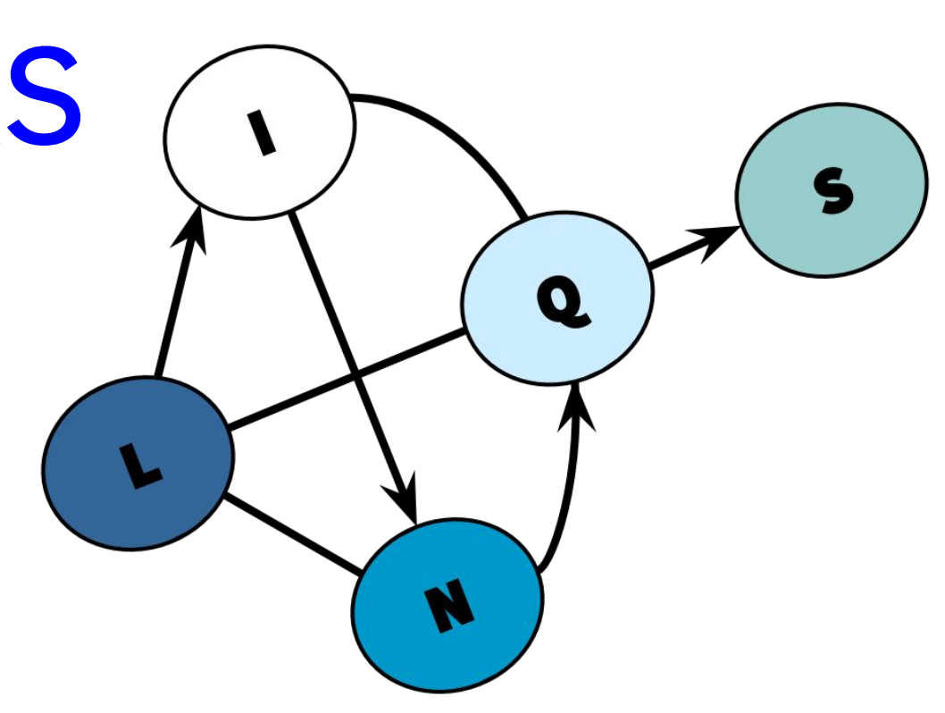
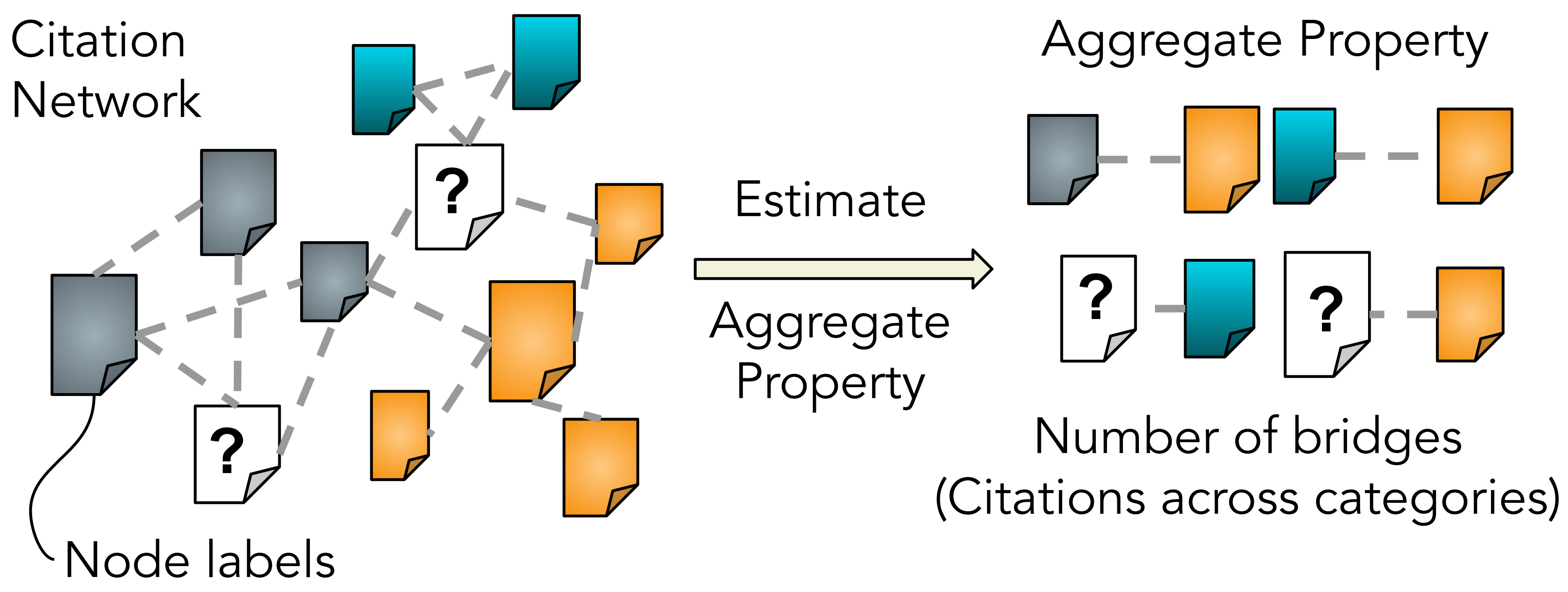
Estimating Aggregate Properties In Relational Networks With Unobserved Data



Varun Embar *, Sriram Srinivasan *, and Lise Getoor
University of California, Santa Cruz
*Equal Contribution

Goal



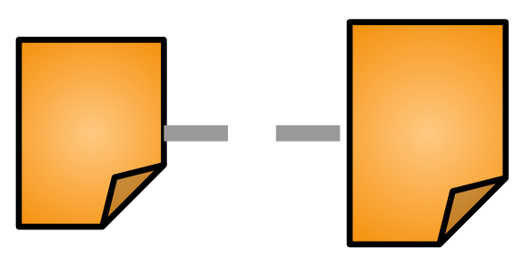
Challenge

- Estimating aggregate properties when network is not fully observed (E.g. missing node labels)

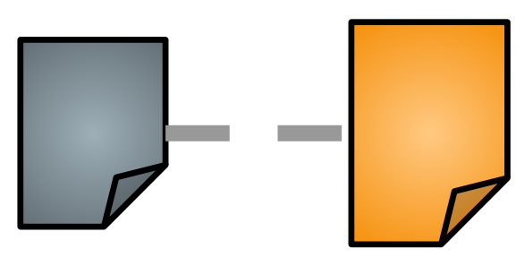
Aggregate properties

- Aggregate property (Q): Aggregate function computed on a set of subgraphs that satisfy given conditions (Q: graph → R)
  - ○ Properties involving multiple nodes, edges and labels

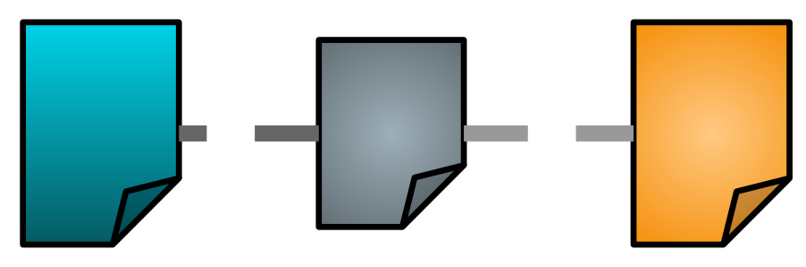
Q1: Category cohesion:
# of links across documents that belong to same category



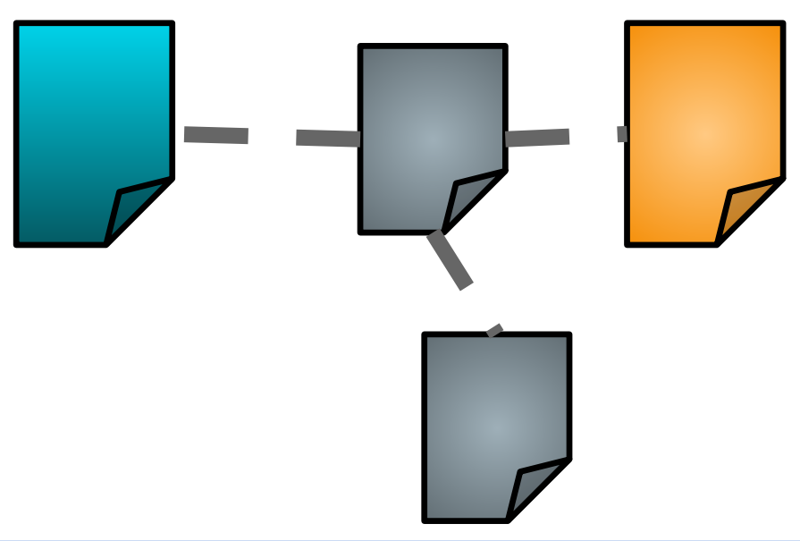
Q2: Category separation:
# of links across documents that belong to different category



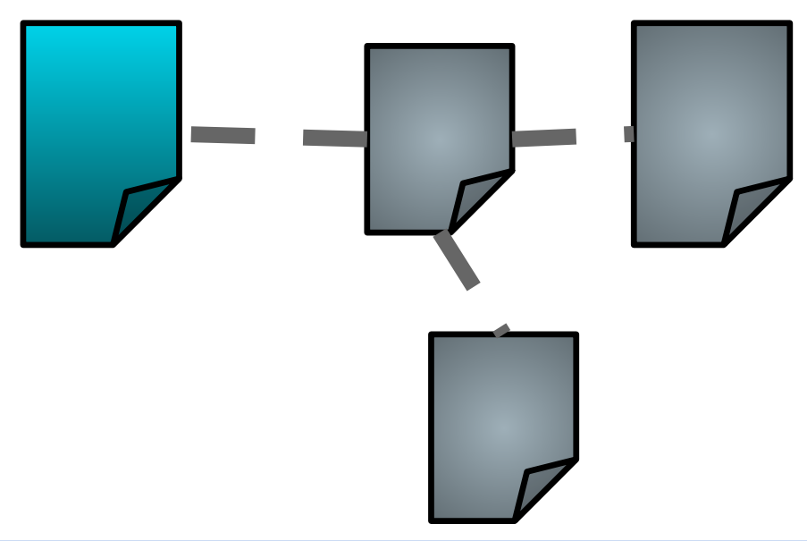
Q3: Diversity of influence:
# of nodes linked to at least half of all categories



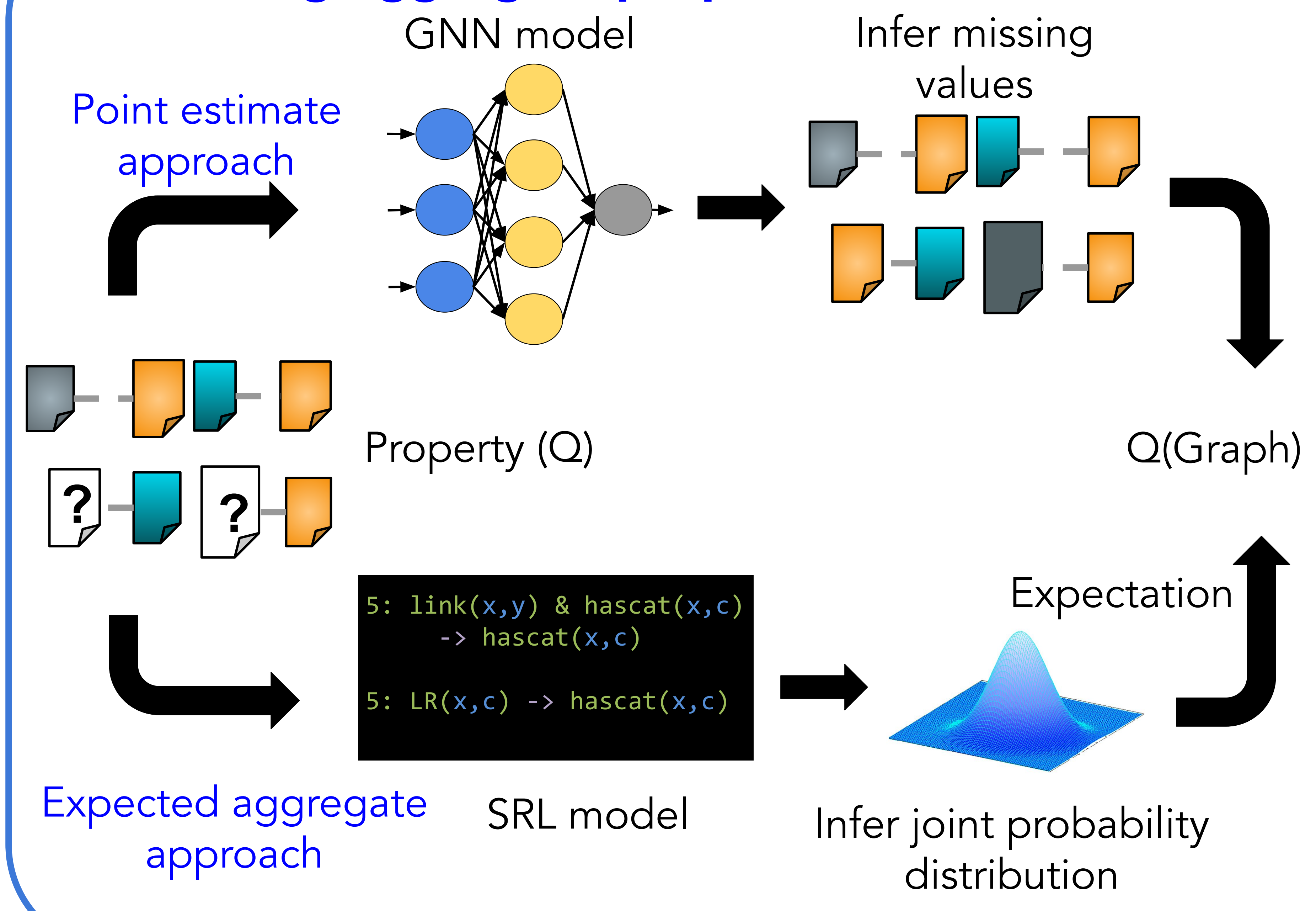
Q4: Exterior documents
# of nodes where half the neighbors belong to different categories



Q5: Interior documents
# of nodes where half the neighbors belong to same categories



Estimating aggregate properties



```
5: link(x,y) & hascat(x,c)
     -> hascat(x,c)

5: LR(x,c) -> hascat(x,c)
```

Tractable expectation computation for PSL

- Probabilistic Soft Logic[1] is a state-of-the-art SRL framework
- Computing expectation is intractable due to integration
- Monte Carlo approximation using samples from Gibbs sampler

Challenge 1

High rule weight ➡ Correlated RVs ➡ Slow convergence

- Identify association blocks from rules using
  - ○ Rule weights
  - ○ Feasible region
- Block sample RV in associated blocks

Challenge 2

Conditional distribution for Gibbs sampler

$$p(y_i \,|\, X, Y_{-i}) \propto exp\{-\sum_{r=1}^{N_i} w_r \phi_r(y_i, X, Y_{-i})\}$$

Hard to sample from

- Single step of Metropolis sampler inside gibbs sampler

$$\alpha = \frac{exp\{-\sum_{r=1}^{N_i} w_r \phi_r(y_i', X, Y_{1:i-1}^{(t+1)}, Y_{i:n}^{(t)})\}}{exp\{-\sum_{r=1}^{N_i} w_r \phi_r(y_i, X, Y_{1:i-1}^{(t+1)}, Y_{i:n}^{(t)})\}}$$

Acceptance ratio

[1] http://psl.linqs.org

Experimental evaluation

Data: Cora, Pubmed and Citeseer

Graph Neural Networks : Graph Convolutional Networks (GCN), Graph Attention Network (GAT), Graph Markov Neural Networks (GMNN)
Statistical Relational Learning: Markov Logic Networks (MLN), Probabilistic Soft Logic (PSL)

Metric: Relative error

Aggregate property estimation:

Methods	Q1	Q2	Q3	Q4	Q5	Average
PSL-MAP	0.13	0.528	0.396	0.714	0.121	0.377
MLN-MAP	0.109	0.491	0.281	0.570	0.102	0.310
PSL-MEAN	0.117	0.474	0.348	0.685	0.115	0.347
MLN-MEAN	0.064	0.261	**0.113**	**0.362**	**0.053**	0.170
GCN	0.089	0.361	0.169	0.626	0.102	0.269
GAT	0.129	0.526	0.293	0.709	0.119	0.355
GMNN	0.156	0.513	0.299	0.679	0.119	0.353
PSL-SAMPLES	0.108	0.441	0.312	0.618	0.105	0.316
MLN-SAMPLES	**0.060**	**0.210**	0.119	0.391	0.061	**0.168**

Pubmed

Methods	Q1	Q2	Q3	Q4	Q5	Average
PSL-MAP	0.175	0.527	0.673	0.57	0.272	0.443
MLN-MAP	0.207	0.648	0.594	0.794	0.392	0.527
PSL-MEAN	**0.134**	**0.403**	0.544	0.551	0.253	0.377
MLN-MEAN	0.137	0.731	0.792	0.691	0.315	0.554
GCN	0.211	0.637	0.712	0.813	0.396	0.553
GAT	0.248	0.747	0.9	0.887	0.416	0.639
GMNN	0.257	0.774	0.881	0.906	0.447	0.653
PSL-SAMPLES	0.137	0.413	**0.539**	**0.499**	**0.236**	**0.364**
MLN-SAMPLES	0.244	0.736	0.793	0.691	0.315	0.555

Citeseer

Methods	Q1	Q2	Q3	Q4	Q5	Average
PSL-MAP	0.047	0.205	0.165	0.1	0.062	0.115
MLN-MAP	0.032	**0.046**	0.412	0.436	0.242	0.234
PSL-MEAN	0.021	0.090	0.027	0.054	0.041	0.047
MLN-MEAN	0.038	0.163	**0.009**	0.174	0.068	0.090
GCN	0.048	0.207	0.137	0.671	0.34	0.28
GAT	0.073	0.313	0.376	0.697	0.355	0.362
GMNN	0.071	0.306	0.174	0.711	0.352	0.322
PSL-SAMPLES	**0.014**	0.061	0.050	**0.053**	**0.031**	**0.041**
MLN-SAMPLES	0.045	0.161	0.042	0.173	0.068	0.097

Cora

Predictive accuracy

Methods	Cora Acc (%)	Pubmed Acc (%)	Citeseer Acc (%)
PSL-MAP	**85.34**	**83.6**	**72.25**
MLN-MAP	77.9	76.75	71.7
PSL-MEAN	84.13	83.16	71.7
MLN-MEAN	82.35	75.14	71.25
GCN	81.96	77.73	68.78
GAT	81.43	76.87	70.41
GMNN	83.26	81.07	70.15
PSL-SAMPLES	83.01	81.88	71.29
MLN-SAMPLES	82.25	73.48	71.11

Runtime

Methods	Cora Time (sec)	Pubmed Time (sec)	Citeseer Time (sec)
PSL-MAP	14	124	37
PSL-MEAN	105	638	124
MLN-MEAN	270	1947	166
MLN-MAP	65	368	36
GCN	24	59	29
GAT	142	138	122
GMNN	30	17	8
PSL-SAMPLES	105	638	124
MLN-SAMPLES	270	1947	166

Conclusion

- Defined a suite of practical aggregate properties
- Proposed a novel sampling framework for PSL
- Extensive evaluation shows SRL approaches outperform GNNs when estimating aggregate properties

Supported by the NSF grants CCF-1740850 and IIS-1703331 and by AFRL and the DARPA.