Foundations of Comparative Analytics for Uncertainty in Graphs

Lise Getoor, University of Maryland

Alex Pang, UC Santa Cruz

Lisa Singh, Georgetown University

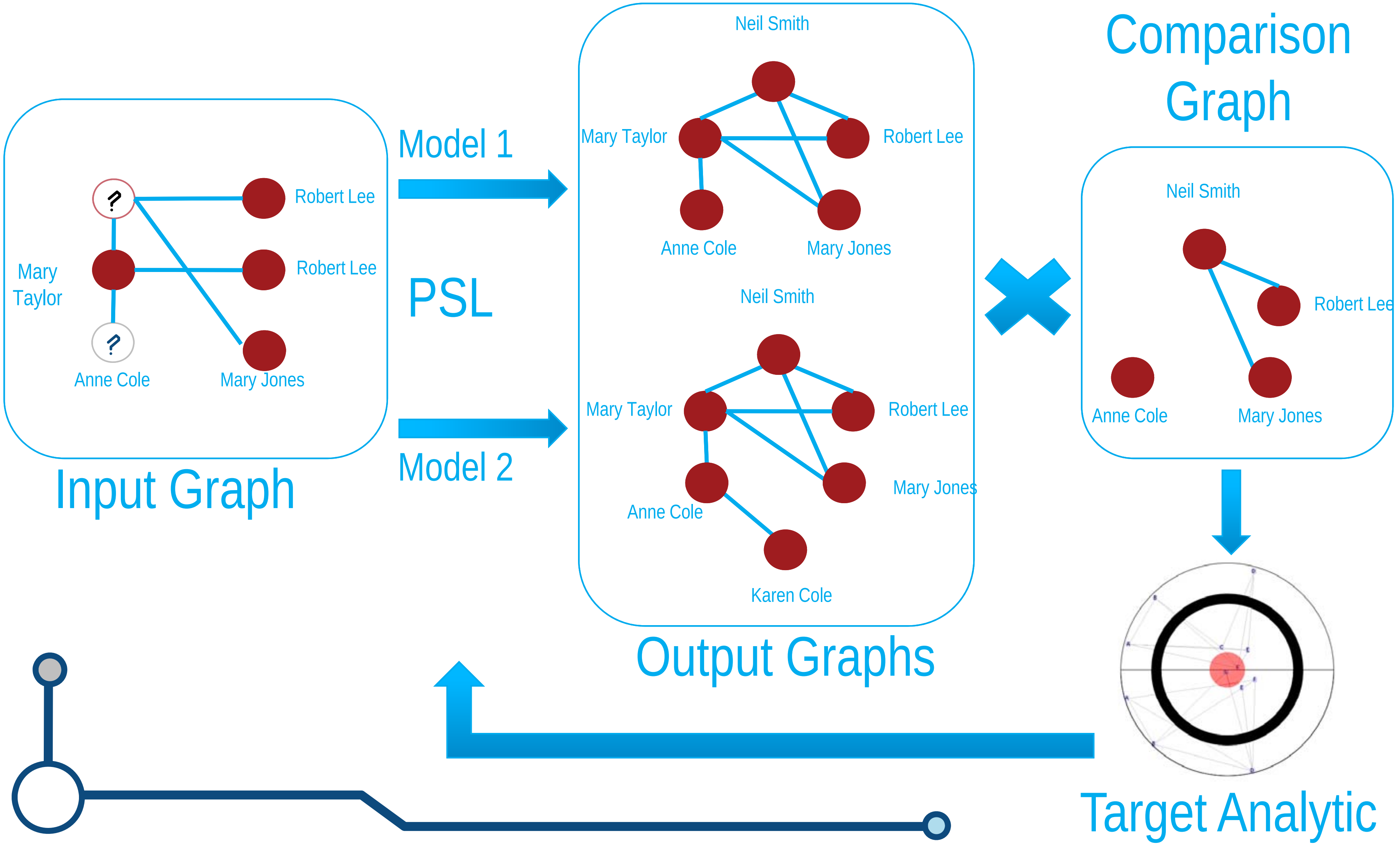
Motivation

- Input to analysis process is mix of structured, semi-structured and unstructured data
- Here, we focus on data that is best described as multi-modal, attributed graph or network
- Input to analysis process is often noisy and incomplete
- In addition, analytic process requires reasoning about similarity, uncertainty and logical conclusions

Needs

- **Mathematical models** which can infer missing values, infer links, and infer matches or duplicates in the data, and can capture the uncertainty and imprecision in the analytic process
- **Comparative analysis methods** that can contrasts the results of different models
- **Visual analytic tools** that support the understanding results of comparison and support the analyst in interactively updating the model/conclusions

The Big Picture



Outline

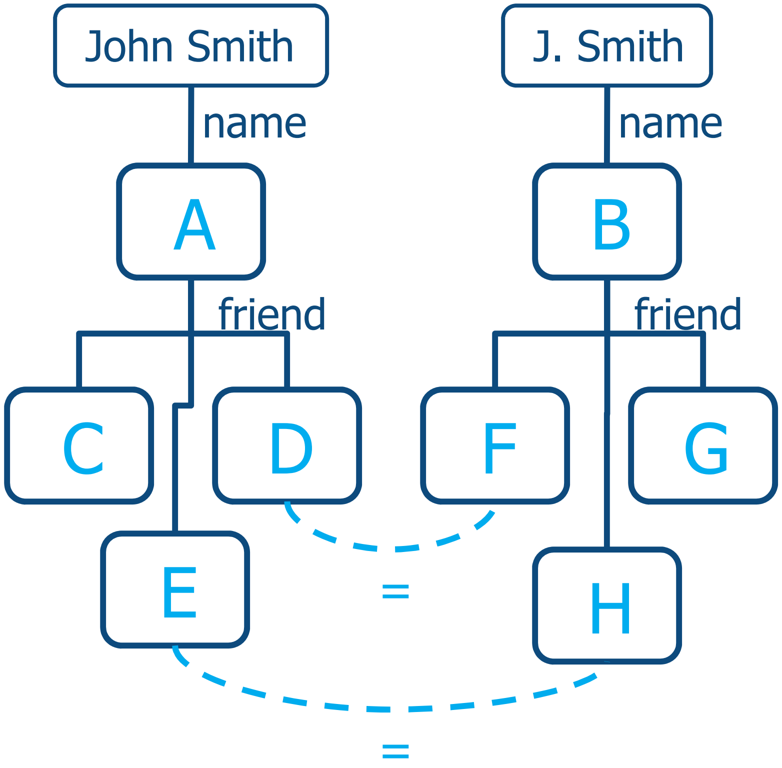
- Motivation
- Mathematical Foundations for Uncertainty in Graphs
  - **Probabilistic Similarity Logic (PSL)**
- Comparative Analysis
- Visual Analytic Support
- Application Domains

Why PSL?

- Collective Reasoning under Uncertainty
  - Combining probabilistic and logical inference
- Reasoning about Similarity
  - Degrees of Similarity vs. Bivalent Logic
- Reasoning with Sets of Objects
- Simplicity, "Vanilla"-version → usability
- Scalability for large data sets
- Integration Framework

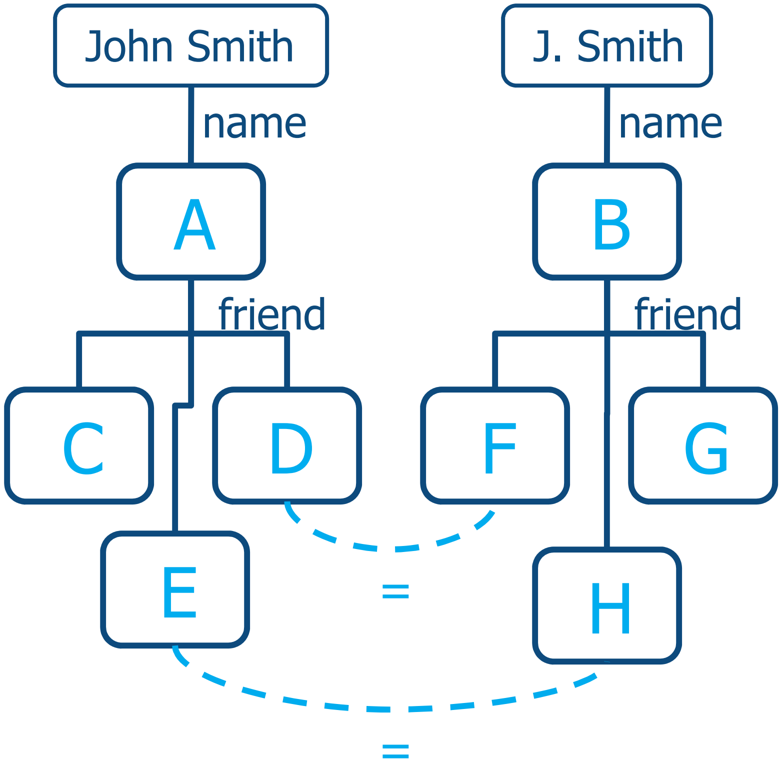
Ex. 1: Entity Resolution

- Entities
  - People
- Attributes
  - Name
- Relationships
  - Friendship



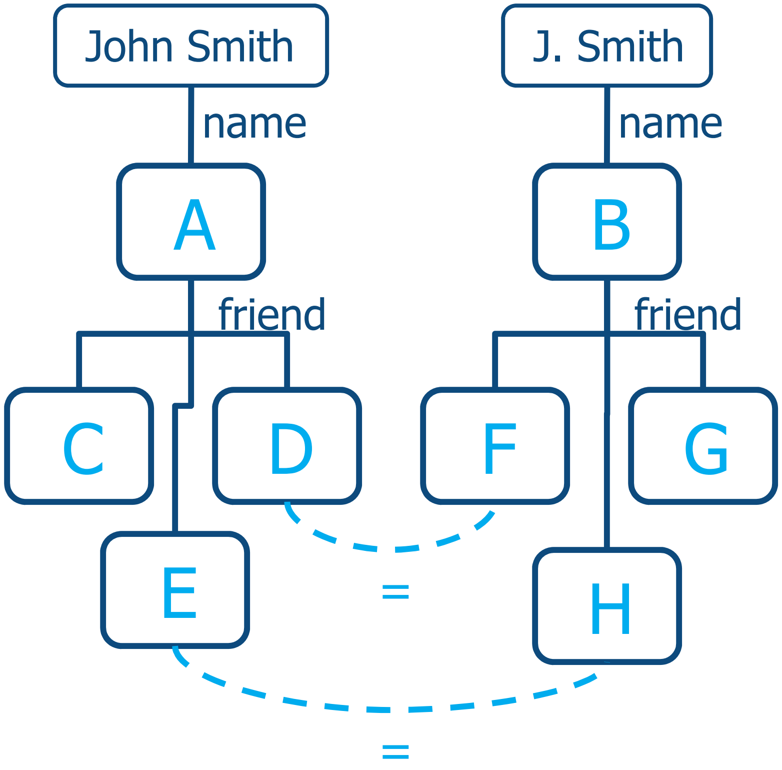
Example: Entity Resolution

- Entities, attributes, relationships
- Use rules to express evidence
  - Modular, simple
  - "If two people have the same name, they are probably identical"
  - "If two people have the same friends, they are probably identical"
  - "If A=B and B=C, then A and C must also denote the same person"



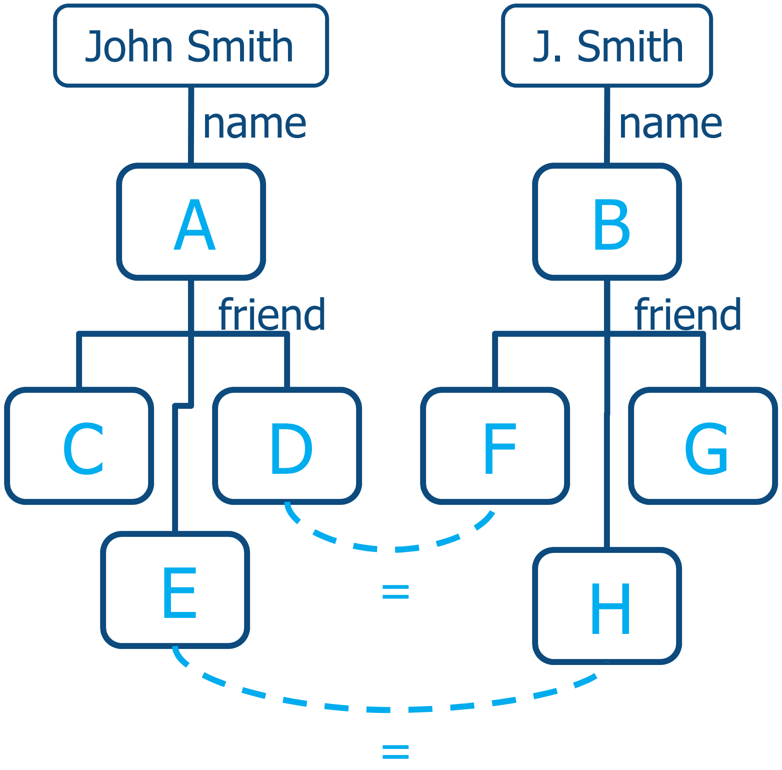
Example: Entity Resolution

- Entities, attributes, relationships
- Use rules to express evidence
  - Modular, simple
  - ”If two people have the same name, they are probably identical”
  - “If two people have the same friends, they are probably identical”
  - “If A=B and B=C, then A and C must also denote the same person”



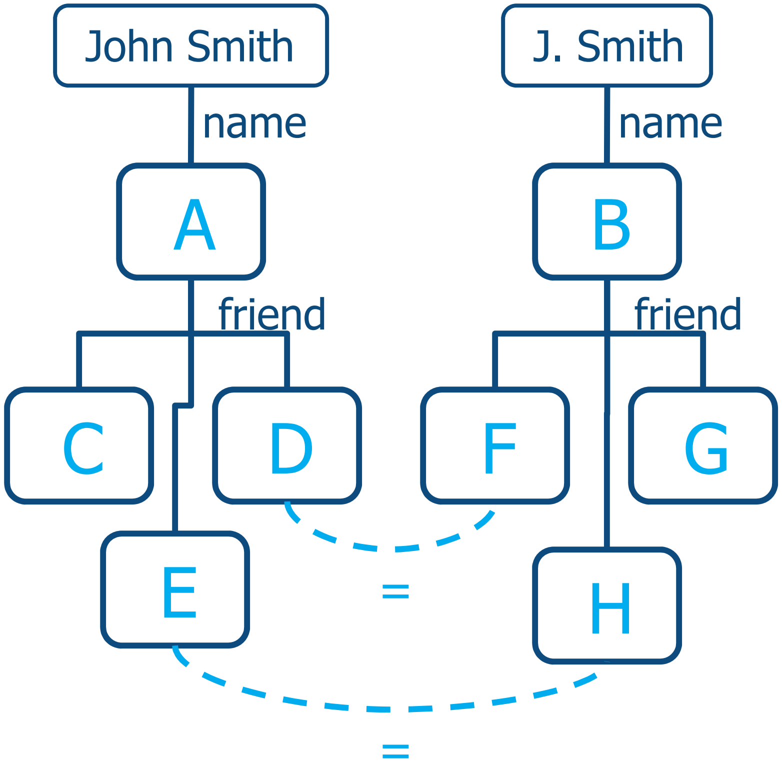
Example: Entity Resolution

- Entities, attributes, relationships
- Use rules to express evidence
  - Modular, simple
  - '"f two people have the same name, they are probably identical"
  - "If two people have the same friends, they are probably identical"
  - "If A=B and B=C, then A and C must also denote the same person"



Example: Entity Resolution

- Entities, attributes, relationships
- Use rules to express evidence
  - Modular, simple
  - "If two people have the same name, they are probably identical"
  - "If two people have the same friends, they are probably identical"
  - "If A=B and B=C, then A and C must also denote the same person"



Syntax Components

- Rules + Weights
  - A / B ^{2} C : w , w real number
- Rules defines evidence
  - Soft Evidence: “If X then likely Y”
    - 0 < w < ∞
  - Conclusive Evidence: “If X then definitely Y”
    - w = ∞
  - Modularized: A model is a set of rules
  - Humanly understandable
- Weight specifies relative probability

Addressing Entities

- Use relational syntax
  - X.name
  - X.father
  - X.friend   (a friend)
- Explicitly handle sets
  - {X.friend} - all friends
  - {X.friend.friend} - all second level friends
  - X.friend.{friend} - all friends of a friend

Example

- X.name $=_s$ Y.name => X = Y : 5
  - Implicit universal quantification
  - $=_s$ denotes a string similarity function
- {X.friend} $=_{\{\}}$ {Y.friend} => X = Y : 3
  - $=_{\{\}}$ denotes a set similarity function

Addressing Entities

- Entity Addressing can consider inferred relationships or be restricted to known ones.
  - Atoms for ''closed'' predicates are always assumed to be known. ''Open'' predicates are subject to inference.

$\{A.\mathrm{groups}\} =_{\{\}} \{B.\mathrm{groups}\} \mathrel{=>} \mathrm{friend}(A,B) : 2$

$\{A.\mathrm{friend}\} =_{\{\}} \{B.\mathrm{friend}\} \mathrel{=>} A=B : 3$

- Consider inferred

$\{A.\$\mathrm{friend}\} =_{\{\}} \{B.\$\mathrm{friend}\} \mathrel{=>} A=B : 4$

- Consider only known

Advanced Addressing

- Qualifications
  - {?X.friend[age>50]}
  - {?Y.friend[gender=female].friend}
  - Like "where" clauses
- Catch-all Global Addressing
  - {?A.friend} = {*[age>65]} => ?A.type=old_representative
- Catch-all relations with qualifications
  - {?X.*[type=association]}={?Y.*[type=association]}

Constraints

- Predicate properties
  - Child = inverse(parent)
  - symmetric(friend)
- Exclusivity Constraints
  - Needed e.g. in alignment problems
  - functional(hasLabel)
    - Each entity is assigned 1 label
  - partialFunctional(equalConcept)
    - Each concept is equivalent to at most one other.

Truth Combiner Functions

- Need to combine truth values for multiple atoms
  - A / B ² C 1 D
- Lukasiewicz T-Norm
  - T(A / B) = max(T(A)+T(B)-1 ,0)
  - T(C 1 D) = min(T(C)+T(D) ,1)

PSL Inference

- Satisfaction Distance
- P = set of rules, KB

All ground rules

- $d(P,I) = \|d(\vec{R},I)\|_x = \left\| \begin{bmatrix} d(R_1,I) \\ \vdots \\ d(R_n,I) \end{bmatrix} \right\|_x$
- $s(\, I \mid P) = {}^1/_Z \exp\left(- d(P,I)\right)$

MAP Inference

- Most Probable Interpretation
  - Most likely truth value assignment given some facts.

$$\underset{I}{\operatorname{argmax}}\ s(\, I \mid P)$$

$$\cdots$$

$$\underset{I}{\operatorname{argmin}}\ d(P,I)$$

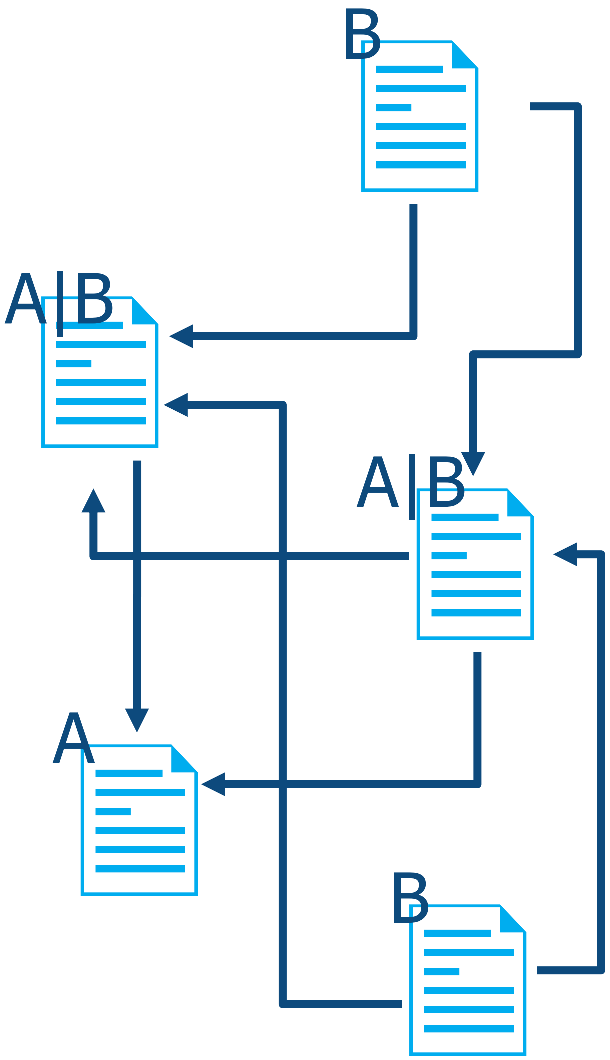
MAP Inference Results

- Exact PSL inference in polynomial time
  - Convex optimization problem

- O($n^{3.5}$) inference for PSL fragment
  - Second Order Cone Program
  - Efficient commercial optimization packages

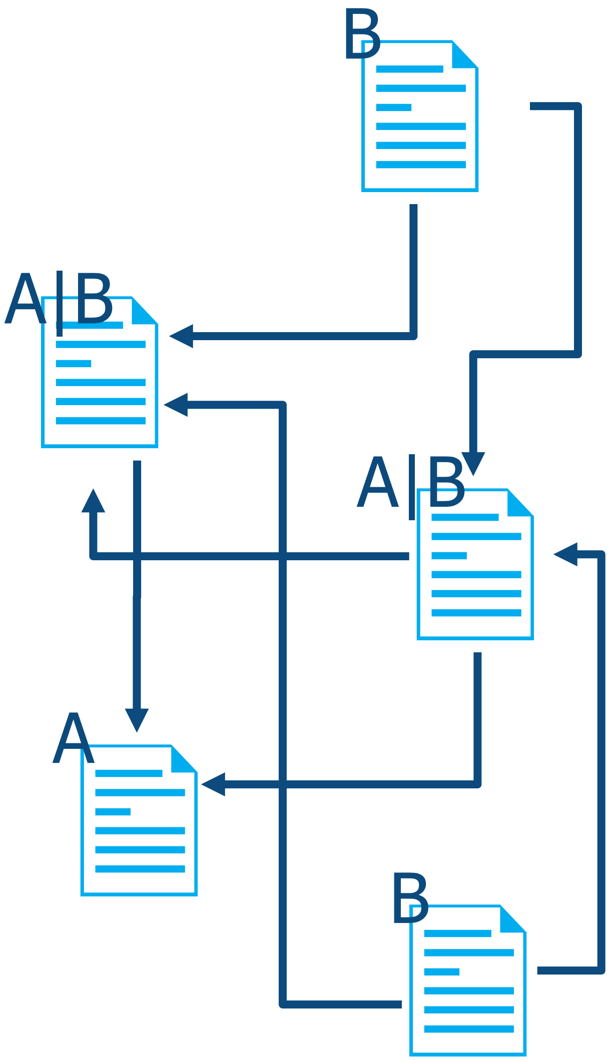
Ex. 2: Collective Classification

- Entities
  - Documents
- Attributes
  - Word occurrence within document
- Relationships
  - Citations
- Goal: Classify documents
  - Fixed number of topics
  - Allow multi-membership



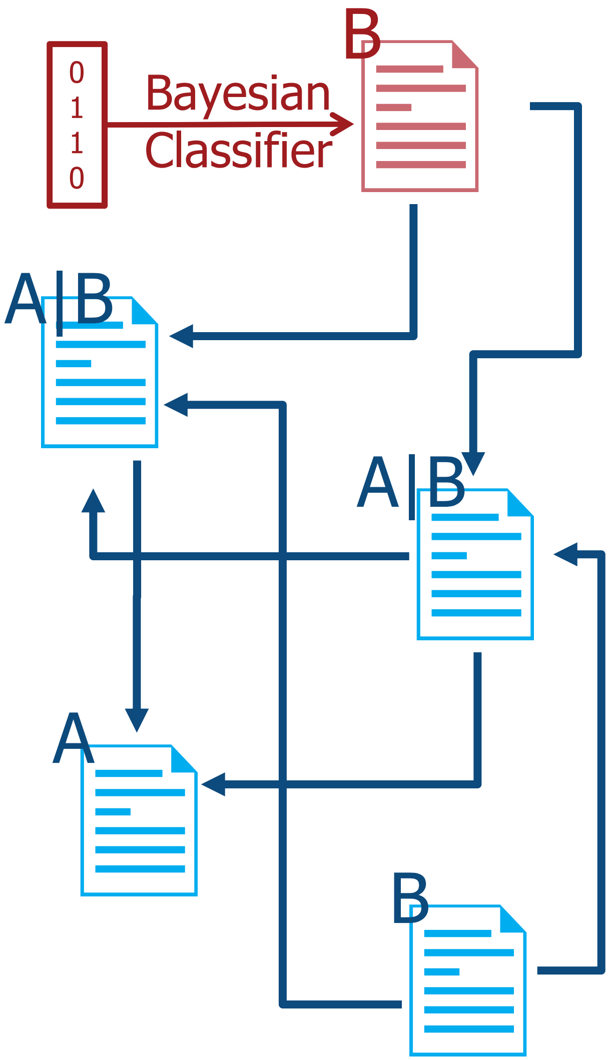
Collective Classification

- Documents, words, links
- Use rules to express evidence
  - "If an attribute-based classifier predicts a document's topic to be X, then it is X"
  - "If a document has topic X, then the majority of documents it links to are also classified as X"
  - "If a document has topic X, then any document that refers to it is also of topic X"



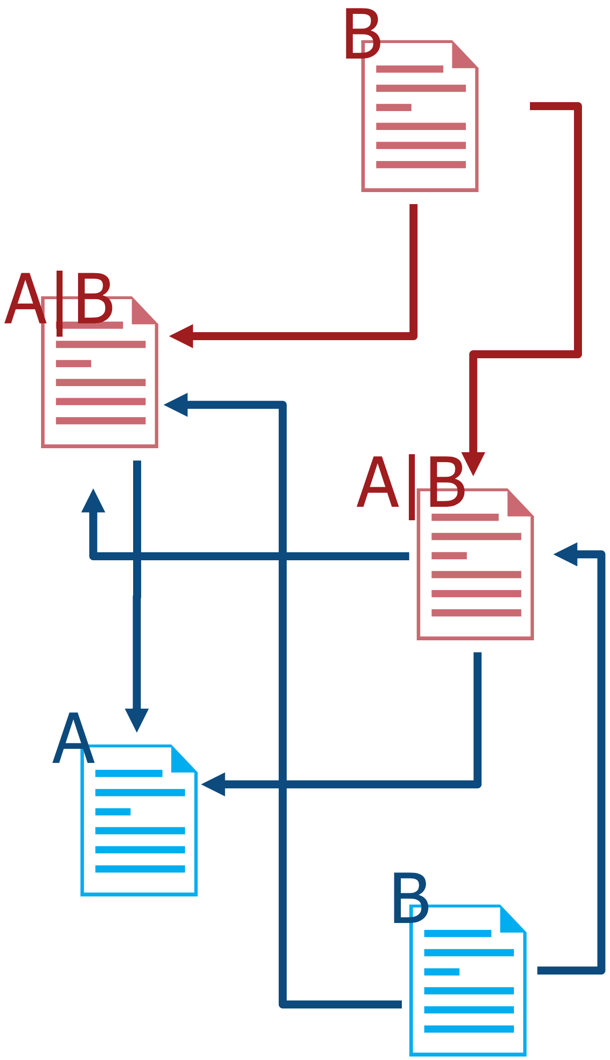
Collective Classification

- Documents, words, links
- Use rules to express evidence
  - "If an attribute-based classifier predicts a document's class to be X, then it is X"
  - "If a document has topic X, then the majority of documents it links to are also classified as X"
  - "If a document has topic X, then any document that refers to it is also of topic X"



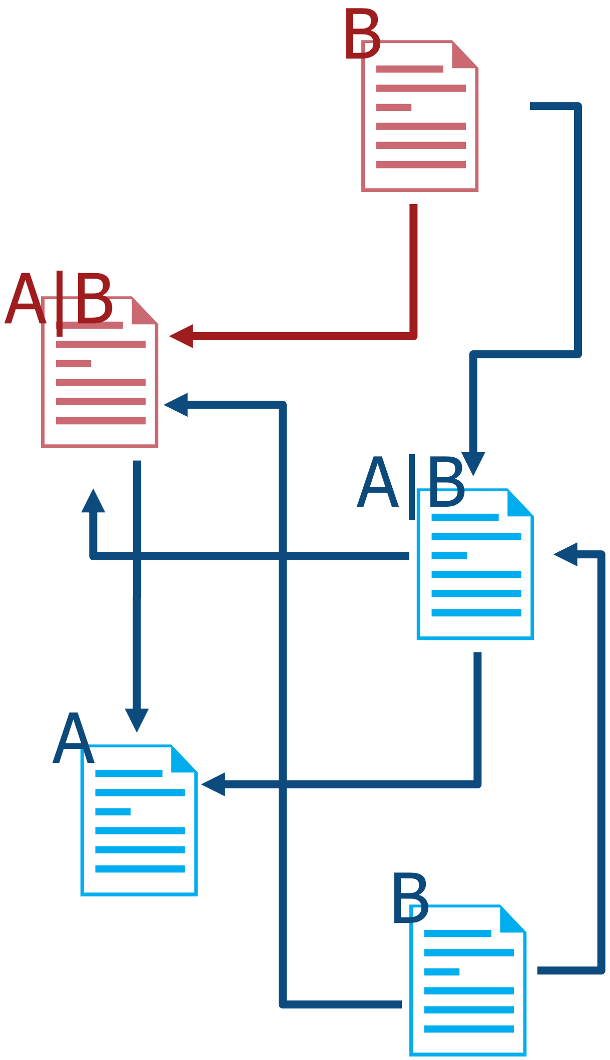
Collective Classification

- Documents, words, links
- Use rules to express evidence
  - "If a classifier predicts a document's topic to be X, then it is X"
  - "If a document has topic X, then the majority of documents it links to are also classified as X"
  - "If a document has topic X, then any document that refers to it is also of topic X"



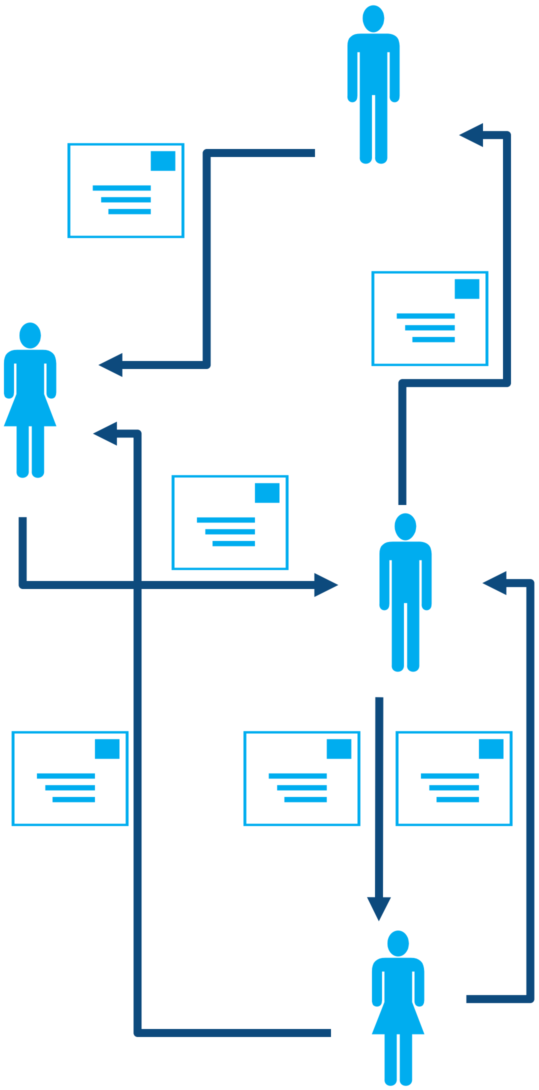
Collective Classification

- Documents, words, links
- Use rules to express evidence
  - "If a classifier predicts a document's topic to be X, then it is X"
  - "If a document has topic X, then the majority of documents it links to are also classified as X"
  - "If a document has topic X, then any document that refers to it is also of topic X"



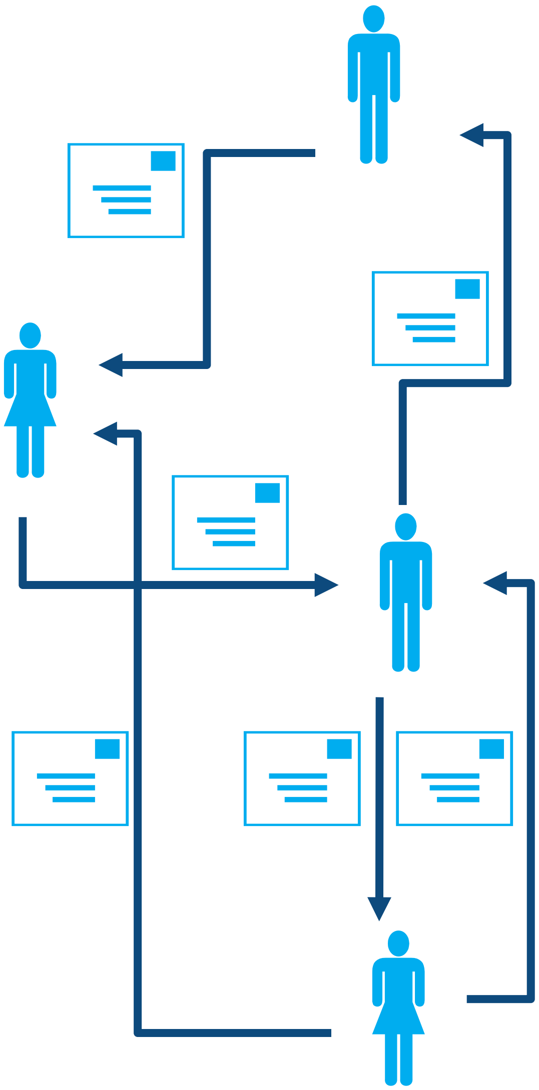
Ex. 3: Link Prediction

- Entities
  - People, Emails
- Attributes
  - Words in emails
- Relationships
  - communication, work relationship
- Goal: Identify work relationships
  - Supervisor, subordinate, colleague



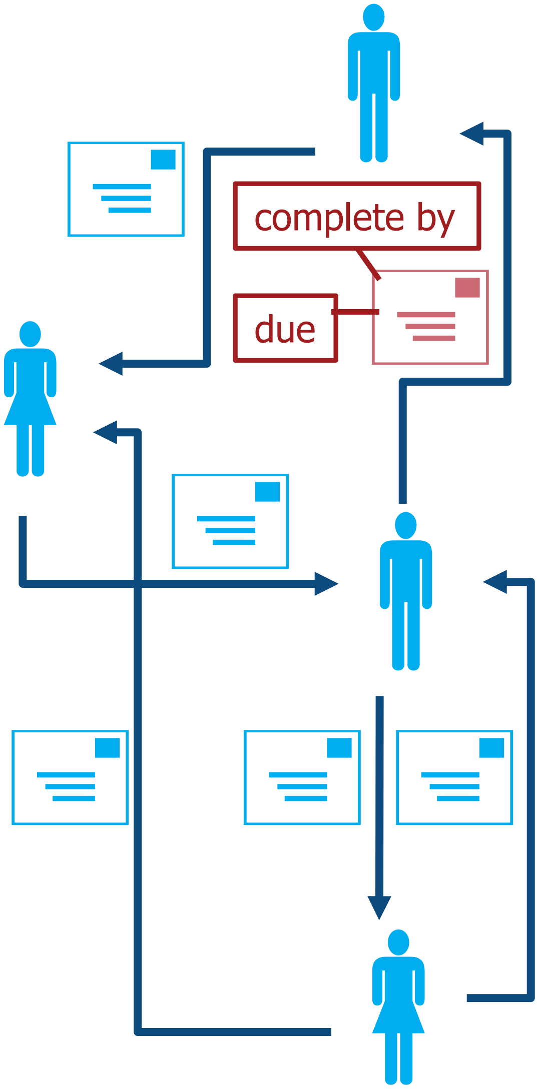
Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If an email is classified as type X, it is of type X"
  - "If A sends deadline emails to B, then A is the supervisor of B"
  - "If A is the supervisor of B, and A is the supervisor of C, then B and C are colleagues"



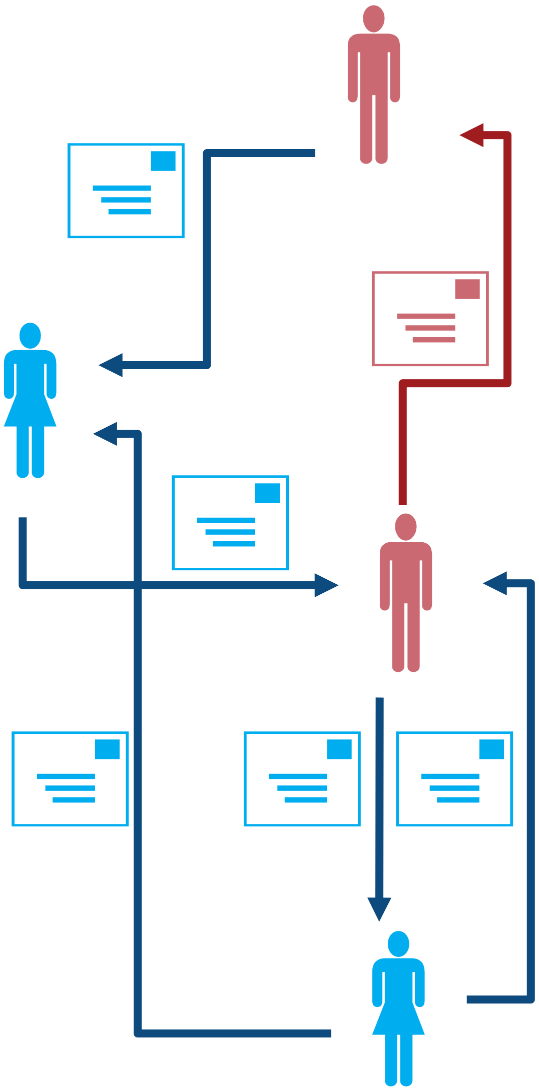
Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If an email is classified as type X, it is of type X"
  - "If A sends deadline emails to B, then A is the supervisor of B"
  - "If A is the supervisor of B, and A is the supervisor of C, then B and C are colleagues"



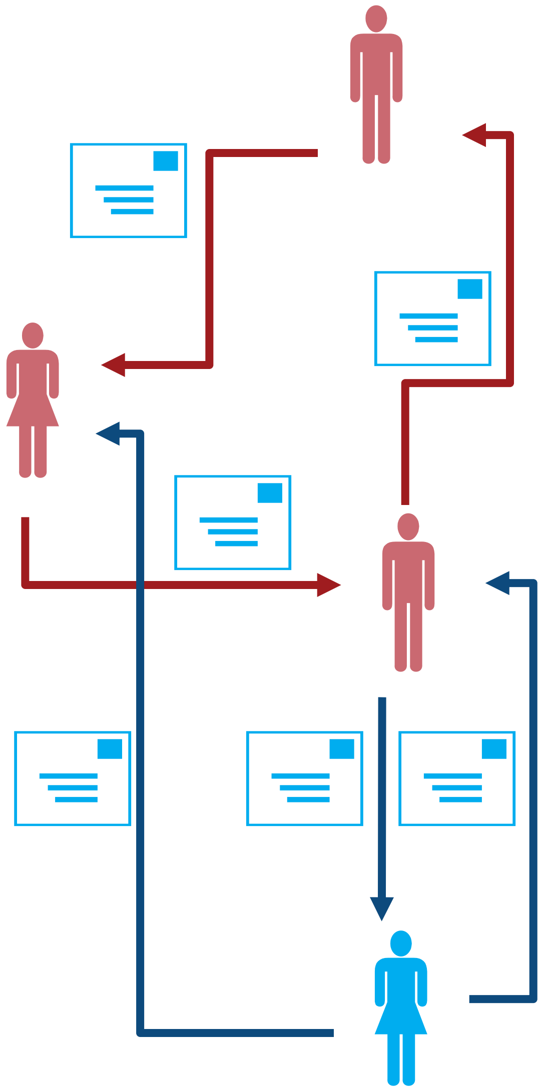
Link Prediction

- People, emails, words, commuication, relations
- Use rules to express evidence
  - “If an email is classified as type X, it is of type X”
  - “If A sends deadline emails to B, then A is the supervisor of B”
  - “If A is the supervisor of B, and A is the supervisor of C, then B and C are colleagues”



Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If an email is classified as type X, it is of type X"
  - "If A sends deadline emails to B, then A is the supervisor of B"
  - "If A is the supervisor of B, and A is the supervisor of C, then B and C are colleagues"



Outline

- Motivation
- Mathematical Foundations for Uncertainty in Graph
  - Probabilistic Similarity Logic (PSL)
- **Comparative Analysis**
- Visual Analytic Support
- Application Domains
- Research Plan

Quantifying Uncertainty in Graphs

- Types of uncertainty
  - Attribute uncertainty
  - Link Uncertainty
  - Entity Uncertainty

- Want to compare distributions
  - Over attribute values
  - Link probabilities
  - Equivalence of objects

Comparative Analysis

- Our comparative operators are expressed using a graph algebra.
- We can compare posterior probabilities of nodes, edges and/or attributes.
- Basic operators serve as building blocks for more complex ones.
- Ranking
  - Unary operator that orders nodes, edges or attributes based on posterior probability, variability, etc.

Comparative Operators

- **Difference**

  Given two uncertain graphs G1 and G2, compute a resultant graph that contains nodes and edges that have a difference in posterior probabilities greater than threshold τ

- **Intersection**

  Given two uncertain graphs G1 and G2, compute a resultant graph that contains nodes and edges that have a difference in posterior probabilities greater than threshold τ

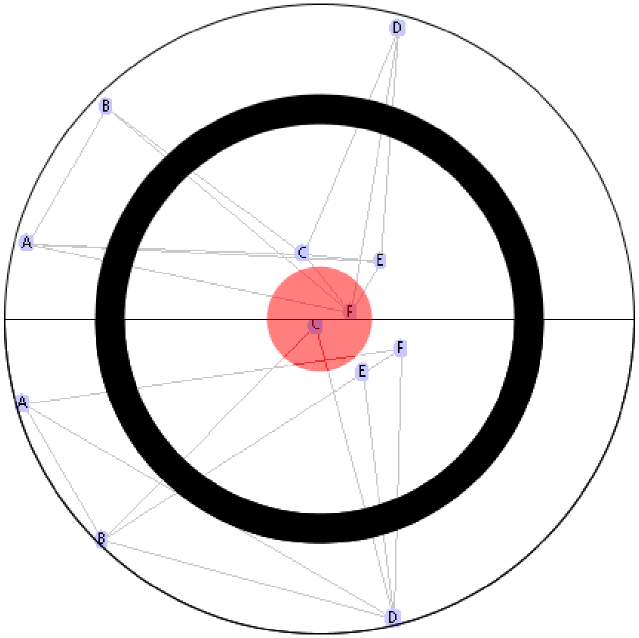
Outline

- Motivation
- Mathematical Foundations for Uncertainty in Graph
  - Probabilistic Similarity Logic (PSL)
- Comparative Analysis
- **Visual Analytic Support**
- Application Domains

Visualization

- Developing open source visual analytic platform for comparing graphs. Platform being built using open source toolkits, Prefuse and Jung.

- Developing specialized visualizations that focus on comparing local uncertainty. We are currently exploring a bullseye metaphor.



Outline

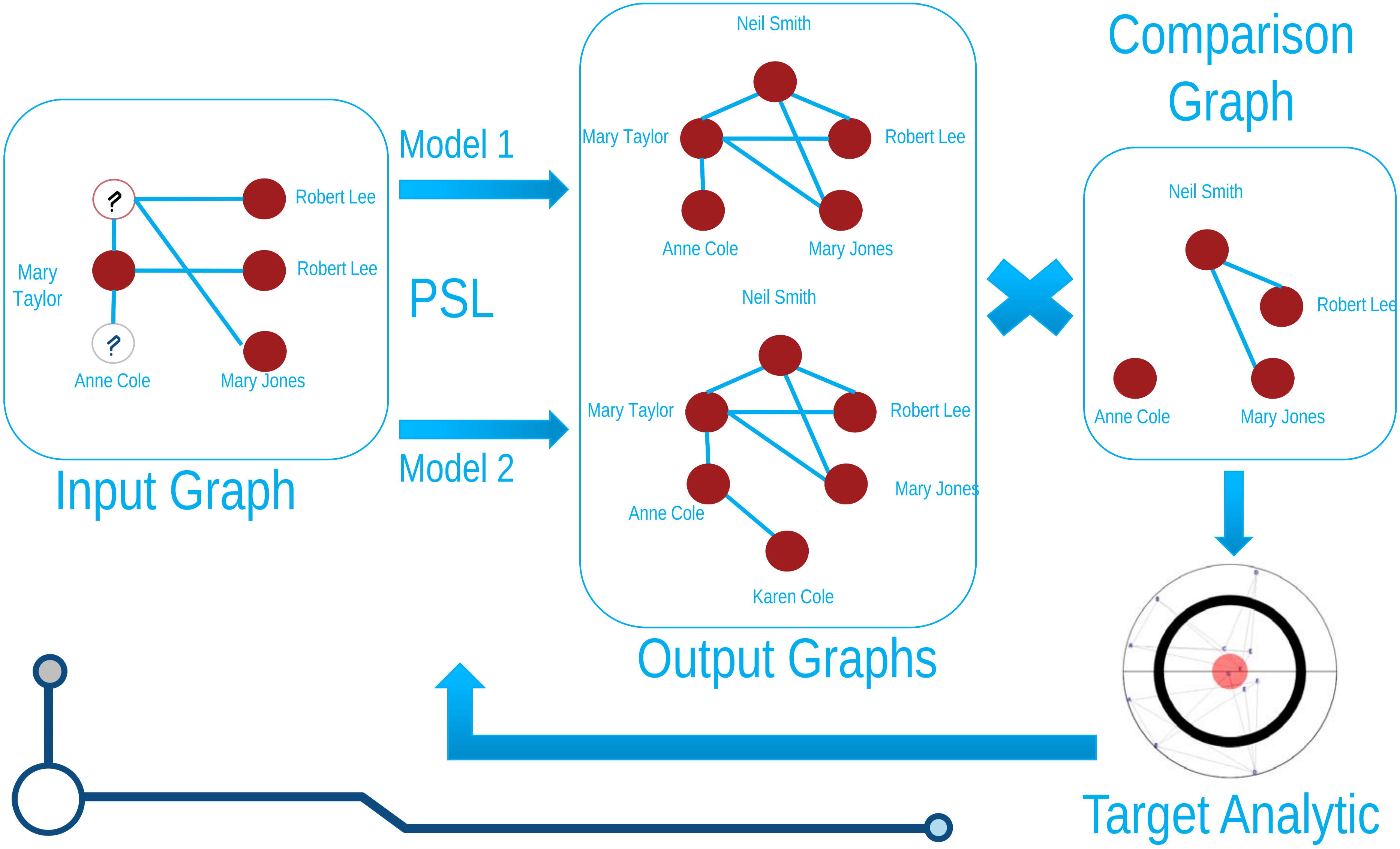
- Motivation
- Mathematical Foundations for Uncertainty in Graph
  - Probabilistic Similarity Logic (PSL)
- Comparative Analysis
- Visual Analytic Support
- **Application Domains**

Shark Bay Dolphin Research Project Overview

- Dolphins monitored by international team of scientists since 1984.
  - 14000 surveys
  - Thousands of hours of focal follows
  - Thousands of pictures
  - GIS spatial data

© Janet Mann

Summary



Input Graph

Model 1

PSL

Model 2

Output Graphs

Comparison Graph

Target Analytic

Questions?
Feedback?