Foundations of Comparative Analytics for Uncertainty in Graphs

Lise Getoor, University of Maryland

Alex Pang, UC Santa Cruz

Lisa Singh, Georgetown University

Students: Steve Bach, Matthias Broecheler, Hossam Sharara, Galileo Namata, Nathaniel Cesario, Awalin Sopan, Denis Dimitrov, Katarina Yang

Objectives

- Develop mathematical models for capturing uncertainty in graphs:
  - node merging uncertainty (entity resolution)
  - edge existence uncertainty (link prediction)
  - node label uncertainty (collective classification)
- Develop visual analytic tools for comparative analysis of uncertainty such models

Proposed Approaches

- Uncertainty in Graphs: Foundations
  - **Probabilistic Soft Logic (PSL)**
  - http://psl.umiacs.umd.edu/
- Uncertainty in Graphs: Comparative Analytics
  - **G-Pare (Graph Compare)**
  - http://www.cs.umd.edu/projects/linqs/gpare

PSL Foundations

- **Declarative language** based on logic to express collective probabilistic inference problems
- **Probabilistic Model**
  - Undirected graphical model
  - Constrained Continuous Markov Random Field (CCMRF)
- **Key distinctions**
  - Continuous-valued random variables
  - Efficiently compute similarity & propagate similarity
  - Ability to efficiently reason about sets and aggregates
  - Scalable inference using consensus optimization

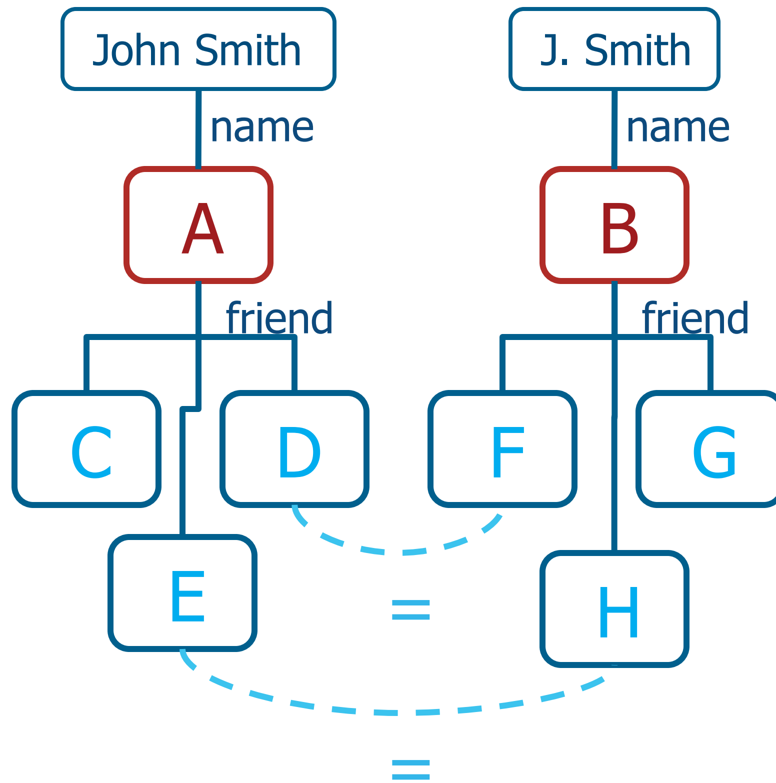
What is PSL Good for?

- Specifying probabilistic models for:
  - Information Alignment
  - Information Fusion
  - Information Diffusion
- Each of these requires:
  - Entity resolution
  - Link prediction
  - Node Labeling

Recent applications:
•Sentiment Analysis
•Models of Group Affiliation
•Graph Summarization
•Role Identification in Online Discussions

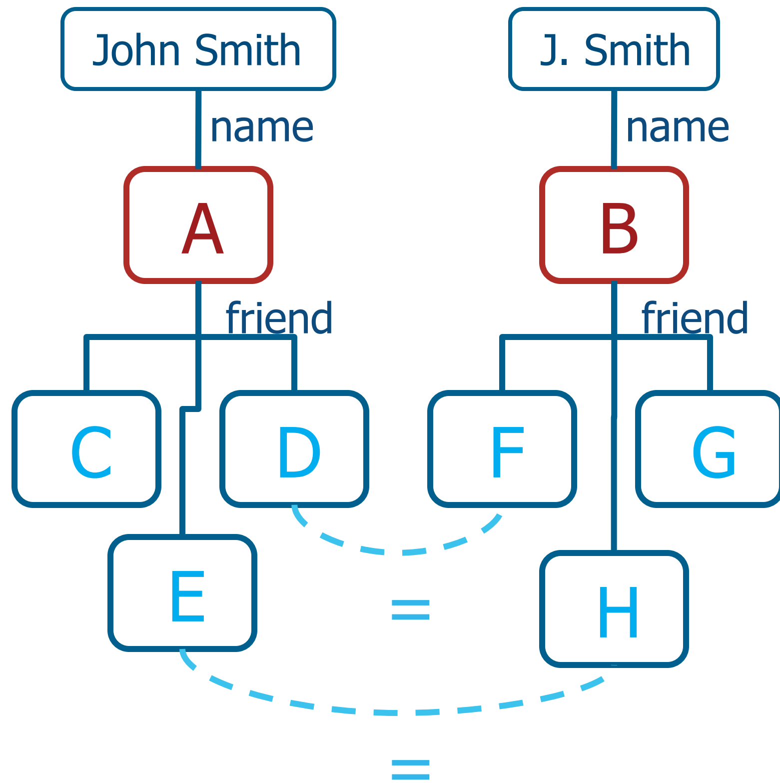
Entity Resolution

- Entities
  - People References
- Attributes
  - Name
- Relationships
  - Friendship
- Goal: Identify references that denote the same person



Entity Resolution

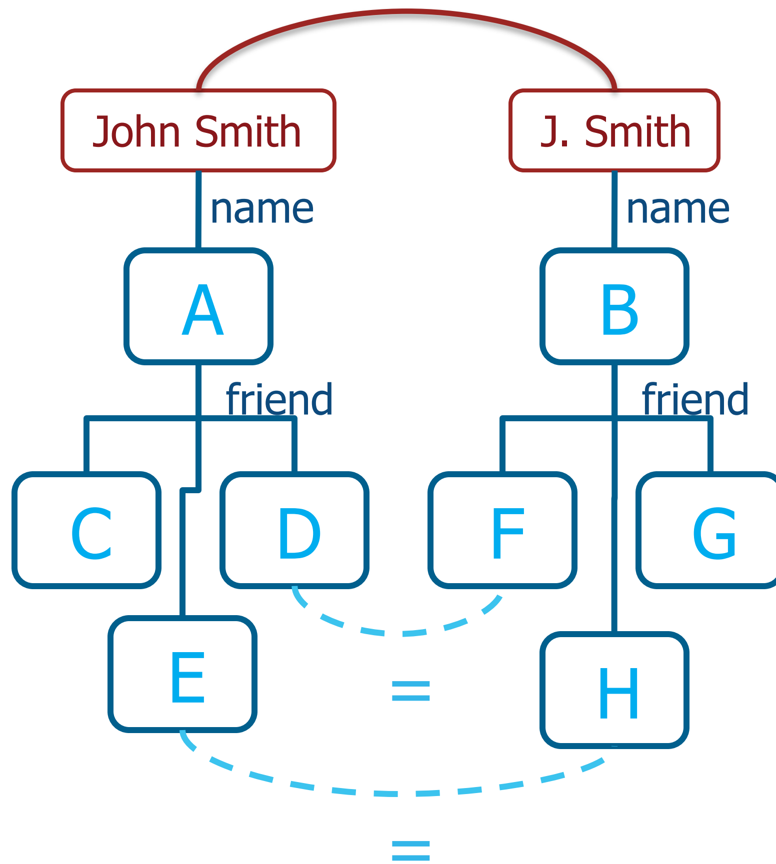
- References, names, friendships
- Use rules to express evidence
  - ‘’If two people have similar names, they are probably the same’’
  - ‘’If two people have similar friends, they are probably the same’’
  - ‘’If A=B and B=C, then A and C must also denote the same person’’



Entity Resolution

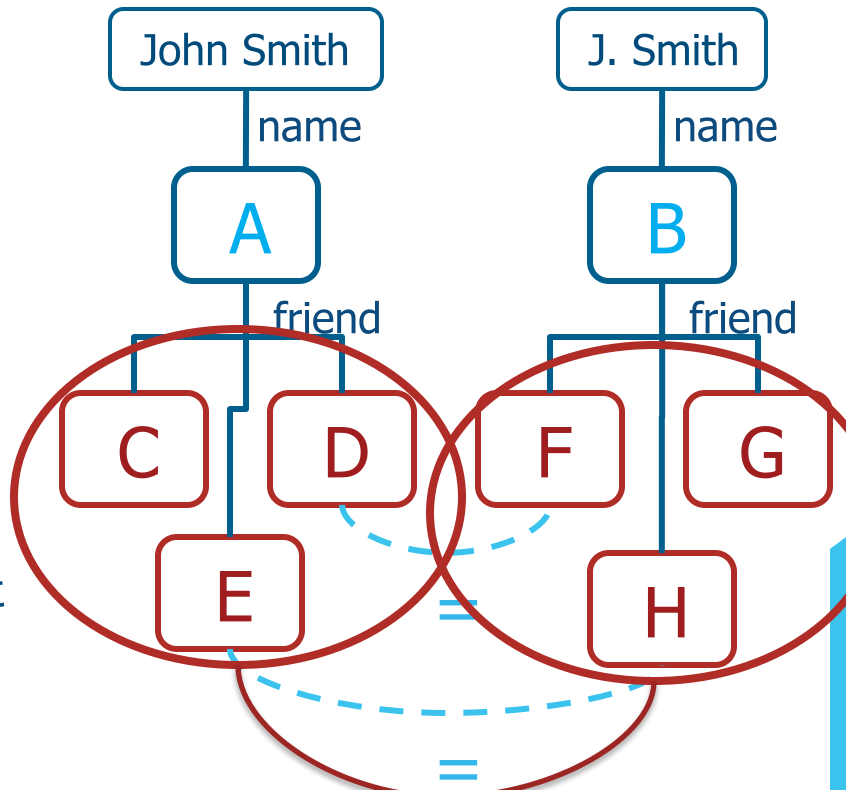
$\text{A.name} \approx_{\{str\_sim\}} \text{B.name} \Rightarrow \text{A} \approx \text{B} : 0.8$

- References, names, friendships
- Use rules to express evidence
  - ''If two people have similar names, they are probably the same''
  - ''If two people have similar friends, they are probably the same''
  - ''If A=B and B=C, then A and C must also denote the same person''



Entity Resolution

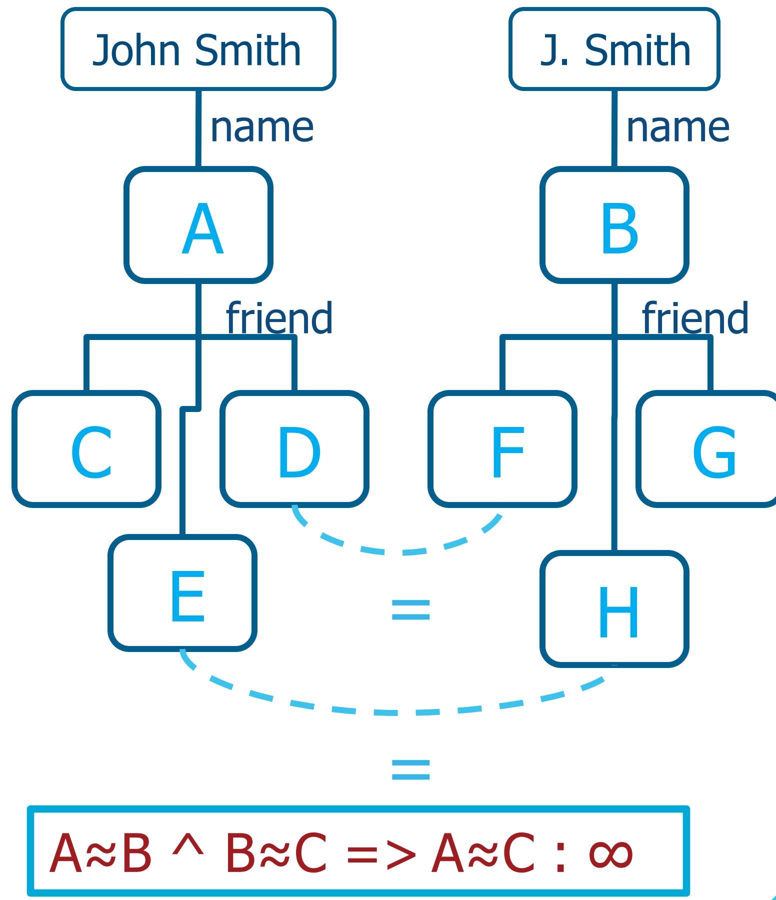
- References, names, friendships
- Use rules to express evidence
  - ''If two people have similar names, they are probably the same''
  - ''If two people have similar friends, they are probably the same''
  - ''If A=B and B=C, then A and C must also denote the same person''



$\{A.friends\} \approx_{\{\}} \{B.friends\} \Rightarrow A \approx B : 0.6$

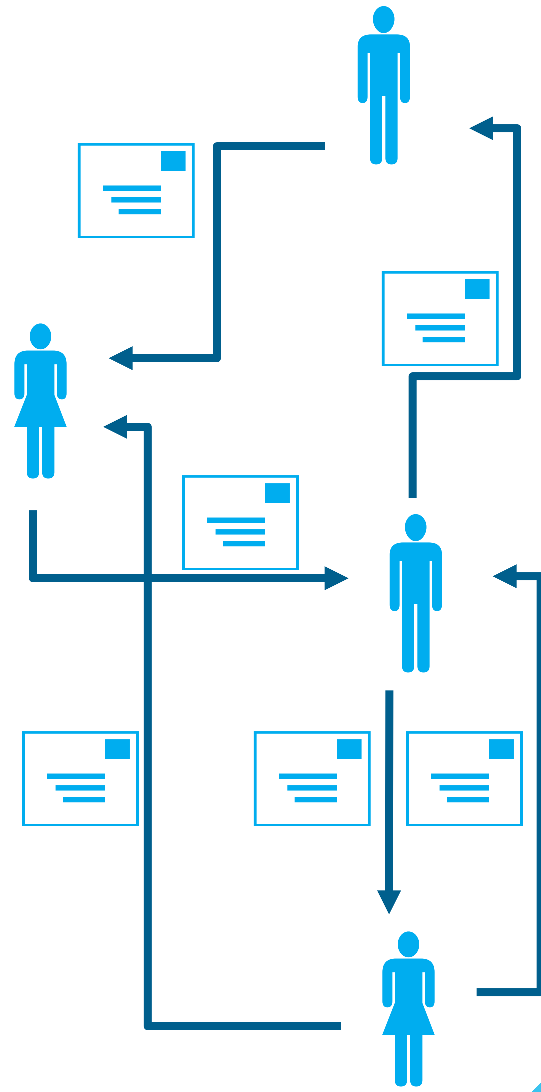
Entity Resolution

- References, names, friendships
- Use rules to express evidence
  - ‘’If two people have similar names, they are probably the same’’
  - ‘’If two people have similar friends, they are probably the same’’
  - ‘’If A=B and B=C, then A and C must also denote the same person’’



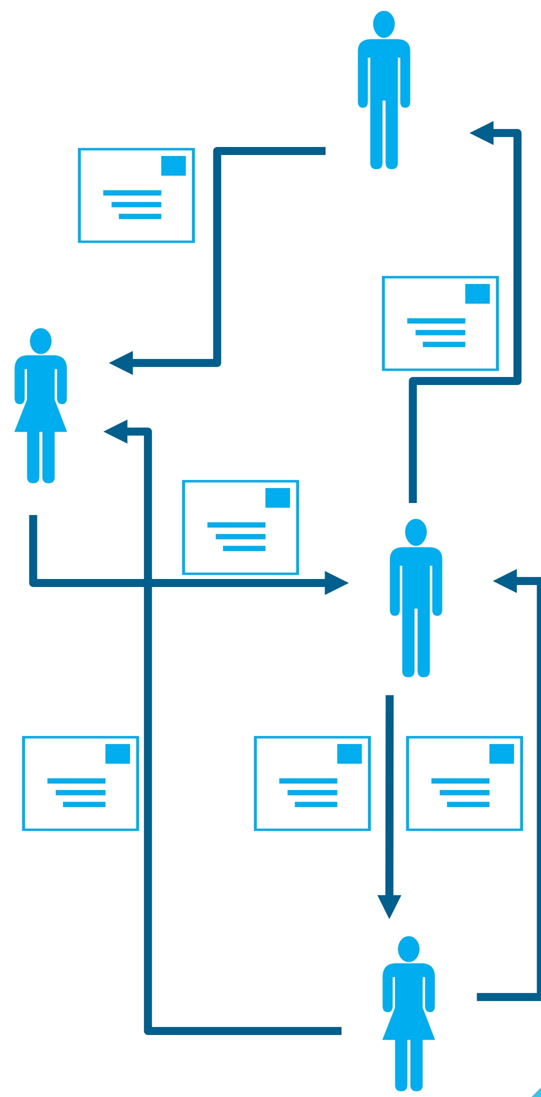
Link Prediction

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



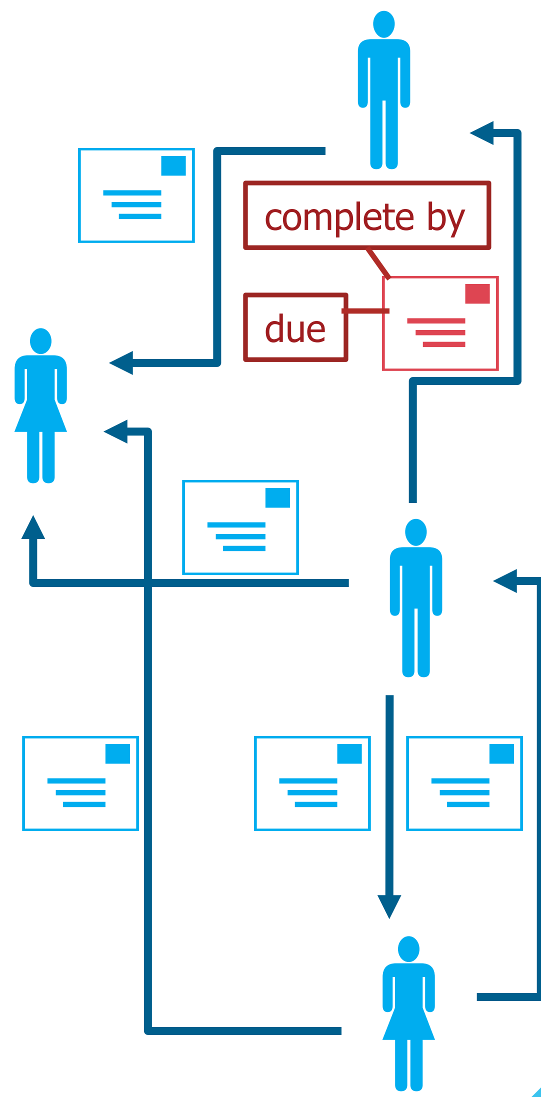
Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If email content suggests role X, person 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"



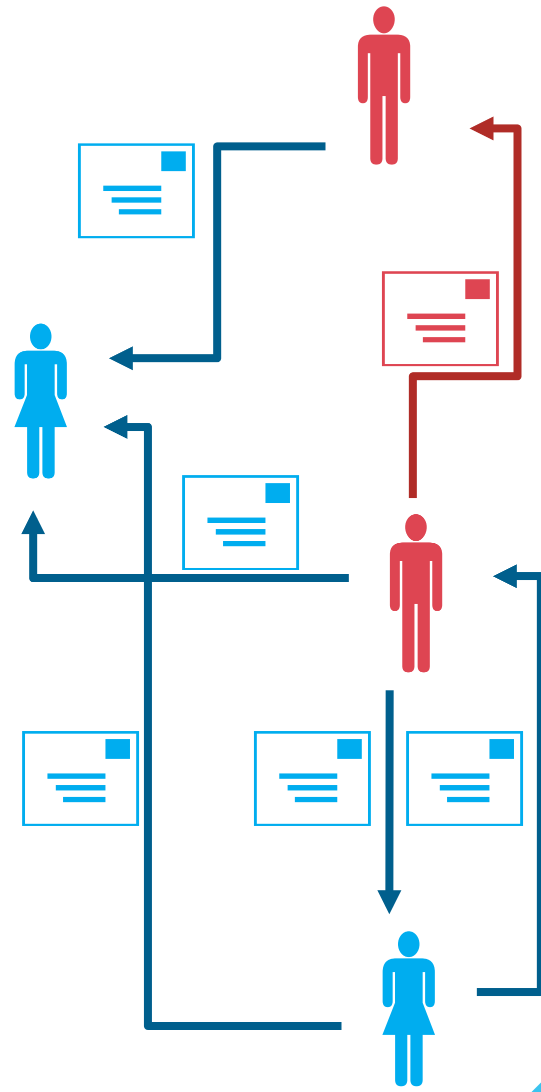
Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If email content suggests 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"



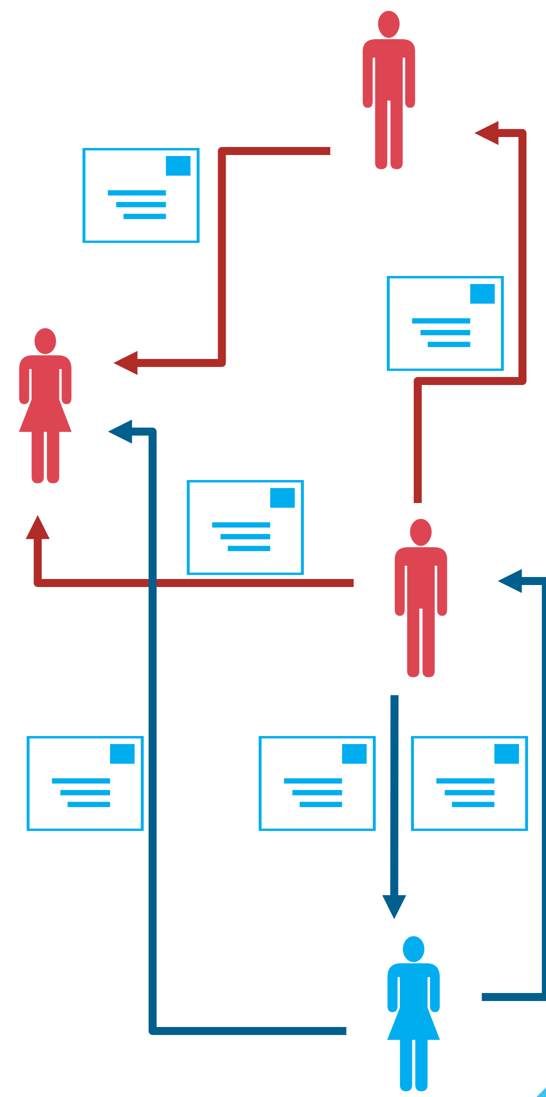
Link Prediction

- People, emails, words, communication, relations
- Use rules to express evidence
  - "If email content suggests 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"

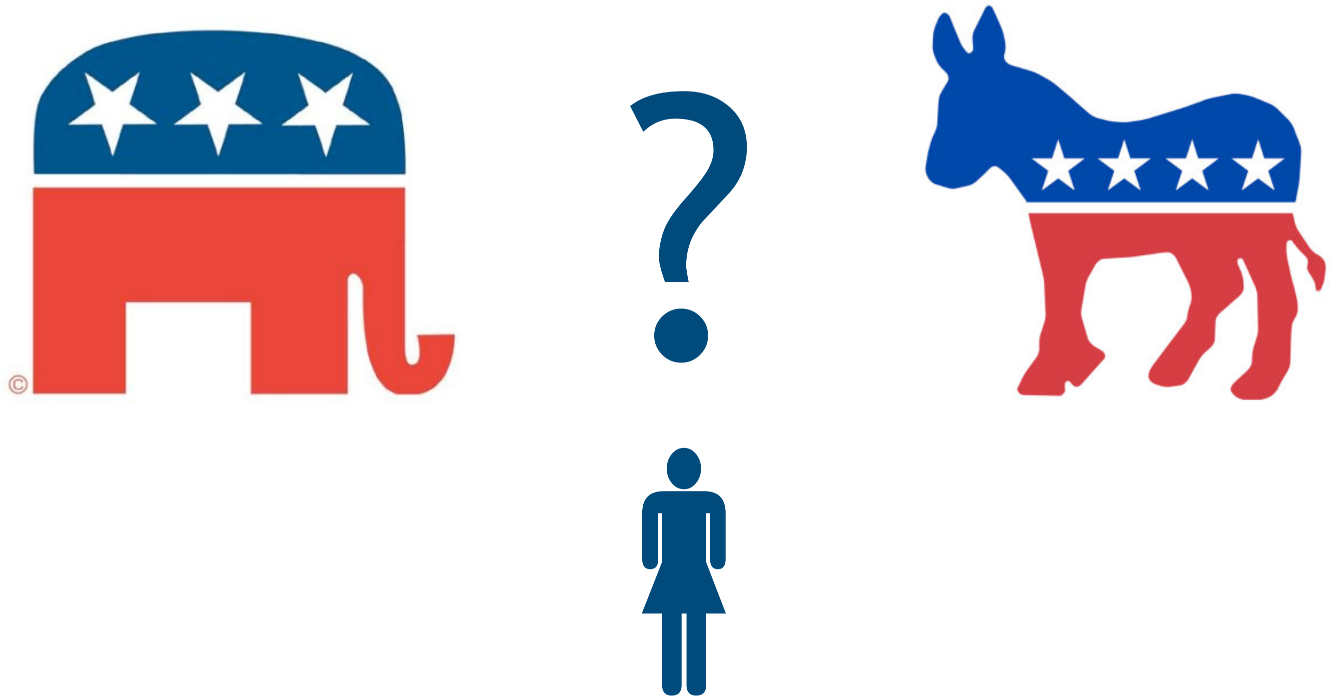


Link Prediction

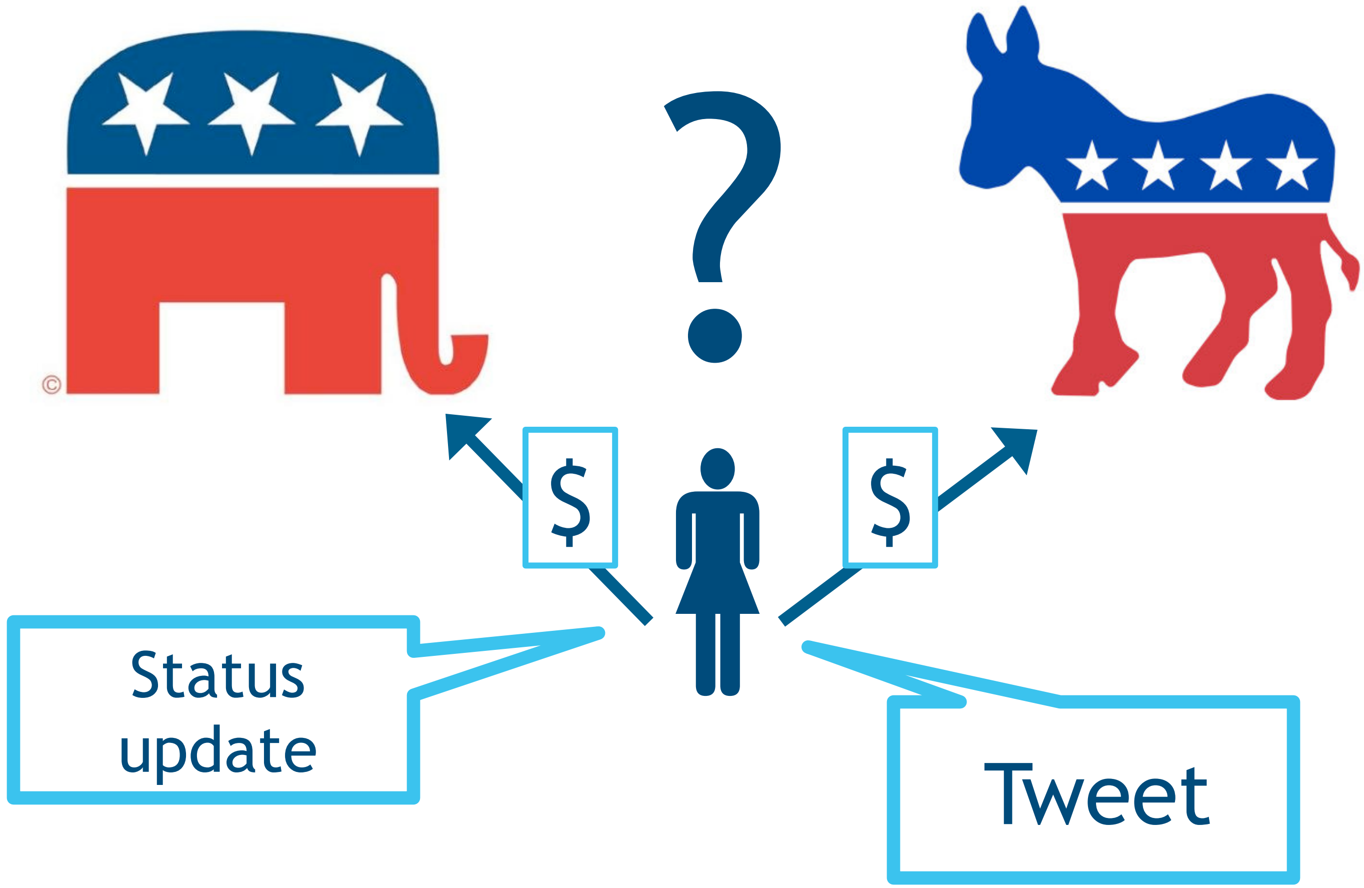
- People, emails, words, communication, relations
- Use rules to express evidence
  - "If email content suggests type X, it is of type X"
  - "If A sends deadline emails to B, then A is the supervisor of B"
  - "If A is the supervisor of B, and A is the supervisor of C, then B and C are colleagues"



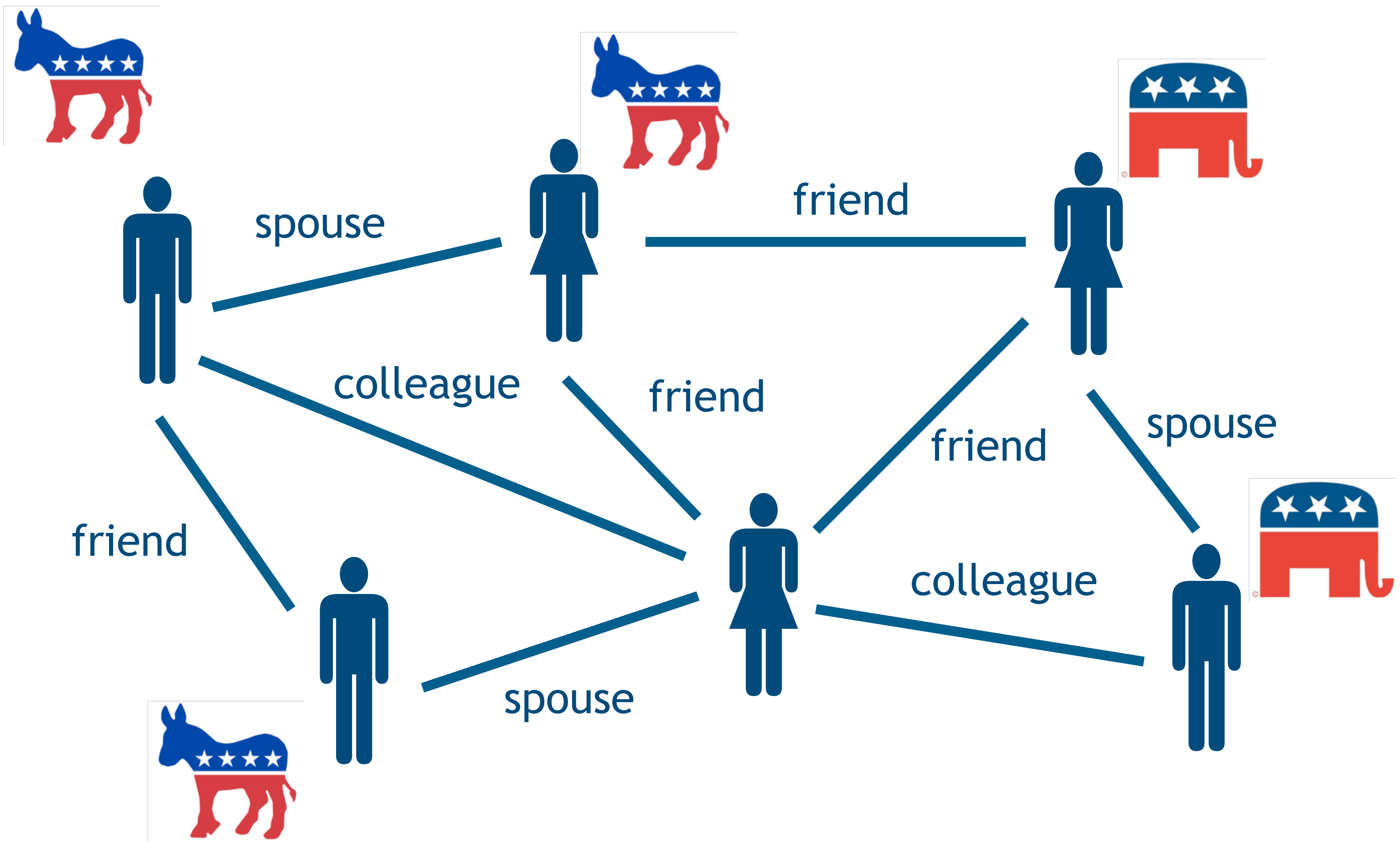
Node Labeling



Voter Opinion Modeling

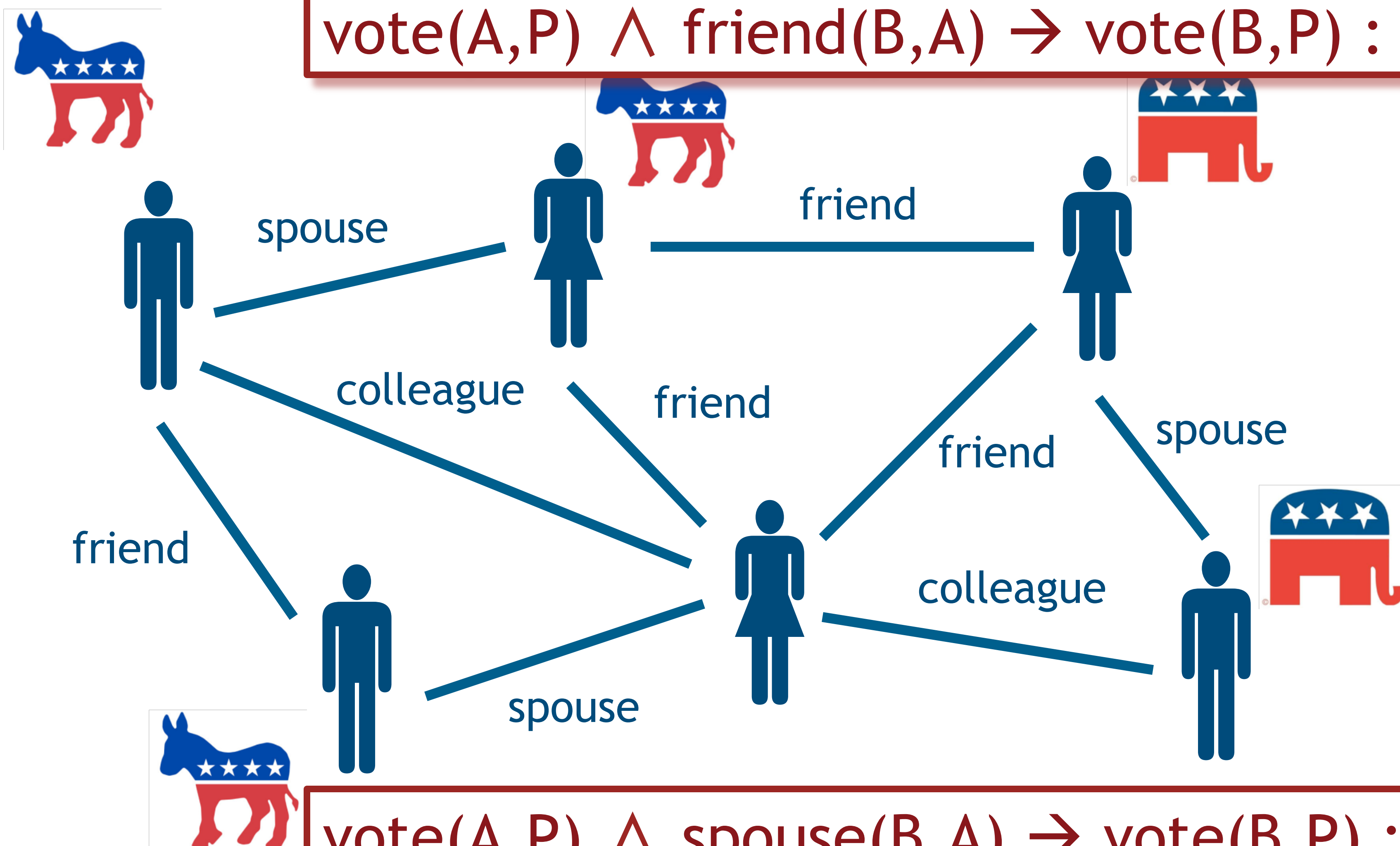


Voter Opinion Modeling



Voter Opinion Modeling

vote(A,P) ∧ friend(B,A) → vote(B,P) : 0.3



vote(A,P) ∧ spouse(B,A) → vote(B,P) : 0.8

Mathematical Foundation

Rules

$$H_1 \vee \ldots H_m \leftarrow B_1 \wedge B_2 \wedge \ldots B_n$$

- Atoms are real valued, [0,1]
- Combination functions, Lukasiewicz T-norm
  - $a_1 \vee a_2 = \min(1, a_1 + a_2)$
  - $a_1 \wedge a_2 = \max(0, a_1 + a_2 - 1)$
- Distance to Satisfaction
  - $h_1 \leftarrow b_1 \wedge b_2$

$$R \approx T \leftarrow A \approx B{:}0.7 \wedge D \approx E{:}0.8$$

Rules

$$H_1 \vee \ldots H_m \leftarrow B_1 \wedge B_2 \wedge \ldots B_n$$

- Atoms are real valued, [0,1]
- Combination functions, Lukasiewicz T-norm
  - $a_1 \vee a_2 = \min(1, a_1+a_2)$
  - $a_1 \wedge a_2 = \max(0, a_1 + a_2 - 1)$
- Distance to Satisfaction
  - $h_1 \leftarrow b_1 \wedge b_2$

$$R{\approx}T{:}\geq 0.5 \leftarrow A{\approx}B{:}0.7 \wedge D{\approx}E{:}0.8$$

Rules

$$H_1 \vee \ldots H_m \leftarrow B_1 \wedge B_2 \wedge \ldots B_n$$

- Atoms are real valued, [0,1]
- Combination functions, Lukasiewicz T-norm
  - $a_1 \vee a_2 = \min(1, a_1 + a_2)$
  - $a_1 \wedge a_2 = \max(0, a_1 + a_2 - 1)$
- Distance to Satisfaction
  - $h_1 \leftarrow b_1 \wedge b_2$

$R{\approx}T{:}0.7 \leftarrow A{\approx}B{:}0.7 \wedge D{\approx}E{:}0.8$	0.0
$R{\approx}T{:}0.2 \leftarrow A{\approx}B{:}0.7 \wedge D{\approx}E{:}0.8$	0.3

Probabilistic Model

Rule's weight

Rule's distance to satisfaction

$$d_r(I) = \max(0, I_{r,\text{body}} - I_{r,\text{head}})$$

Probability density over interpretation I

$$f(I) = \frac{1}{Z}\exp[-\sum_{r \in R} \lambda_r (d_r(I))^p]$$

Distance exponent in {1, 2}

Normalization constant

Set of ground rules

Constrained Continuous Markov Random Field (CCMRF)

PSL Inference

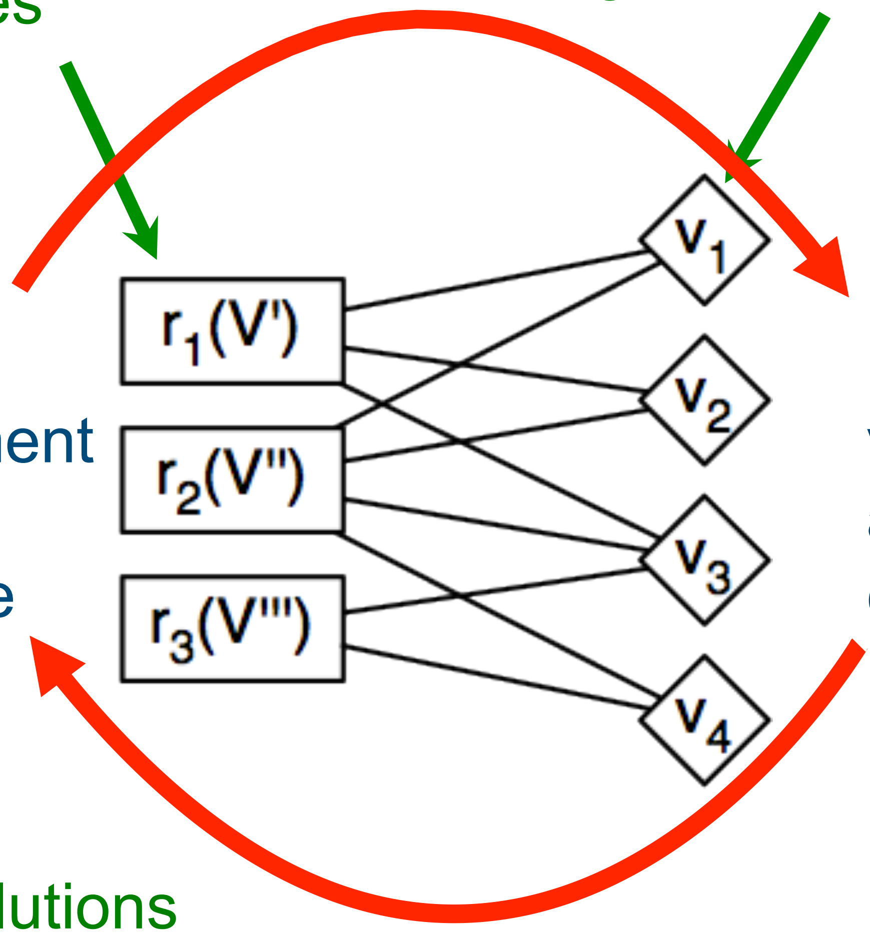
- CCMRF translates to a conic program in which:
  - MAP inference is tractable ($O(n^{3.5})$) using off-the-shelf interior point methods (IPM) optimization packages [Broecheler et al. UAI 2010]
  - Margin inference is based on sampling algorithms adapted from computational geometry methods for volume computation in high dimensional polytopes [Broecheler & Getoor, NIPS 2010]
- While a naïve approach is tractable, it still suffers from problems of scalability
  - IPMs operate on matrices. These matrices become large and dense when many variables are all interdependent, such as is common in alignment problems.
  - Scaling to large data requires an alternative to forming and operating on such matrices

Consensus Optimization

[Bach et al, NIPS 12]

rules with local copies of random variables

original random variables

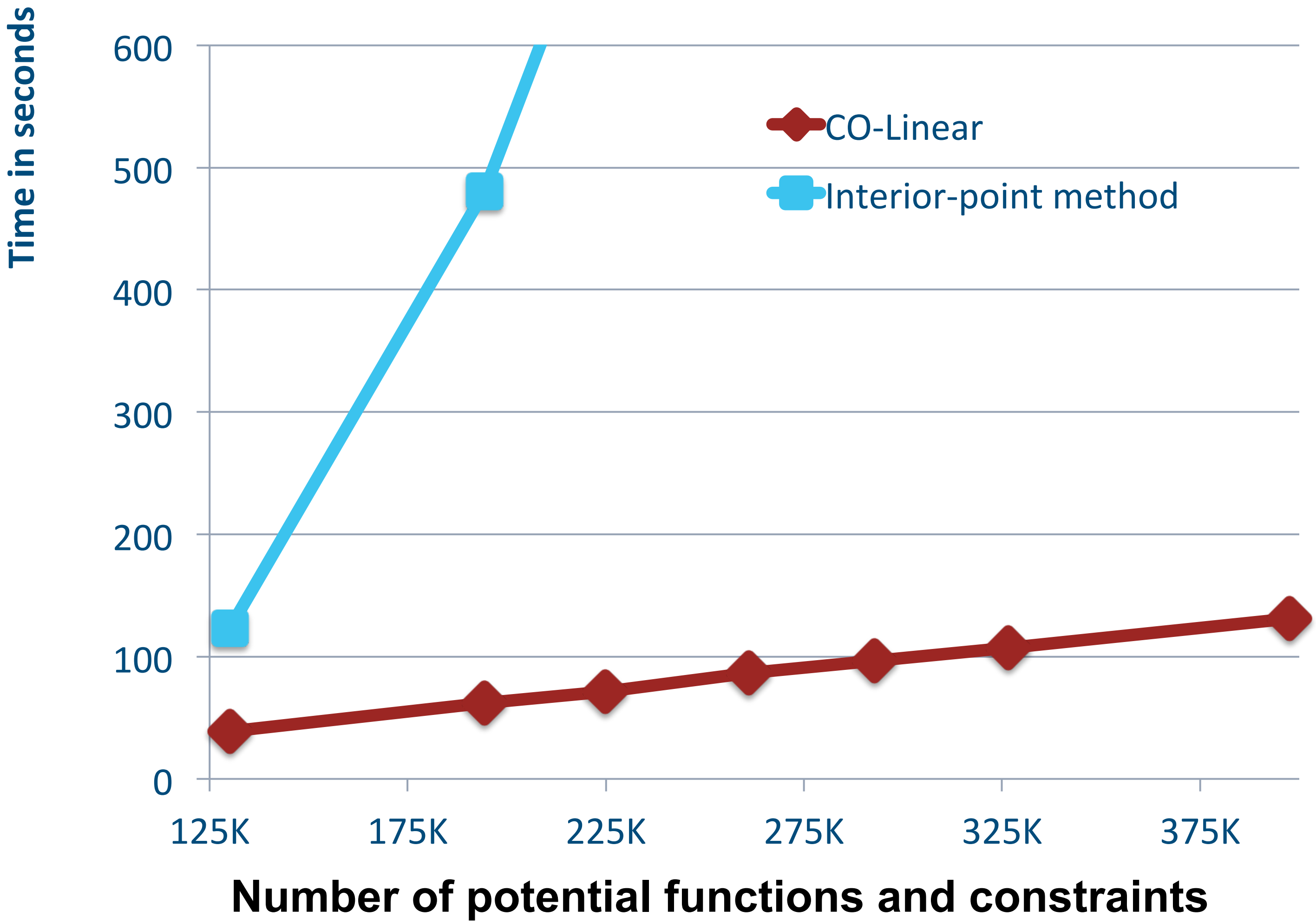


optimize truth values & agreement with original variables per rule

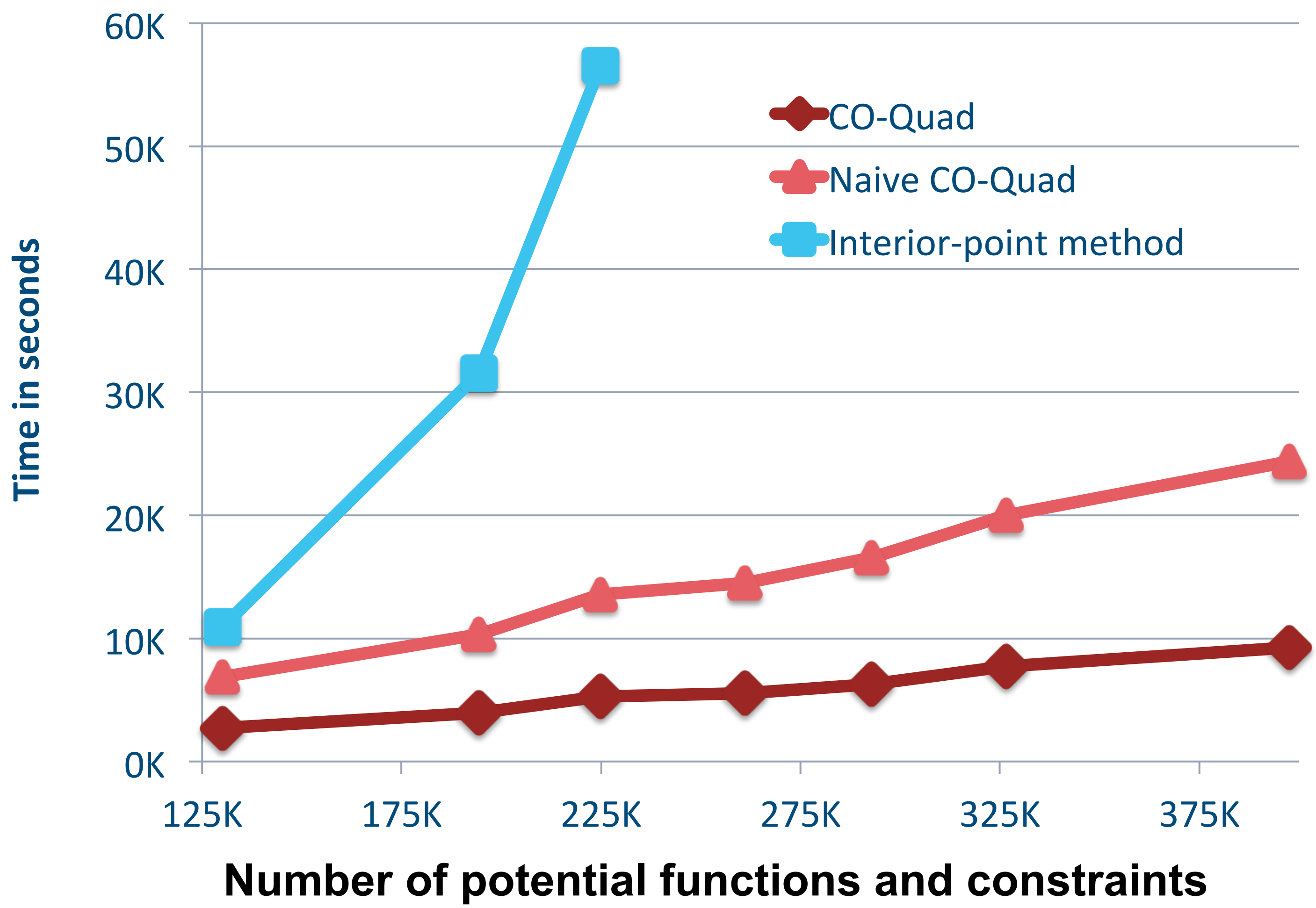
update variables to average of copies

key: fast solutions

Linear Constraints



Quadratic Constraints

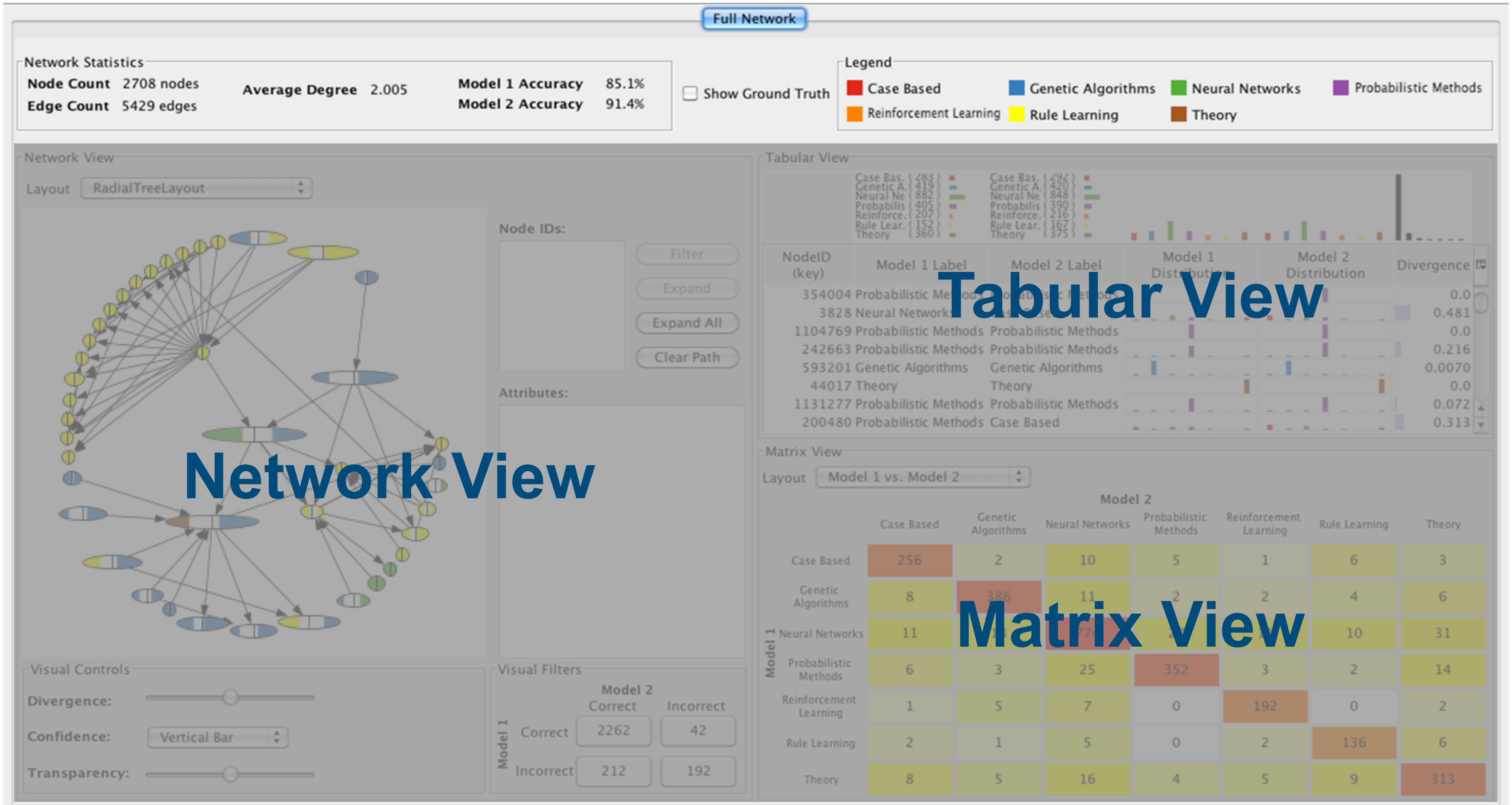


Comparative Visual Analytics

G-Pare

- A visual analytic tool that:
  - Supports the comparison of uncertain graphs
  - Integrates three coordinated views that enable users to visualize the output at different abstraction levels
  - Incorporates an adaptive exploration framework for identifying the models' commonalities and differences

G-Pare



Full Network
Network Statistics
Node Count 2708 nodes
Edge Count 5429 edges
Average Degree 2.005
Model 1 Accuracy 85.1%
Model 2 Accuracy 91.4%
Show Ground Truth
Legend
Case Based
Genetic Algorithms
Neural Networks
Probabilistic Methods
Reinforcement Learning
Rule Learning
Theory
Network View
Tabular View
Matrix View

Node Visualization

- Model 1 prediction: “Neural Networks”
  Model 2 prediction: “Theory”
- Model 1 is more confident in its prediction than Model 2
- Distributions of the two models vary significantly
- Model 1’s prediction matches the ground truth

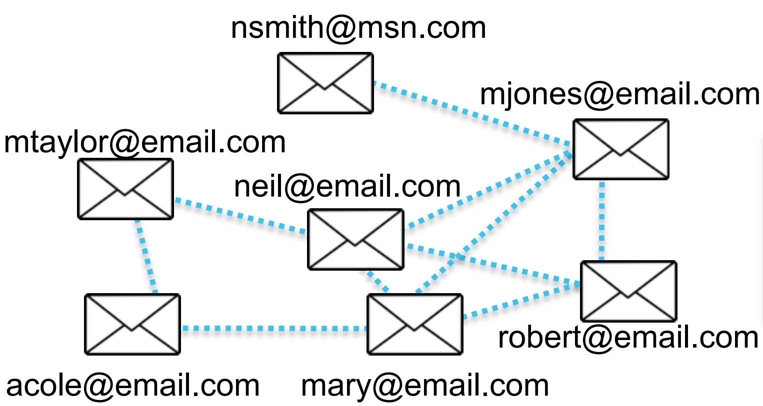
Summary

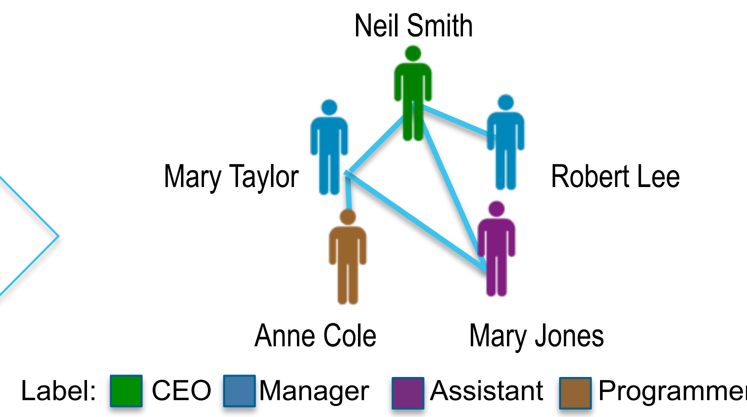
- Uncertain Graphs: Foundations
  - Probabilistic Soft Logic (PSL)
  - **http://psl.umiacs.umd.edu/**
- Visual Analytics for Model Comparison
  - G-Pare
  - **http://www.cs.umd.edu/projects/linqs/gpare**
- Key supporting publications: VAST 2009, UAI 2010, NIPS 2010, NIPS WS 2010, VAST 2011, VDA 2011, NIPS 2012, PAKDD 2012, ISWC WS 2012, UAI WS 2012, 3 NIPS WS 2012

Impact: Graph Identification

- Analytic Goal:
  - Given a partially observed **input graph** infer a distribution over **output graphs**
- Major components:
  - **Entity Resolution (ER):** Infer the set of nodes
  - **Link Prediction (LP):** Infer the set of edges
  - **Collective Classification (CC):** Infer the node labels

e.g., Communication -> Social Network





<u>Communication Network</u>
Nodes: Email Address
Edges: Communication
Node Attributes: Words

<u>Organizational Network</u>
Nodes: Person
Edges: Manages
Node Labels: Title

Extensions and Outreach

- Funding
  - Maryland Industrial Partners w/ Optimal Solutions ($130K), OSI IARPA sub to Vtech ($2M), NSF III Small ($500K)
- 20+ Invited Talks
  - CMU, NYU, Notre Dame, Minnesota, Rutgers, UCI, CRA-W, Microsoft Research, Google, Sante Fe Institute, IMA, DIMACS/CCICADA, NEH/IPAM, etc.
  - Invited Talk NIPS WS on Challenges in Data Visualization
- 9 Tutorials & 2 Workshops
  - NIPS 2012, VLDB 2012, AAAI 2012, ASONAM 2012, VizWeek 2012, WSDM 2011, SDM 2011, SIGMOD 2011, IEEE Visualization 2011 and SRL/ISSDM Research Symposium 2011, AAAI 2010
- Incorporated Visual Analytics into 3 courses
- Grant has supported 5 PhD students, 2 Master's students, 4 undergraduates

Thanks!

Questions?

Comments?

Come to posters!

References

References

[1] *Computing marginal distributions over continuous Markov networks for statistical relational learning*, Matthias Broecheler, and Lise Getoor, Advances in Neural Information Processing Systems (NIPS) 2010

[2] *A Scalable Framework for Modeling Competitive Diffusion in Social Networks*, Matthias Broecheler, Paulo Shakarian, and V.S. Subrahmanian, International Conference on Social Computing (SocialCom) 2010, Symposium Section

[3] *Probabilistic Similarity Logic*, Matthias Broecheler, Lilyana Mihalkova and Lise Getoor, Conference on Uncertainty in Artificial Intelligence 2010

[4] *Decision-Driven Models with Probabilistic Soft Logic*, Stephen H. Bach, Matthias Broecheler, Stanley Kok, Lise Getoor, NIPS Workshop on Predictive Models in Personalized Medicine 2010

[5] *Probabilistic Similarity Logic*, Matthias Broecheler, and Lise Getoor, International Workshop on Statistical Relational Learning 2009

[6] *G-PARE: A Visual Analytic Tool for Comparative Analysis of Uncertain Graphs* Hossam Sharara, Awalin Sopan, Galileo Namata, Lise Getoor, Lisa Singh IEEE Conference on Visual Analytics Science and Technology, 2011 (VAST '11).

Probabilistic
Soft
Logic

psl.umiacs.umd.edu