Functional Dependencies for Graphs

Wenfei Fan[1,2] Yinghui Wu[3] Jingbo Xu[1,2]

[1]University of Edinburgh [2]Beihang University [3]Washington State University

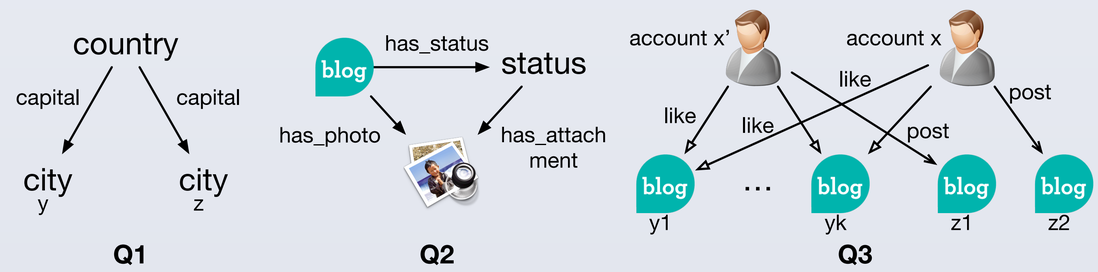
INTRODUCTION

- We propose a class of functional dependencies for graphs, referred to as GFDs. GFDs capture both attribute-value dependencies and topological structures of entities, and subsume conditional functional dependencies (CFDs) as a special case.
- We settle **satisfiability** and **implication** problem for GFDs.
- As one of applications of GFDs, we study the validation problem, to detect errors in graphs by using GFDs as data quality rules.
- We experimentally verify the effectiveness and efficiency of our GFD techniques.

GFDS: SYNTAX AND SEMANTICS

GFDs. A GFD ϕ is a pair $(Q[\bar{x}], X \rightarrow Y)$, where

- $Q[\bar{x}]$ is a graph pattern, called the pattern of ϕ; and
- X and Y are two (possibly empty) sets of literals of $\bar{x}$.



- GFD $\phi 1$ = (Q1[x, y, z], $\emptyset \rightarrow$ y.val = z.val). It is to ensure that for all country entities x, if x has two capital entities y and z, then y and z share the same name.
- GFD $\phi 2$ = (Q2[x, y, z], $\emptyset \rightarrow$ x.text = y.desc). It states that if entities x, y and z satisfy the topological constraint of Q2, then the annotation of status x of blog z must match the description of photo y included in z.
- GFD $\phi 3$ = (Q3[x, x′ , $y_1, \ldots, y_k, z_1, z_2$], $X_3 \rightarrow Y_3$), where X_3 includes x′.is_fake = true, z_1.keyword = c, z_2.keyword = c, and Y_3 is x.is_fake = true; here c is a constant indicating a peculiar keyword. It states that for accounts x and x′, if the conditions in X_3 are satisfied, including that x′ is confirmed fake, then x is also a fake account.

Special cases:

- Relational FDs and CFDs are special cases of GFDs.
- **constant GFDs** subsume constant CFDs, and **variable GFDs** are analogous to traditional FDs
- GFDs can specify certain type information.

REASONING ABOUT GFDS

Satisfiability Problem: A set Σ of GFDs is satisfiable if Σ has a model; that is, a graph G such that (a) G |= Σ, and (b) for each GFD $(Q[\bar{x}], X \rightarrow Y)$ in Σ, there exists a match of Q in G.

Theorem: The satisfiability problem is coNP-complete for GFDs.

Implication Problem: A set Σ of GFDs implies another GFD ϕ, denoted by Σ |= ϕ, if for all graphs G such that G |= Σ, we have that G |= ϕ, i.e., ϕ is a logical consequence of Σ.

Theorem: The implication problem for GFDs is NP-complete.

PARALLEL QUANTIFIED MATCHING

An algorithm is parallel scalable if

$$T(|A|, |G|, n) = O\left(\frac{t(|A|, |G|)}{n}\right) + (n|A|)^{O(1)}$$

- $T(|A|, |G|, n)$: worst case running time for solving problem A over graph G using n processors
- $t(|A|, |G|)$: worst case running time of sequential algorithm

(1) Parallel scalable algorithm `repVal` for replicated graph:

- *Graph G is replicated at each processors*
- Workload balancing: Workload estimation and partition
- Local error detection: Upon receiving the assigned $W_i(\Sigma)$, procedure `localVio` computes the local violation set $Vio_i(\Sigma, G)$ at each processor S_i in parallel.

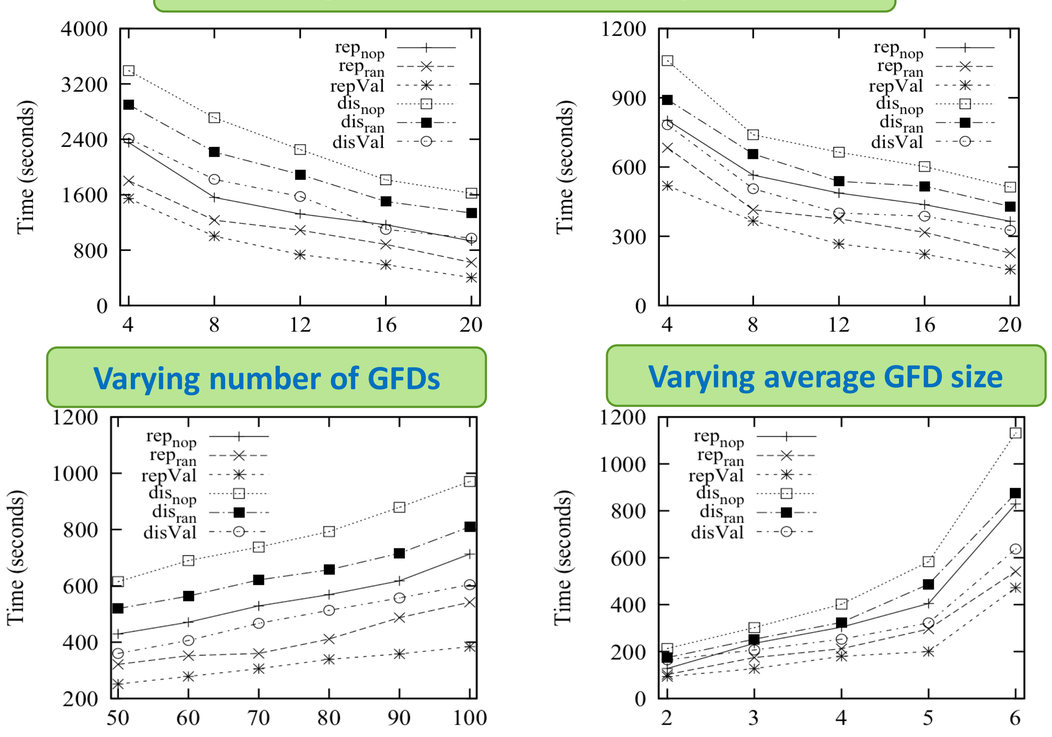
(2) Parallel scalable algorithm `disVal` for distributed graph:

- *Graph G may have already been fragmented and distributed across processors.*
- Bi-criteria balancing: Workload estimation and partition.
- Local error detection: (a) pre-fetch, (b) partial detection.

EXPERIMENTAL STUDY

- **DBPedia**: knowledge graph with 28 million entities of 200 types and 33.4 million edges of 160 types.
- **Pokec**: 1.63M nodes of 269 types, and 30.6M edges of 11 types.
- **Yago2**: 3.5M entities of 13 types and 7.35M links of 36 types.

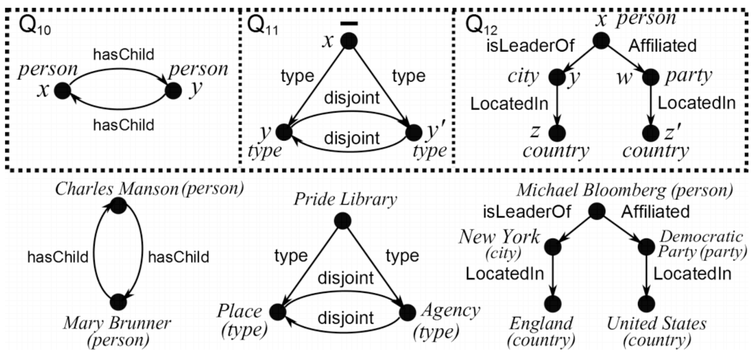
On average 3.7 (2.4) time faster repVal (disVal)



Varying number of GFDs

Varying average GFD size

Real-life GFDs and captured inconsistencies in the datasets



CONCLUSION

We have proposed GFDs, established complexity bounds for their classical problems, and provided parallel scalable algorithms for their application. Our experimental results have verified the effectiveness of GFD techniques.