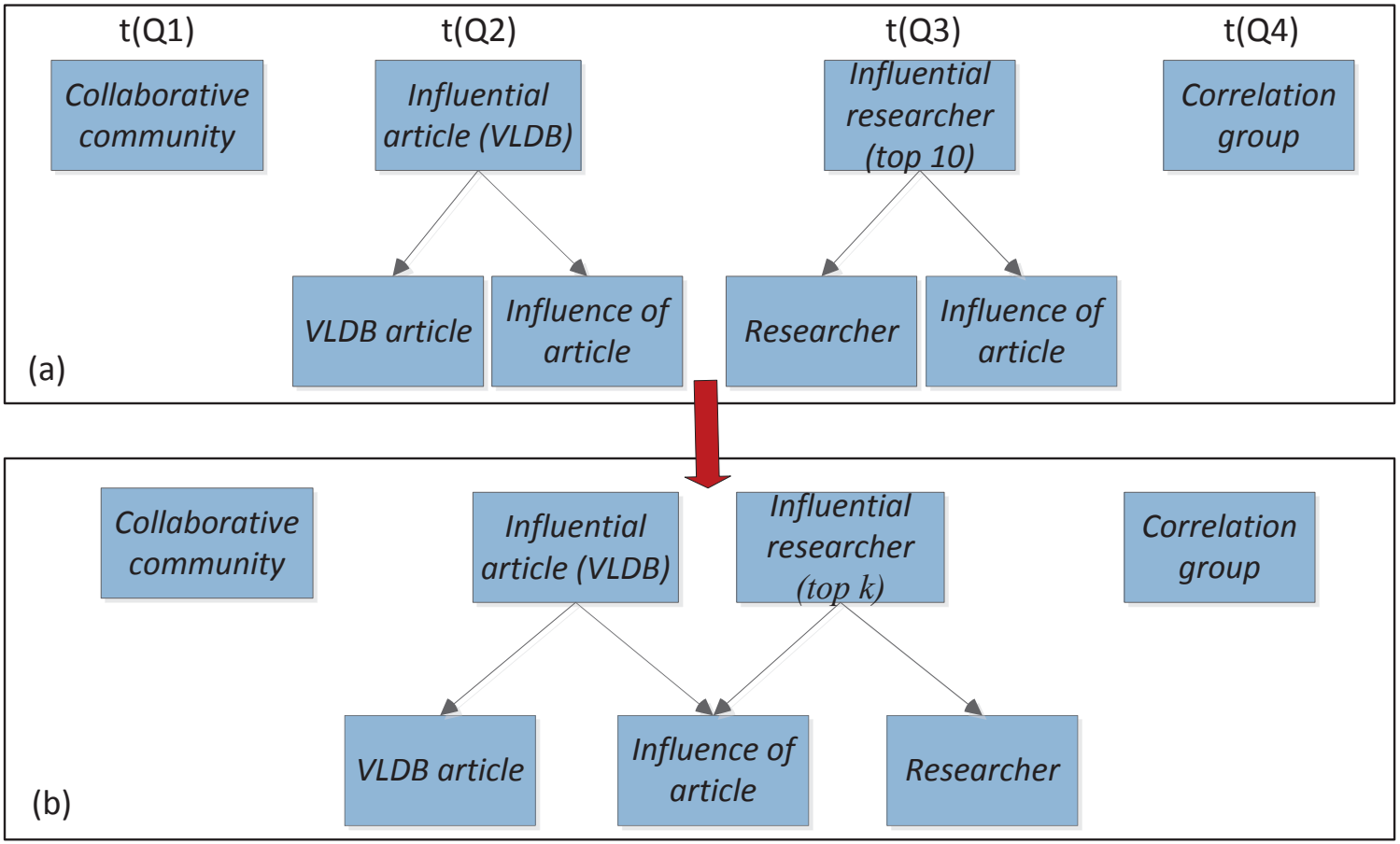
- Data and queries are two different kinds of requirements.
- Queries in NA applications may exist in various forms, e.g.,
  - **database queries** in the traditional sense
  - **analysis queries** from a topological perspective
  - a combination of database and analysis queries
- When designing a conceptual model for NA applications, we are particularly interested in analysis queries.

Design Principles – Questions

Question II: How are query requirements and query topics related?

- Queries need to be analyzed to unravel:
  - The **semantic structure** of **a query**
  - The **semantic structure** among **a set of queries**
- Each query Q is associated with a **query topic tree** $t(Q)$.
  - If $t(Q_1)$ and $t(Q_2)$ coincide over some nodes, then it means that two queries Q_1 and Q_2 are related.

Design Principles – Questions



Design Principles – Questions

Question III: How are the core and topology schemas designed?

- Central idea:

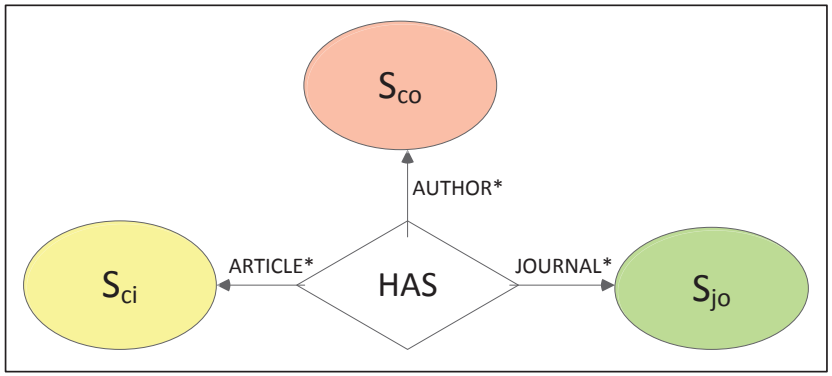
  (1) Data requirements should be captured by **the core schema**.

  (2) Query requirements should be captured by **a collection of topology schemas**.

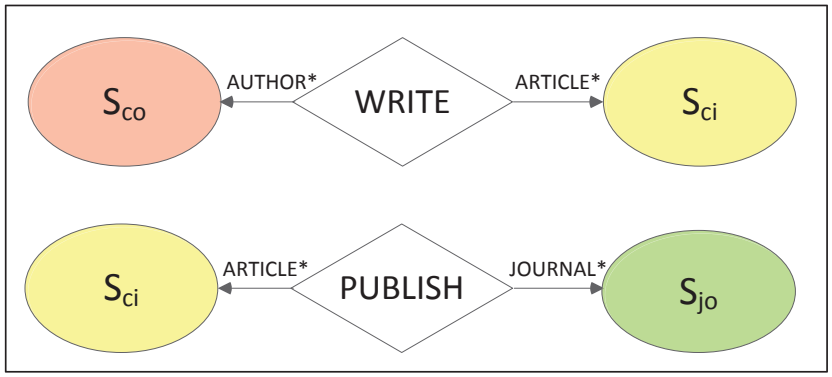
- Two criteria for designing topology schemas:
  - Topology schemas should be **small**.
  - Topology schemas should be **dynamic**.

Composition of Topology Schemas

(a) Composed through an analytical type, i.e., HAS



(b) Composed through a base type, i.e., WRITE and PUBLISH



Conclusions and Future Work

- We proposed the NAER model – a conceptual modelling paradigm that incorporates both data and query requirements of network analysis.
  - Enable us to better understand the semantics of data and queries, and how they interact with each other;
  - Avoid unnecessary computations in network analysis queries;
  - Support comparative network analysis.

- We plan to implement the NAER model over network analysis applications.
  - Establish an analytical framework;
  - Incorporate a query engine for processing topic-based queries.