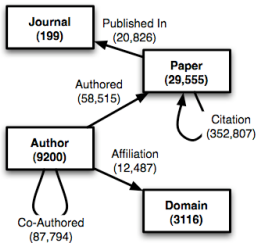
Relational Learning

David Jensen

Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst

Understanding physics publishing

- Physics Preprint Archive (www.arxiv.org)
  30,000 papers in high-energy physics theory (HEP-Th)
- Questions
  - What are the major areas of research in this field?
  - Whose work is most central?
  - Whose work is currently under-rewarded?
  - Does the 80/20 rule hold in this field?
  - What factors determine whether a paper will be published?

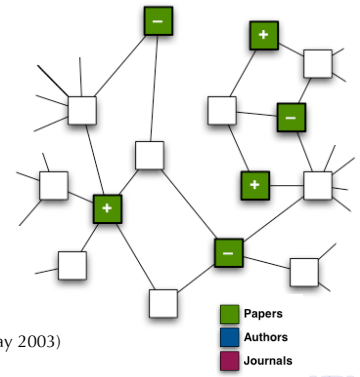


KDD Cup 2003 competition

- **What is it?** — Most widely recognized competitive evaluation of the technology and practices of knowledge discovery
- **Who competed?** — 57 teams from universities and companies in 10 countries competing on four tasks
- **'Open Task'** — Define and answer questions about the physics literature based on the HEP-Th data
- **Evaluation** — Questions and answers judged by a panel of experts
- **Result** — First place
  (McGovern, Friedland, Hay, Gallagher, Fast, Neville & Jensen 2003)

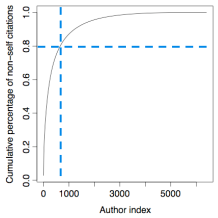
Example analyses of HEP-Th papers

- Consolidated authors with graph queries
  (Blau, Immerman, & Jensen 2001)
- Identified topics with spectral clustering
  (Neville, Adler, & Jensen 2003)
- Identified authoritative authors with graph calculations
- Built statistical models of publication success
  (Neville, Jensen, Friedland & Hay 2003)



Example results from HEP-Th papers

- Citations among papers in HEP-Th defines cohesive paper topics (e.g., tachyon condensation)
- Edward Witten is the most influential figure in theoretical high-energy physics today.
- Two physicists (I. Klebanov and A. Strominger) may be due for awards soon
- An '80/20 rule' applies
- Single-author papers are much less likely to be published in journals



KDL

What is knowledge discovery?

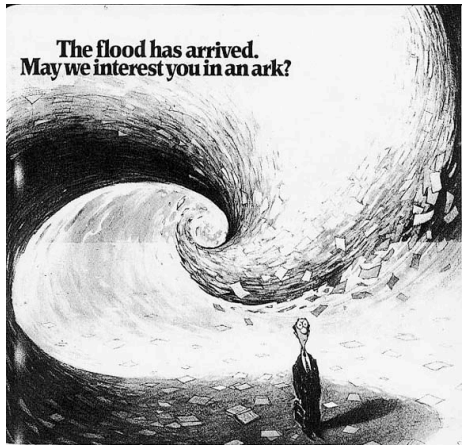
- "Computational tools for extracting previously unknown and potentially useful information from large sets of data."
- *Software for 'sensemaking'* — Computational tools that help people bring meaning to the huge volumes of data that flood the modern world. (Waldrop 2003)
- Draws on work in statistics, artificial intelligence, databases, psychology, and philosophy of science (and social network analysis and graph theory)

KDL

Why is knowledge discovery important?

- Critical tasks in business, science, and government already require 'sensemaking' from large and complex databases
  - Stock analysis and fraud detection
  - Citation analysis
  - Intelligence analysis and government oversight
- ...soon we all may need sensemaking help
  - *Web search* returns thousands of documents
  - *Citation databases* access vast citation networks
- Often want *understanding*, not just predictions

KDL



KDL

Complementary areas of research

- Searching and retrieving useful data
  - "Information retrieval" or "Database querying"
  - KD helps us understand the deep structure of the Web
- Extracting structured data from text and other sources
  - "Information extraction" and "image understanding"
  - KD can use extracted data
- Merging many smaller databases into a large one
  - "Database integration" or "Data fusion"
  - KD constructs models from large and small databases
- Autonomous model building
  - "Agent learning" or "Robot learning"
  - KD focuses on complementing human abilities

KDL

Taking 'sensemaking' seriously

- People are…
  …rich in knowledge about the world
  …poor at probabilistic learning and reasoning
- Tools are…
  …poor in knowledge about the world
  …rich in probabilistic learning and reasoning
- One recipe for knowledge discovery
  - Leverage human knowledge of the world
  - Provide computational support for statistical learning and reasoning
- *Are we there yet?*

KDL

The "big ideas" of relational learning

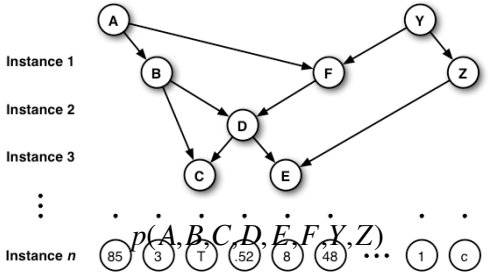
- Joint models of attributes in relational data
  - "PRMs" or relational Bayesian networks (RBNs) (Getoor, Friedman, Koller & Pfeffer 2001)
  - Relational Markov Networks (RMNs) (Taskar et al. 2002)
  - Relational dependency networks (RDNs) (Neville & Jensen 2003, 2004)
- Statistical biases in relational learning
  - Autocorrelation & feature selection (Jensen & Neville 2002)
  - Aggregation & feature selection (Jensen, Neville, & Hay 2003)
- Collective inference
  - Hypertext classification (e.g., Chakrabarti, Dom & Indyk 1998)
  - General relational data (e.g., Neville & Jensen 2000; Taskar, Segal & Koller 2001; Jensen, Rattigan, & Blau 2003; Jensen, Neville & Gallagher 2004)

KDL

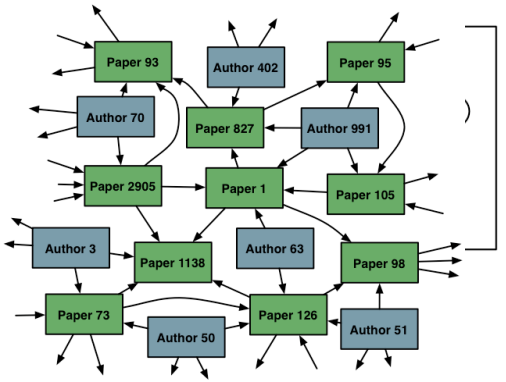
Joint Models of Attributes in Relational Data

KDL

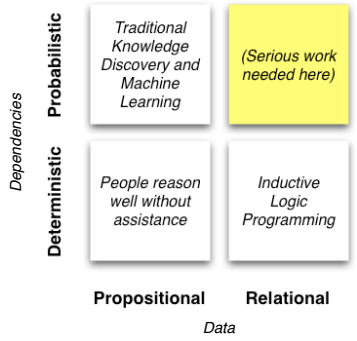
Propositional models



Assumptions of propositional models



Problem space for knowledge discovery

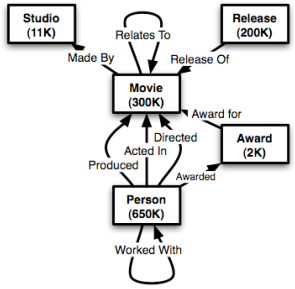


Growing research interest

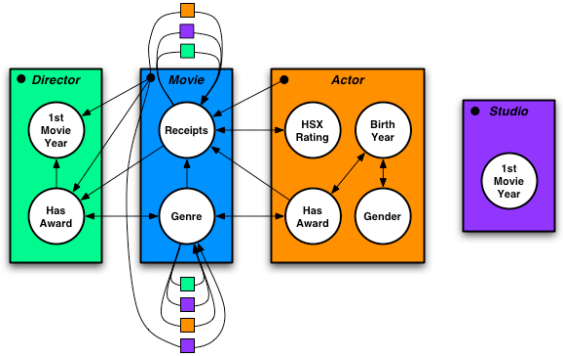
- New work in knowledge discovery & machine learning
  - "Relational revolution" (Dietterich 2003)
  - Growing frequency of specialized workshops — AI&LA 1998, SRL 2000, MRDM 2001, SRL 2003, MRDM 2003, SRL 2004
  - Major topic area for technical conferences (ICML, KDD)
  - Focus area for two DARPA programs (2001-2003; 2004-)
- Investigation of emergent properties of networks
  - Condensed matter physics and social network analysis
  - "New science of networks" (Watts 2003)
- Growing interactions
  - Example: Domingos & Richardson 2001 (best paper KDD 2001); D. Kempe, J. Kleinberg, & E. Tardos (best paper KDD 2003)

Internet Movie Database (IMDb)

- The Internet Movie Database (www.imdb.com)
- Questions
  - What predicts box office receipts?
  - Are awards important?
  - What about previous commercial success?
  - Do ticket buyers care about studios, or only about actors and directors?

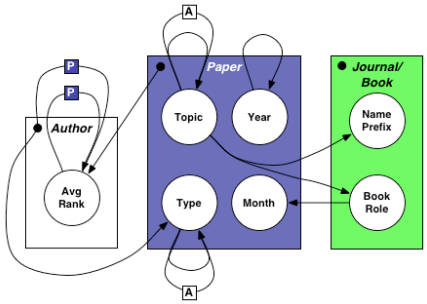


RDN for IMDb

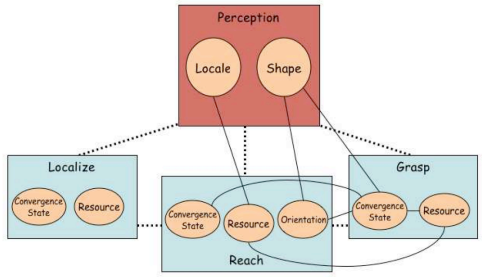


(Neville & Jensen 2003, 2004; builds on Heckerman et al. 2000)

RDN for Cora



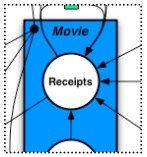
RDN for Robotic Localize-Reach-Grasp



(Hart, Grupen, & Jensen 2004)

The conditional models inside RDNs

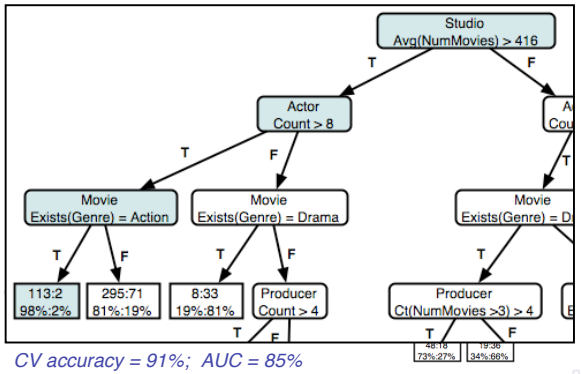
- RDNs are learned by composing a set of conditional models $p(y|Y^{-},X,S)$
- For example, $p(Receipts)$ given
  - Receipts of related movies
  - Movie genre
  - Ratings of the actors in the movie
  - …



- To obtain advantages, the conditional models must be accurate, valid, and parsimonious

KDL

Relational probability trees (RPT)



CV accuracy = 91%; AUC = 85%

KDL

RDN Strengths

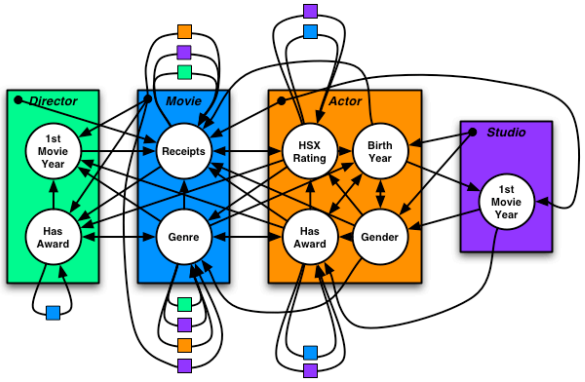
	RDNs	DTs, etc.	ILP	RBCs (KDD'03)	PRMs (Friedman et al '99)	RMNs (Taskar et al '02)
Relational	√	X	√	√	√	√
Probabilistic	√	√	X	√	√	√
Collective Inference	√	X	X	X	√	√
Autocorrelation	√	X	√	X	X	X
Efficient learning	√	√	X	√	√	X

KDL

Statistical Biases in Relational Learning

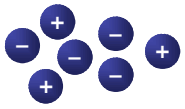
KDL

Pathological learning

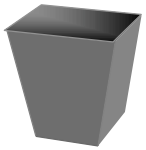


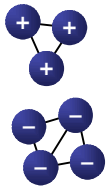
KDL

Independence assumption



Nearly all techniques assume data instances are independent random samples



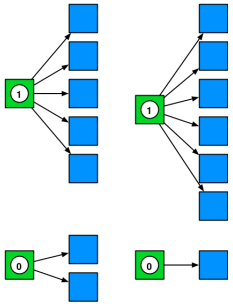


In reality, data instances in many relational data sets are interconnected and *autocorrelated*

KDL

Some causes for pathological learning

- Autocorrelation
  The value of a variable on one object depends on the values of the same variable on related objects
  $p(y) \neq p(y|Y\text{-})$
- Structural dependence
  The attributes and the structure of data are correlated
  $p(y) \neq p(y|S)$

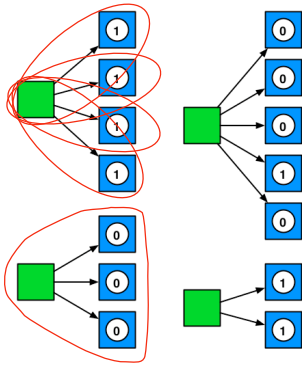


(Jensen & Neville 2002; Jensen, Neville & Hay 2003)

KDL

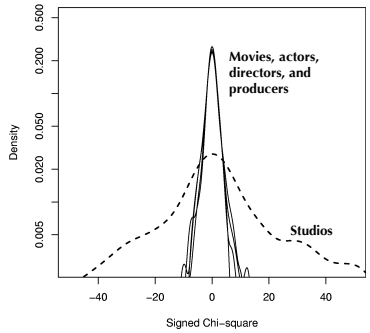
Autocorrelation and effective sample size

- The confidence of any statistical association varies with sample size (N)
- Consider evaluating the association between characteristics of groups and their members
- What is the "effective" sample size?
  - N = |members|
  - N =|groups|
  - |members| ≥ N ≥ |groups|

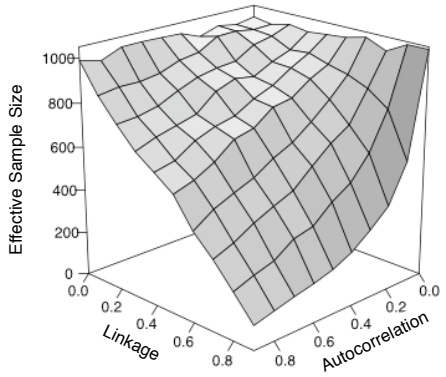


KDL

Differing variance of feature scores

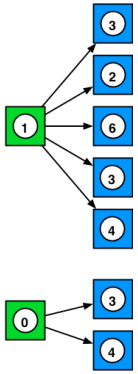


Autocorrelation and effective sample size



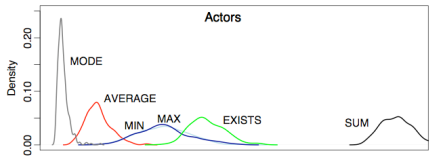
Structural dependence

- Of the few existing techniques for relational learning, nearly all use aggregation functions (e.g., MAX)
- Degree disparity can cause nearly any aggregated variable to produce apparent correlation
- For MAX and SUM,
  E(X|high-degree) > E(X|low-degree)
- For AVE and MODE,
  Var(X|high-degree) < Var(X|low-degree)



Effects of structural dependence

- Thus, measures of relational correlation are drawn from different sampling distributions that depend on graph structure
- This can biases selection toward features with the *least* statistical evidence

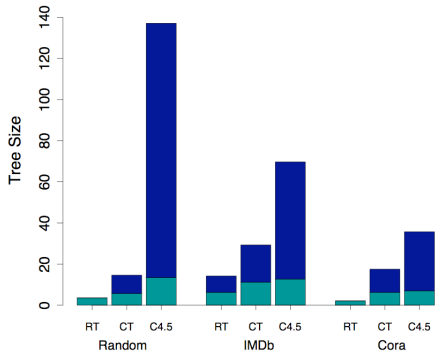


(Jensen, Neville & Hay 2002)

Adjusting for relational dependence

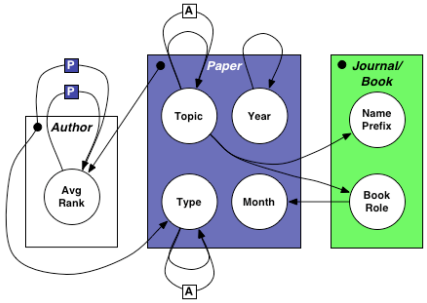
- Randomization tests — Produce empirical sampling distributions by randomizing key elements of the data
- Sample size corrections — Estimate effective sample size using observed autocorrelation and graph structure.
- Conditional hypothesis tests — Use conventional hypothesis tests that explicitly account for correlation between attributes and structure.

Corrections produce smaller models

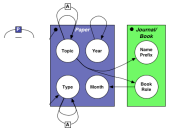


Collective Inference

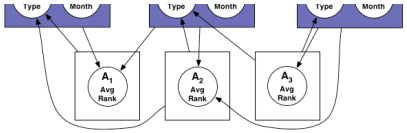
RDN Model for Cora



Collective inference for RDNs



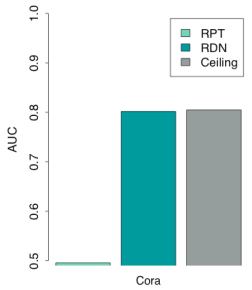
- Use Gibbs sampling to estimate a full joint distribution and calculate probabilities of interest (e.g., marginals)
- Complexity ≈ *objects* • *attributes* • *log(dependencies)*



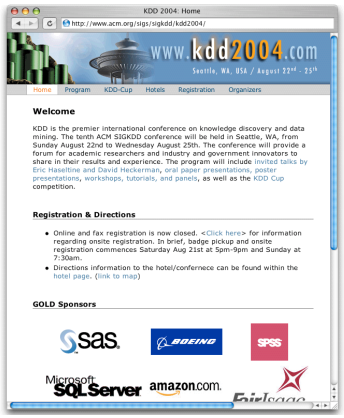
KDL

Model Performance

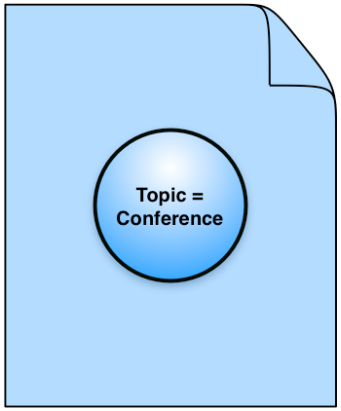
- Comparison #1:
  - RDN vs. RPT learned without class labels of related entities
  - *Collective classification significant improvement over individual classification*
- Comparison #2:
  - RDN vs. RDNs applied with true class labels of related entities (*Ceiling*)
  - *Joint inference with Gibbs sampling is effective*



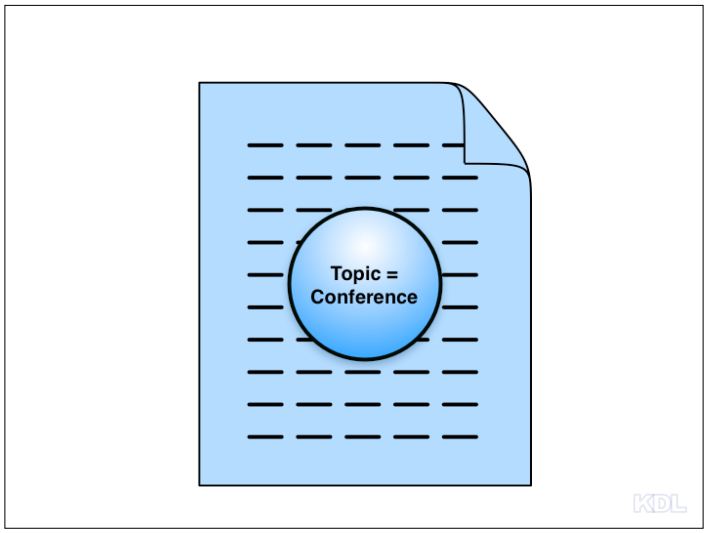
KDL

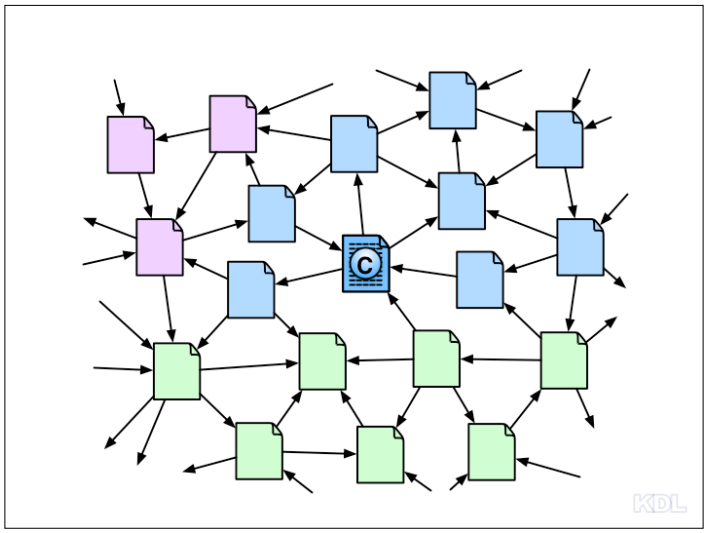


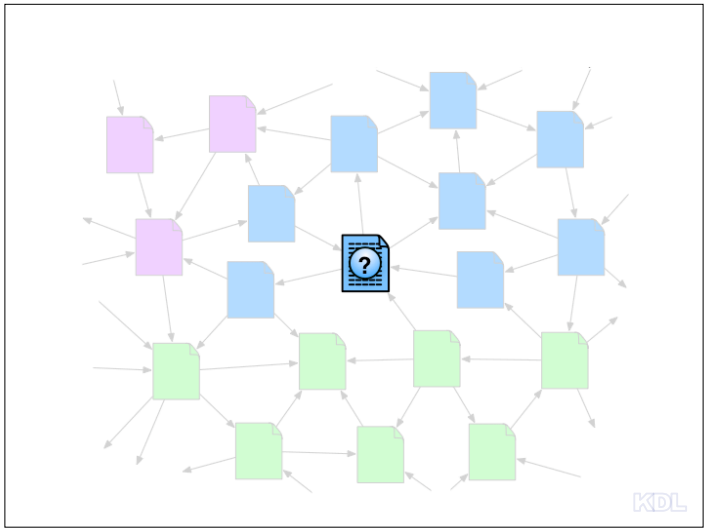
KDL

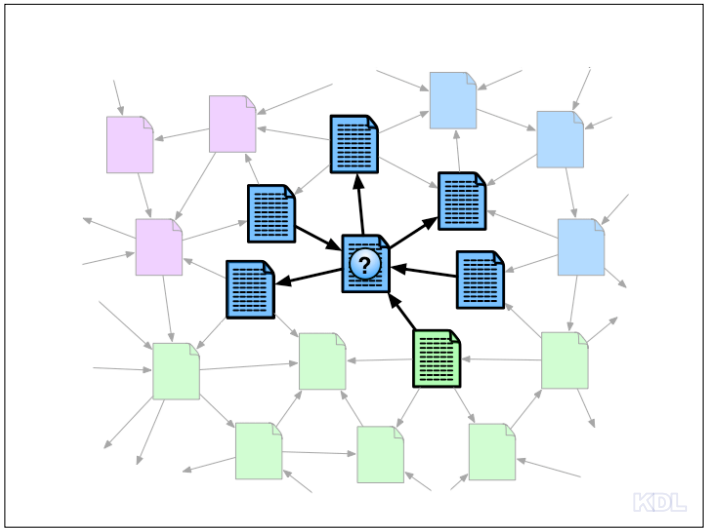


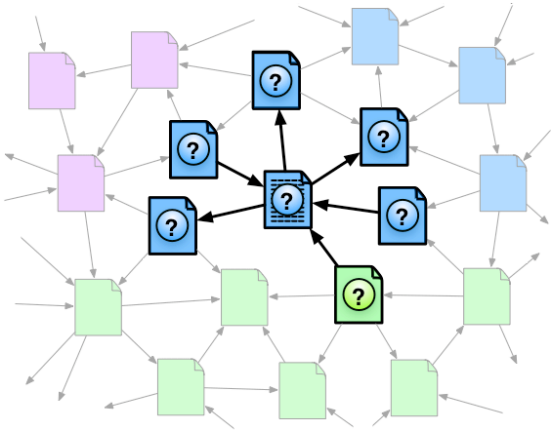
KDL


Topic =
Conference
KDL

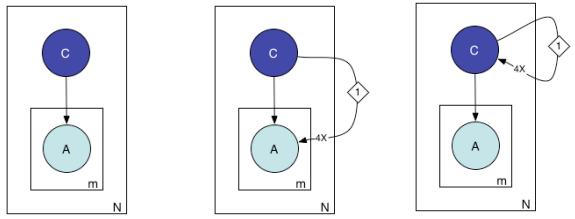

C
KDL


?
KDL


?
KDL



Models

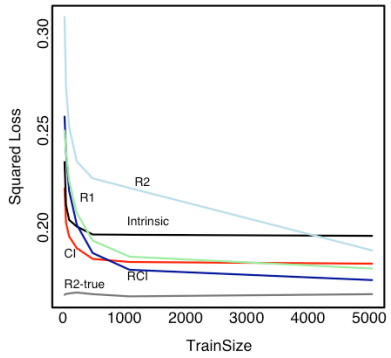


Intrinsic Relational 1(R1) Collective (CI)

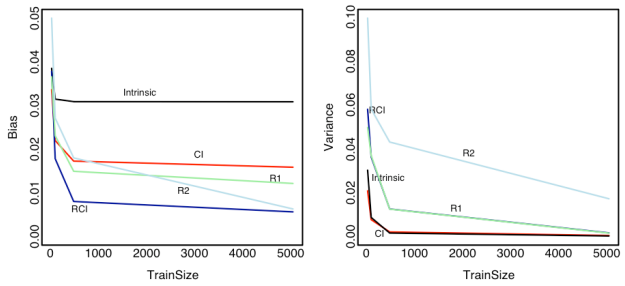
Collective inference

- Joint models of relational data can exploit *collective inference*, in which inferences about all variables in a data set are made jointly (Chakrabarti et al. 1998; Taskar et al. 2001)
- The influence of highly confident inferences can travel substantial distances in the graph
- Collective inference exploits a clever factoring of the space of dependencies to reduce variance, thus improving performance over considering all relational attributes (Jensen, Neville & Gallagher 2004).

CI has lowest small-sample error

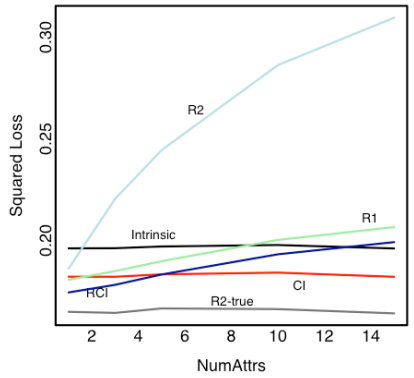


CI reduces bias with minimal variance



KDL

Increasing number of attributes



KDL

Open Topics

KDL

Open research topics

- Learning models that infer the existence of objects, links, and groups
- Representing, learning, and reasoning with temporal and spatial knowledge
- Active learning
- Learning causal dependencies in relational data
- Incremental learning and reasoning
- Connecting learning and simulation
- Diagnosis and repair of compositional models

KDL

Open source software

- Nearly all techniques developed in KDL are implemented within PROXIMITY, our environment for relational knowledge discovery
- Implementation
  - 30,000+ lines of Java,
  - Built on Monet, an open-source database by CWI
  - Runs on all major platforms
  - 80-page tutorial and additional documentation
- Open-source release of v.3 on 15 April 2004; Released 3.1 in September.

KDL

Thanks to…

Jennifer Neville
Brian Gallagher

Michael Hay
Amy McGovern
Matthew Rattigan
Özgür Şimşek
Pippin Wolfe

Hannah Blau
Dan Corkill
Matthew Cornell
Ross Fairgrieve
Andrew Fast
Lisa Friedland
Cindy Loiselle
Agustin Schapira

KDL

Further information

jensen@cs.umass.edu
kdl.cs.umass.edu

KDL