INTRODUCTION

Knowledge bases (KBs) constructed by human collaboration and information extraction (IE) are often huge but still incomplete or uncertain. To facilitate applications of these knowledge bases, we build a system for end users to query uncertain knowledge. We study the following key problems:

- **Incompleteness**: Knowledge cannot be fully captured and represented by structured data;
- **Uncertainty**: IE algorithms generate uncertain knowledge;
- **Scalability**: Today’s information scale results in huge KBs.

CONTRIBUTIONS

- Web interface for loading, querying, and updating uncertain knowledge bases;
- Query-driven inference by K-hop approximation;
- Efficient KB expansion by applying rules in batches;
- In-database integration.

PROBABILISTIC KNOWLEDGE BASES

- At $t_1$, public information showed Barack Obama was born in Hawaii.
- At $t_2$, anonymous emails questioned Obama’s birthplace and indicated that “Obama isBornInCountry Kenya.”
- At $t_3$, the local newspaper Star-Bulletin advertised the birthplace of Obama.
- At $t_4$, the Hawaii Department of Health released Obama’s birth certificate.
- The query “Obama isBornInCountry USA” returns probabilities 1.00, 0.84, 0.90 and 0.97 at $t_1$ to $t_4$, respectively.

EFFICIENT KNOWLEDGE EXPANSION

We design a relational model of probabilistic facts and rules;
We apply the inference rules in batches by performing join queries on the facts and rules tables;
We store the ground factor graph in database tables with UDA-GIST integration;
We achieve 300 times of speedup compared to the state-of-the-art.

QUERY-DRIVEN INFERENCE

- We use query-driven inference by extracting K-hop sub-networks to approximate the ground graphical model;
- We compute probabilities with efficient MCMC, utilizing the UDA-GIST big data analytics framework;
- We achieve 18 times of speedup, with an acceptable error of 0.04 in the computed probabilities.

DEMONSTRATION

- Dataset: Reverb-Sherlock (15M facts, 30912 rules), NELL-Sports (84.6M facts, 1828 rules);
- Query result with update probabilities;
- Visualization of factor graph, lineage, and explanations.