【ICDE 2014勉強会】

Session 25:
Uncertain and Probabilistic Data

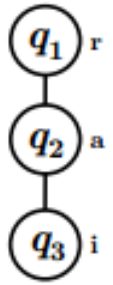
担当：胡（名大）

Some figures are copied from ICDE 2014 proceedings.

Subgraph Pattern Matching over Uncertain Graphs with Identity Linkage Uncertainty

- Walaa Eldin Moustafa (U. Maryland), Angelika Kimming (KU Leuven),
- Amol Deshpande, Lise Getoor (U. Maryland)
- 不確実グラフへのクエリ
- 背景：

クエリグラフ：



情報抽出と統合によるグラフデータの不確実性：

- 例：同じ実世界エンティティは異なるデータソースに引用され、統合されたとき、不確実性を生みだす

3種類の不確実性

- identity uncertainty（同一性）
- attribute value uncertainty

（ラベルの値の不確実性）

- edge existence uncertainty

（エッジは存在するかどうか）

動機：共同でこれらの不確実性を対処する方法はまだ不足している

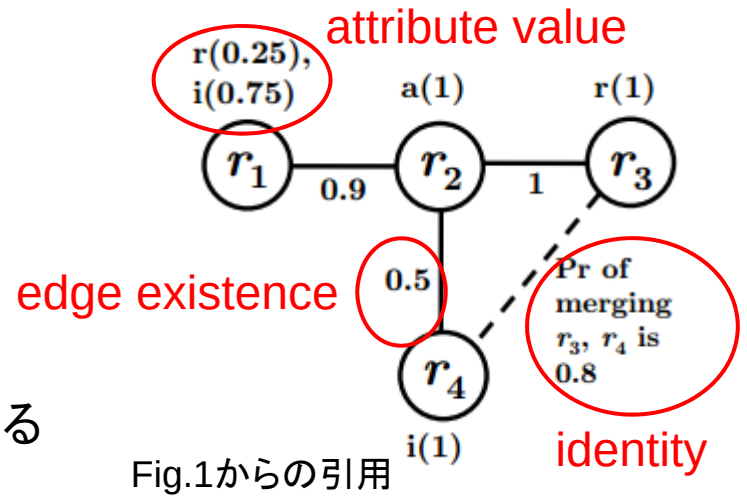
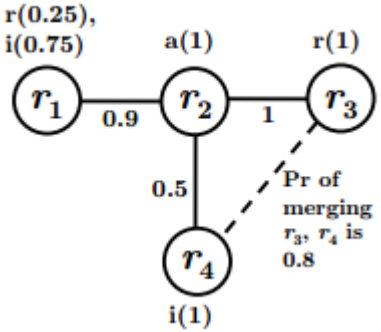


Fig.1からの引用

PEGの提案

- PEG（probabilistic entity graph）：エンティティレベルで不確実グラフの分布を定義する
  - 同一不確実性の定義(f^N(s1.n=v1,…,sk.n=vk))

$$f^N(s_1.n = v_1, \ldots, s_k.n = v_k) =$$
$$\begin{cases} p^s(s_i.x = T) & \text{if } v_i = T \text{ and, for all } j \neq i, v_j = F \\ 0 & \text{otherwise.} \end{cases}$$



例：
1)f^N(s1.n=T,other=F)=1
2) f^N(s3.n=T,other=F)=0.25
3) f^N(s34.n=T,other=F)=0.5

  - ラベル値の不確実性の定義　(Pr(s.l))

$$Pr(s.\boldsymbol{l}) = \left[m^{\Sigma}(\{p^r | r \in s\})\right](s.\boldsymbol{l})$$

  - エッジの存在の不確実性の定義　(Pr((s$_1$,s$_2$).e))

$$Pr((s_1, s_2).\boldsymbol{e}) = \left[m^{\{T,F\}}(\{p^{(r_1,r_2)} | r_i \in s_i\})\right]((s_1, s_2).\boldsymbol{e})$$

例:

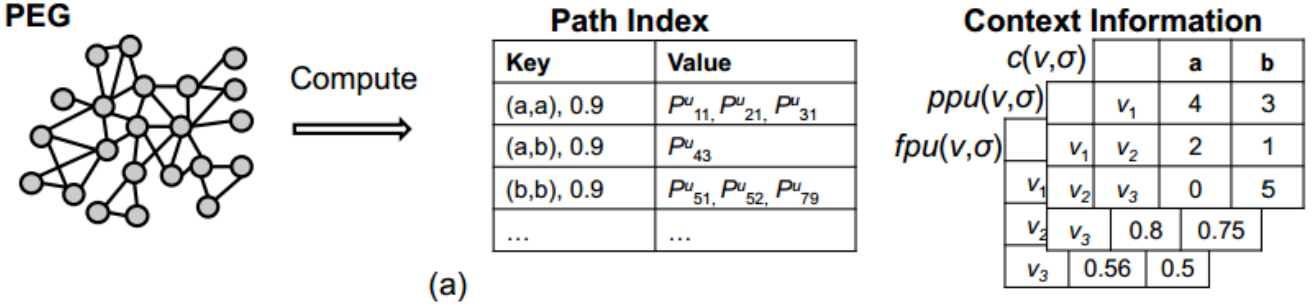
Pr(s1.l = r) = 0.25 ; Pr(s1.l = i) = 0.75

Pr((s1,s2).e=T)=0.9

サブグラフパターンマッチングアルゴリズム

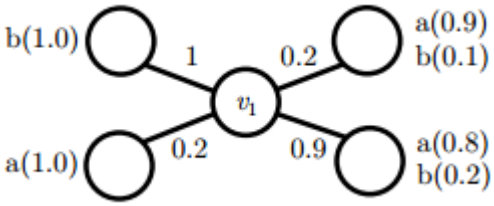
手法：コンテキストアウェアパスインデキシング
候補結合による削減

Offline Phase



- コンポーネントの確率を事前計算
- パスだけでなく、そのパスの近傍の文脈情報も含めてのインデックス

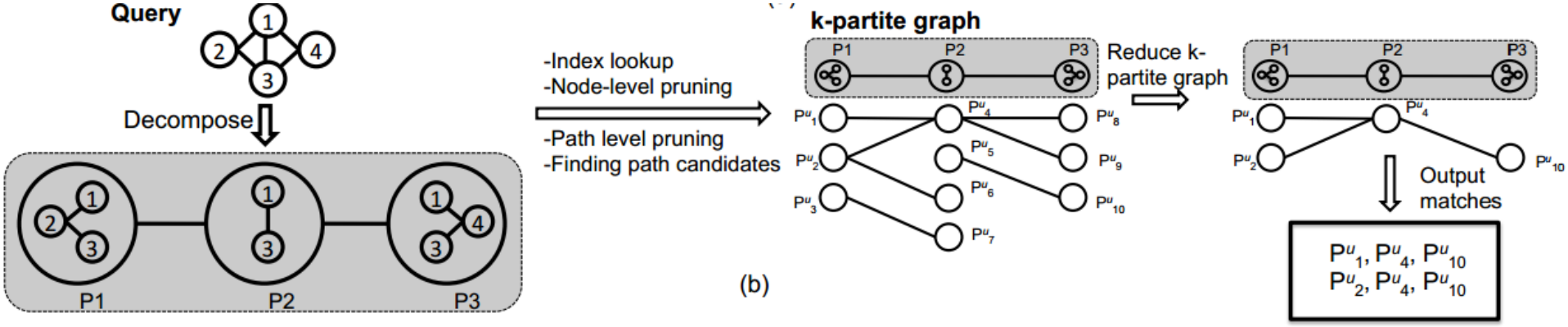
例：
Cardinality a|bを含めて近傍ノードの個数
Partial Probability Upperbound
a=0.9; b=1.0
Full Probability Upperbound:
a=0.9X0.8; b=1X1.0



σ	a	b
$c(v_1,\sigma)$	3	3
$ppu(v_1,\sigma)$	0.9	1.0
$fpu(v_1,\sigma)$	0.72	1.0

サブグラフパターンマッチングアルゴリズム

Online Phase



1)クエリからの分解

2)分解したクエリ毎にパス候補を探す

3)結合できる候補パスを探し出す

4)k-partite graphに基づく候補の削減

5)フールクエリのマッチを探し出す

【ICDE 2014勉強会】

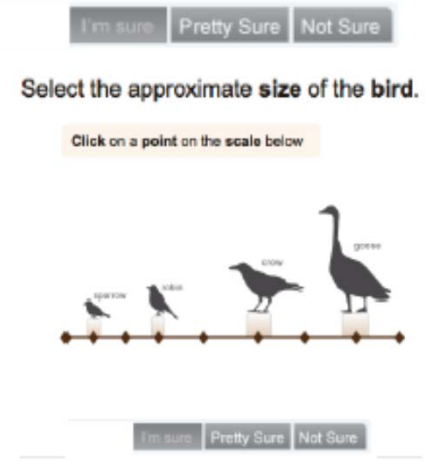
Session 25: Uncertain and Probabilistic Data

担当：石川佳治（名大）

Some figures are copied from ICDE 2014 proceedings.

User-Driven Refinement of Imprecise Queries

- Bahar Qarabaqi, Mirek Riedewald (Northeastern U., USA)
- 目的：曖昧な検索条件を対話的に洗練
- 特徴：確率的なアプローチ
- 例：鳥の観察画像の検索
  - 複数の属性：羽の色，大きさ，．．．
  - ユーザは特定の画像を検索したいのでなく，意図する鳥の種別（例：ツグミ）で絞りこみたい
  - ユーザは属性に対し信念確率を付与可能
  - 例：意図している鳥の多くが赤い羽根を持つとき，hasRedColorWing属性に（yes = 80%, no = 20%）という確率分布を指定



インタフェースのイメージ

- 属性 X_2 = Size と属性 X_4 = MainColor に分布を指定した後の状況
- 類似度によるランキング（中）
- 未指定の属性で結果を改善しそうなものの順位
- 既指定の属性の感度（sensitivity）

感度：確率分布の値をいじるとどのくらい敏感に反応するか

Attribute	Quality Improvement
X_3:ShapeGroup	83.27
X_{14}:BillLength	81.12
X_{11}:WingColor	74.98
X_1:Time	69.47
X_2:Location	65.24
X_5:BreastColor	64.02
X_6:BreastPattern	63.18
X_9:BackColor	57.79
X_{16}:LegColor	49.35
X_8:BellyPattern	48.81

Rank	Species	Image
1	C_{245}:Eastern Bluebird	
2	C_{211}:Blue Jay	
3	C_{223}:Barn Swallow	
4	C_{212}:Western Scrub-Jay	
5	C_{233}:Red-breasted Nuthatch	
6	C_{242}:Blue-grey Gnatcatcher	
7	C_{221}:Tree Swallow	
8	C_{331}:Indigo Bunting	
9	C_{246}:Western Bluebird	
10	C_{210}:Steller's Jay	

Specified Attribute	Sensitivity Score
X_2:Size	206.88
X_4:MainColor	18.01

技術的な部分

- 感度分析(Sensitivity Analysis)
  - 取りうる確率分布の値の変化で，ランクが最大限どの程度変化するかを評価：この論文で頑張っている部分(面倒)
  - ランクの違いはMinkowski距離で評価
  - 確率に関する性質をうまく使って効率的に判定
- 追加条件のランキング(画面の左側)
  - 改善度の期待値(expected improvement)を評価
    - 注目する画像(エンティティ)を高く，他をより低く評価するのがよい
  - エントロピーの概念を用いた定式化
  - 実直な評価手法では対話性に難あり
    - 決定木のアンサンブルによる近似的な評価