Cost-efficient Data Acquisition on Online Data Marketplaces for Correlation Analysis

VLDB'19

Yanying Li[1] Haipei Sun[1] Boxiang Dong[2]
Hui (Wendy) Wang[1]

[1]Stevens Institute of Technology
Hoboken, NJ

[2]Montclair State University
Montclair, NJ

August 28, 2019

Data Marketplace

The rising demand for valuable online datasets has led to the emergence of data marketplace.

Data seller Specify data views for sale and their prices.

Data shopper Decide which views to purchase.

Data Acquisition

We consider data shopper's need as correlation analysis.

Age	Zipcode	Population
[35, 40]	10003	7,000
[20, 25]	01002	3,500
[55, 60]	07003	1,200
[35, 40]	07003	5,800
[35, 40]	07304	2,000

(a) D_S: *Source* instance owned by data shopper Adam

Zipcode	State	
07003	NJ	correct
07304	NJ	correct
10001	NY	correct
10001	NJ	wrong

D_1: Zipcode table
(FD: Zipcode → State)

State	Disease	# of cases
MA	Flu	300
NJ	Flu	400
Florida	Lyme disease	130
California	Lyme disease	40
NJ	Lyme disease	200

D_2: Data and statistics of diseases by state

Age	Address	Insurance	Disease
[35, 40]	10 North St.	UnitedHealthCare	Flu
[20, 25]	5 Main St.	MedLife	HIV
[35, 40]	25 South St.	UnitedHealthCare	Flu

D_3: Insurance & disease data instance

(b) Relevant instances on data marketplace

Need: Find correlation between age groups and diseases in New Jersey

Data Acquisition

- **Requirement 1: Meaningful join**

$D_S \bowtie D_3$ is meaningless, as it associates the aggregation data with individual records.

Age	Zipcode	Population	Address	Insurance	Disease
[35, 40]	10003	7,000	10 North St.	UnitedHealthCare	Flu
[35, 40]	10003	7,000	25 South St.	UnitedHealthCare	Flu
[20, 25]	01002	3,500	5 Main St.	MedLife	HIV
[35, 40]	07003	5,800	10 North St.	UnitedHealthCare	Flu
[35, 40]	07003	5,800	10 North St.	UnitedHealthCare	Flu
[35, 40]	07304	2,000	25 South St.	UnitedHealthCare	Flu
[35, 40]	07304	2,000	25 South St.	UnitedHealthCare	Flu

$D_S \bowtie D_3$

Data Acquisition

- Requirement 1: Meaningful join
- **Requirement 2: High data quality**

We consider data inconsistency as the main quality issue.

Zipcode	State	
07003	NJ	correct
07304	NJ	correct
10001	NY	correct
10001	NJ	wrong

FD: $\mathit{Zipcode} \rightarrow \mathit{State}$

Data Acquisition

- Requirement 1: Meaningful join
- Requirement 2: High data quality
- **Requirement 3: Budget constraint**

The data shopper has a purchase budget. The price of the purchased datasets must be within the budget.

Our Contributions

We design a middleware service named DANCE, a Data Acquisition framework on oNline data market for CorrElation analysis that

- provides cost-efficient data acquisition service;
- enables budget-conscious search of the high-quality data;
- maximizes the correlation of the desired attributes.

Outline

Related Work

Data Market

- Query-based pricing model [KUB+15]
- History-aware pricing model [U+16]
- Arbitrage-free pricing model [KUB+12, LK14, DK17]

Data Exploration via Join

- Summary graph [YPS11]
- Reverse engineering [ZEPS13]

Do not consider data quality and budget.

Preliminaries - Data Pricing

- In this paper, we mainly focus on query-based pricing functions [KUB+15].

  Input Explicit prices for a few views

  Output The derived price for any view

- DANCE is compatible with any pricing model.

Preliminaries - Data Quality

We define data quality as the fraction of tuples that are correct with regard to all the functional dependencies.

TID	A	B	C	D	E
t_1	a_1	b_2	c_1	d_1	e_1
t_2	a_1	b_2	c_1	d_1	e_1
t_3	a_1	b_2	c_2	d_1	e_1
t_4	a_1	b_2	c_3	d_1	e_2
t_5	a_1	b_3	c_3	d_2	e_2

FD: $A \rightarrow B$, $D \rightarrow E$

Preliminaries - Data Quality

We define data quality as the fraction of tuples that are correct with regard to all the functional dependencies.

TID	A	B	C	D	E
t_1	a_1	b_2	c_1	d_1	e_1
t_2	a_1	b_2	c_1	d_1	e_1
t_3	a_1	b_2	c_2	d_1	e_1
t_4	a_1	b_2	c_3	d_1	e_2
t_5	a_1	b_3	c_3	d_2	e_2

FD: $A \rightarrow B$, $D \rightarrow E$

$$C(D, A \rightarrow B) = \{t_1, t_2, t_3, t_4\}$$

Preliminaries - Data Quality

We define data quality as the fraction of tuples that are correct with regard to all the functional dependencies.

TID	A	B	C	D	E
t_1	a_1	b_2	c_1	d_1	e_1
t_2	a_1	b_2	c_1	d_1	e_1
t_3	a_1	b_2	c_2	d_1	e_1
t_4	a_1	b_2	c_3	d_1	e_2
t_5	a_1	b_3	c_3	d_2	e_2

FD: $A \rightarrow B$, $D \rightarrow E$

$$C(D, A \rightarrow B) = \{t_1, t_2, t_3, t_4\}$$

$$C(D, D \rightarrow E) = \{t_1, t_2, t_3, t_5\}$$

Preliminaries - Data Quality

We define data quality as the fraction of tuples that are correct with regard to all the functional dependencies.

TID	A	B	C	D	E
t_1	a_1	b_2	c_1	d_1	e_1
t_2	a_1	b_2	c_1	d_1	e_1
t_3	a_1	b_2	c_2	d_1	e_1
t_4	a_1	b_2	c_3	d_1	e_2
t_5	a_1	b_3	c_3	d_2	e_2

FD: $A \rightarrow B$, $D \rightarrow E$

$$C(D, A \rightarrow B) = \{t_1, t_2, t_3, t_4\}$$

$$C(D, D \rightarrow E) = \{t_1, t_2, t_3, t_5\}$$

$$Q(D) = \frac{3}{5} = 0.6$$

Preliminaries - Join Informativeness

Definition (Join Informativeness)

Given two instances D and D', let J be their join attribute(s). The join informativeness of D and D' is defined as

$$JI(D, D') = \frac{Entropy(D.J, D'.J) - I(D.J, D'.J)}{Entropy(D.J, D'.J)},$$

by using the joint distribution of $D.J$ and $D'.J$ in the output of the full outer join of D and D', where I calculates the mutual information.

- It penalizes those joins with excessive numbers of such unmatched values [YPS09].
- $0 \leq JI(D, D') \leq 1$.
- The smaller $JI(D, D')$ is, the more important is the join connection between D and D'.

Preliminaries - Correlation Measurement

Definition (Correlation Measurement)

Given a dataset D and two attribute sets X and Y, the *correlation* of X and Y $CORR(X, Y)$ is measured as

- $CORR(X, Y) = Entropy(X) - Entropy(X|Y)$ if X is categorical,
- $CORR(X, Y) = h(X) - h(X|Y)$ if X is numerical,

where $h(X)$ is the cumulative entropy of attribute X

$$h(X) = -\int P(X \leq x) log P(X \leq x) dx,$$

and

$$h(X|Y) = -\int h(X|y) p(y) dy.$$

Problem Statement

Input A set of data instances $\mathcal{D} = \{D_1, \ldots, D_n\}$, source attributes $\mathcal{A}_\mathcal{S}$, and target attributes $\mathcal{A}_\mathcal{T}$, purchase budget B, join informativeness threshold α, quality threshold β

Output A set of data views $\mathbb{T} \subseteq \mathcal{D}$ s.t.

$$\underset{\mathbb{T}}{\text{maximize}} \quad CORR(\mathcal{A}_\mathcal{S}, \mathcal{A}_\mathcal{T}) \backslash\backslash \textit{correlation}$$

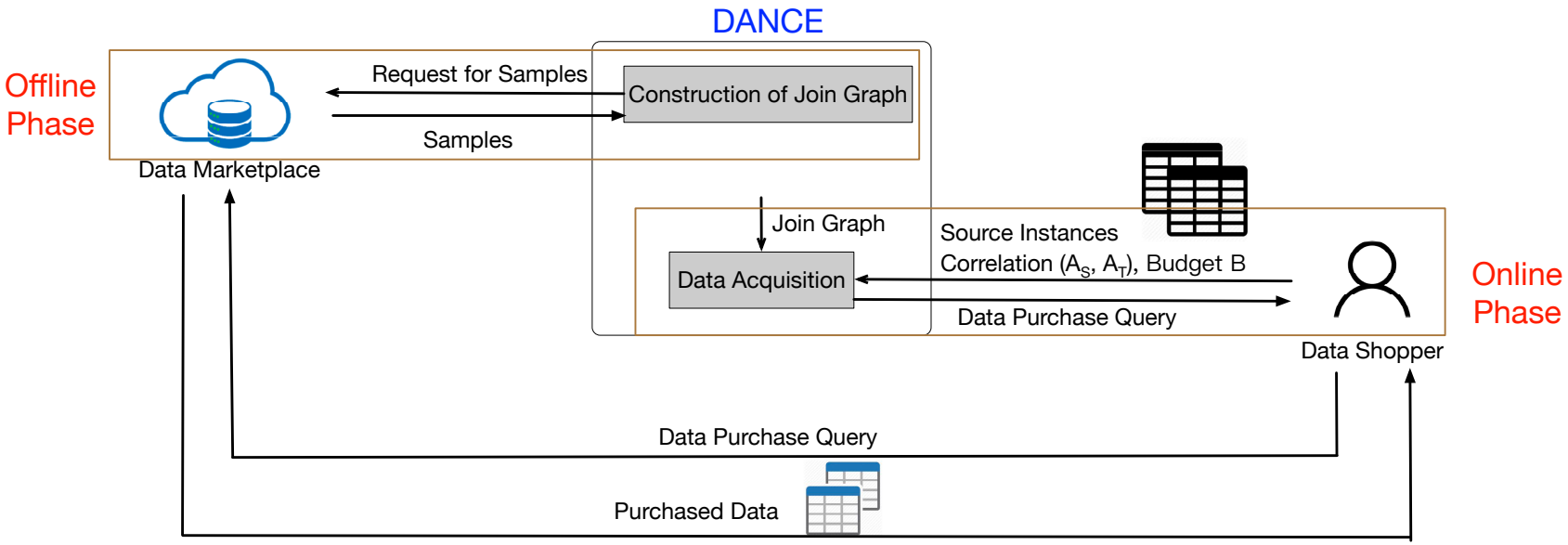
$$\text{subject to} \quad \forall T_i \in \mathbb{T}, \exists D_j \in \mathcal{D} \; s.t. \; T_i \subseteq D_j,$$

$$\sum_{T_i \in \mathbb{S} \cup \mathbb{T}} JI(T_i, T_{i+1}) \leq \alpha, \backslash\backslash \textit{informativeness}$$

$$Q(\mathbb{T}) \geq \beta, \backslash\backslash \textit{quality}$$

$$p(\mathbb{T}) \leq B. \backslash\backslash \textit{budget}$$

Framework of DANCE



Offline Phase Construct a two-layer join graph of the datasets on the marketplace.

Online Phase Process data acquisition requests.

Dealing with Large-scale Data

Correlated Sampling $S = \{t_i \in D \mid h(t_i[J]) \leq p\}$

Estimation from Samples

- $E(JI(S_1, S_2)) = JI(D_1, D_2)$
- $E(Q(S_1 \ltimes S_2)) = Q(D_1 \ltimes D_2)$
- $E(CORR_{S_1 \ltimes S_2}(\mathcal{A}_\mathcal{S}, \mathcal{A}_\mathcal{T})) = CORR_{D_1 \ltimes D_2}(\mathcal{A}_\mathcal{S}, \mathcal{A}_\mathcal{T})$

Re-sampling We design a correlated-resampling method to deal with large join result from samples in case of long join paths.

Offline Phase: Construction of Join Graph

Construct a two-layer join graph from the data samples.

Instance layer

Nodes data instances

Edges join attribute and the minimum informativeness

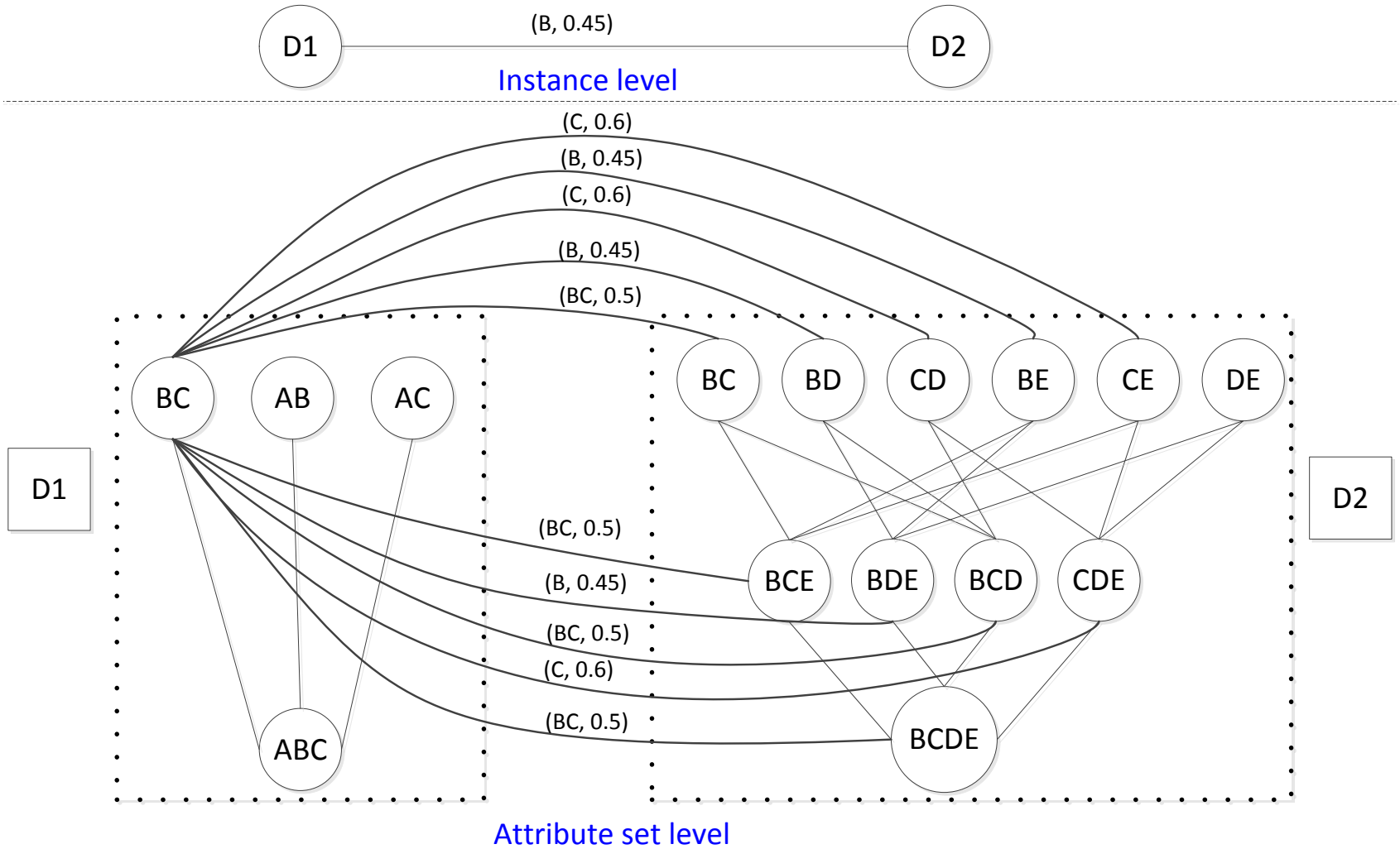
Attribute set layer

Nodes attribute sets

Edges join attribute and informativeness

Offline Phase: Construction of Join Graph

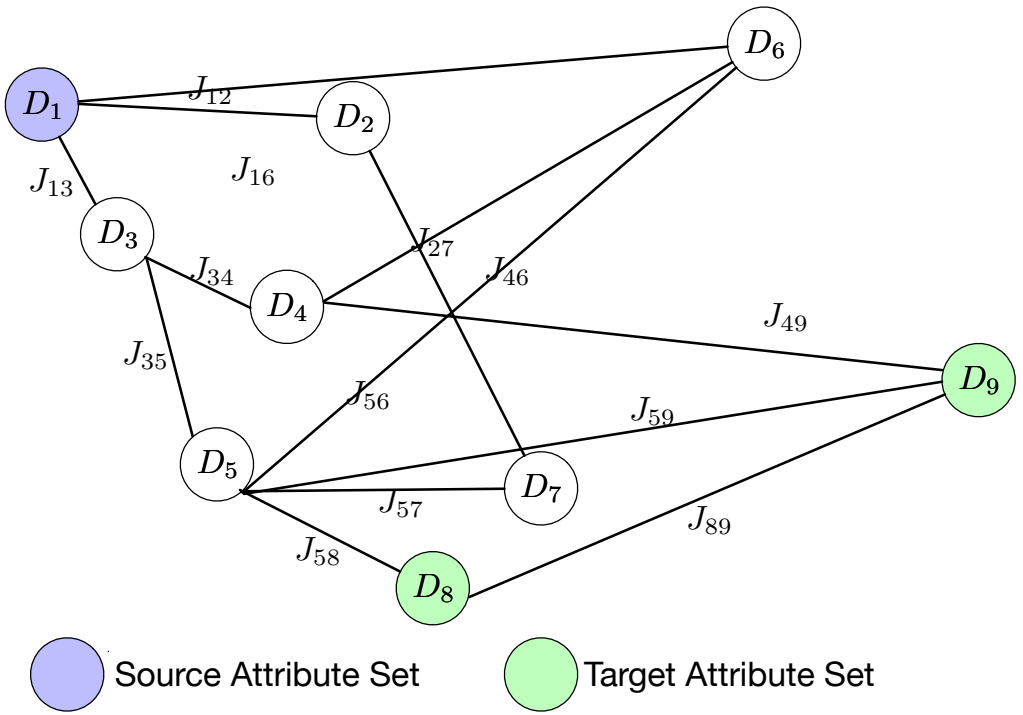
Construct a two-layer join graph from the data samples.



Online Phase: Data Acquisition

We design a two-step algorithm to search for the data views.

Step 1 Find minimal weighted graphs at instance layer.

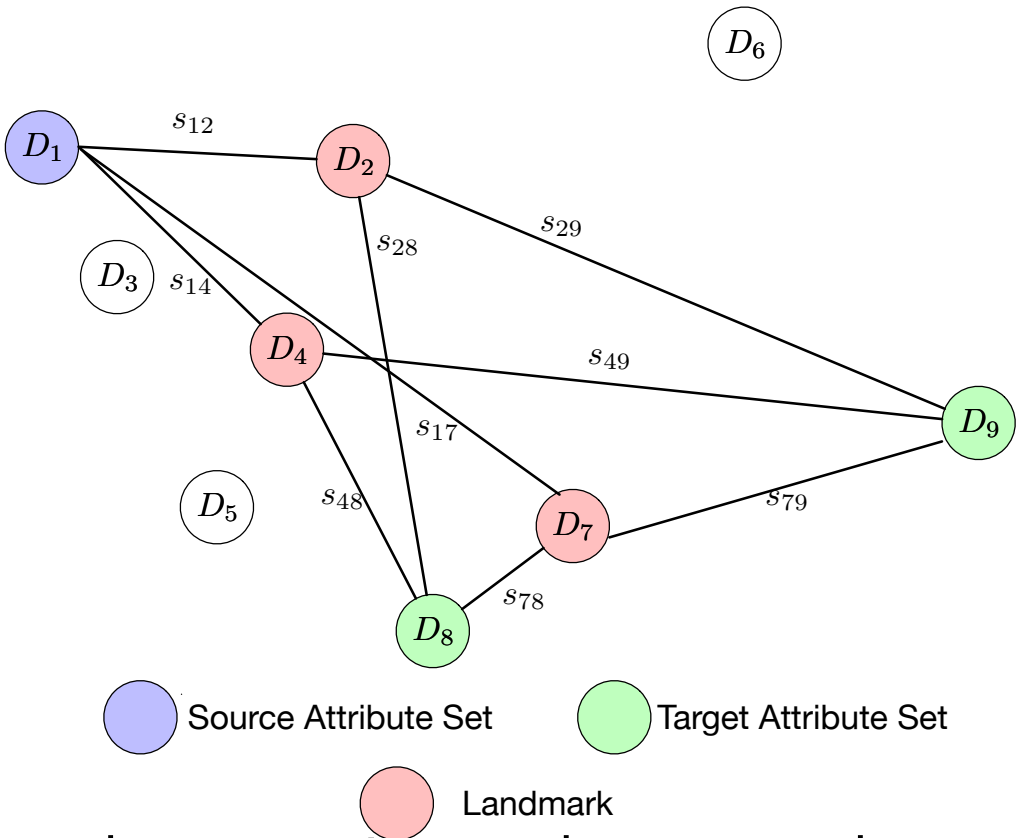


- It is equivalent to the Steiner tree problem and is NP-hard [Vaz13].

Online Phase: Data Acquisition

We design a two-step algorithm to search for the data views.

Step 1 Find minimal weighted graphs at instance layer.



- We adapt the approximate shortest path search algorithm [GBSW10] based on landmarks.

Online Phase: Data Acquisition

We design a two-step algorithm to search for the data views.

Step 1 Find minimal weighted graphs at instance layer.

Step 2 Find optimal target graphs at attribute set layer based on Markov chain Monte Carlo (MCMC).

Input : A minimal weighted I-graph $\mathcal{IG}$
Output: A target graph G^* at the AS-layer

$G^* = NULL$;
$Max = 0$;
$TG = \mathcal{IG}$;
for $i = 1$ *to* ℓ **do**
  Randomly pick an edge $e_{i,j} \in TG$;
  Randomly pick a different edge $e'_{i,j}$ between (v_i, v_j);
  Let TG' be the new target graph;
  if $p(TG') \leq B \wedge w(TG') \leq \alpha \wedge Q(TG') \geq \beta$ **then**
    if *accept* $e'_{i,j}$ *by probability* $\min(1, \frac{CORR(TG')}{CORR(TG)})$ **then**
      $TG = TG'$;
      if $CORR(TG) > Max$ **then**
        $G^* = TG$;
        $Max := CORR(G^*)$;
      end
    end
  end
end
Return G^*;

Experiments

Datasets

- *TPC-E* benchmark
- *TPC-H* benchmark

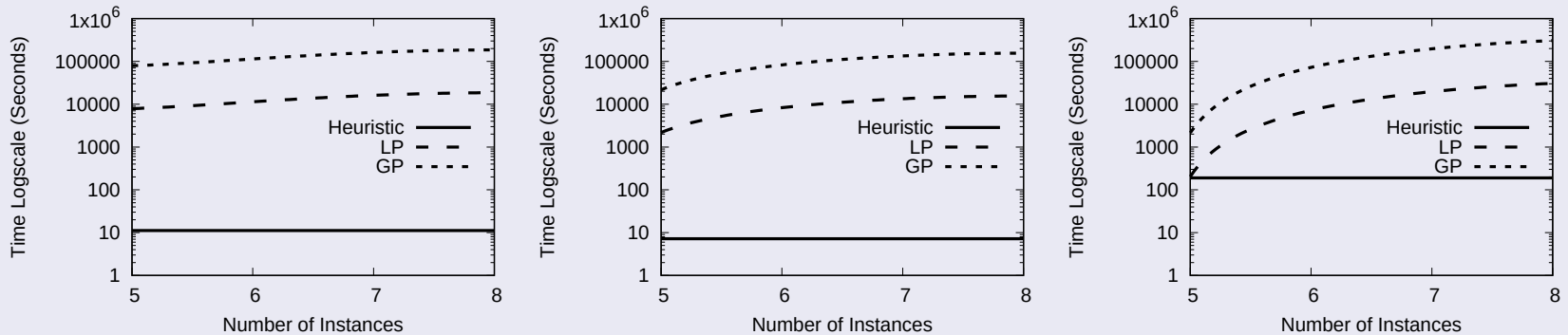
Baselines

- *LP:* enumerate all join paths on samples
- *GP:* enumerate all join paths on original datasets

	# of instances	Max. instance size (# of records)	max. # of attributes	Avg # of FDs per table
TPC-H	8	6,000,000 (Lineitem)	20 (Lineitem)	39
TPC-E	29	10,001,048 (Watchitem)	28 (Customer)	33

Experiments

Time Performance



(a) Q_1 (b) Q_2 (c) Q_3

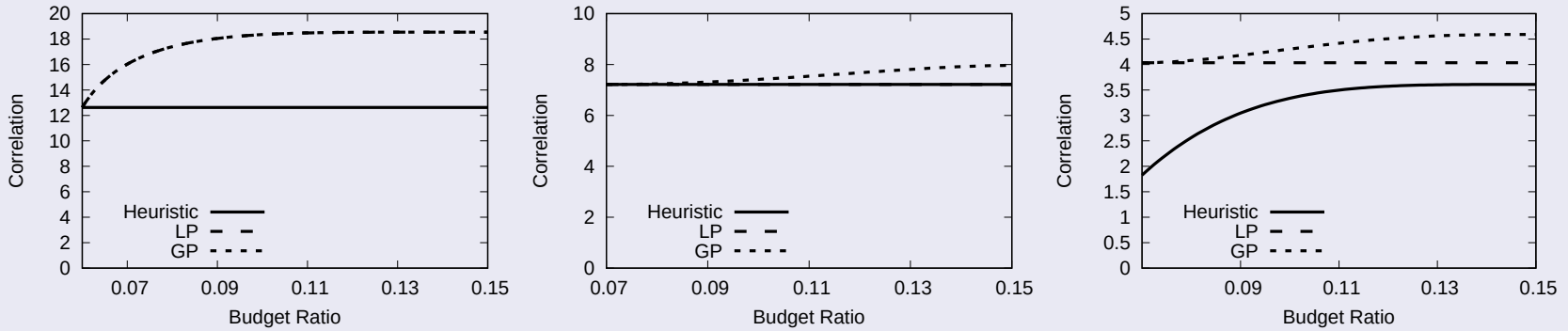
TPC-H dataset

- Our heuristic algorithm can be 2,000 times more efficient than LP, and 20,000 times more efficient than GP.

Query	Source	Target	Explanation
Q1	customer.account_balance	orders.clerk	link customers' account with responsible clerks
Q2	nation.name	partsupp.availqty	link parts with the nation of their suppliers
Q3	orders.total_price	region.name	associate orders' price with the origin region

Experiments

Correlation



(a) Q_1 (b) Q_2 (c) Q_3

TPC-H dataset

- In most cases, the difference of the correlation by our heuristic algorithm and LP/GP is no larger than 15% (our heuristic algorithm is at least 2000 times faster).

Conclusion

We design a middleware service named DANCE, a Data Acquisition framework on oNline data market for CorrElation analysis that

- provides cost-efficient data acquisition service;
- enables budget-conscious search of the high-quality data;
- maximizes the correlation of the desired attributes.

Q & A

Thank you!

Questions?

References I

[BHS11] Magdalena Balazinska, Bill Howe, and Dan Suciu.

Data markets in the cloud: An opportunity for the database community.

Proceedings of the VLDB Endowment, 4(12):1482–1485, 2011.

[DK17] Shaleen Deep and Paraschos Koutris.

The design of arbitrage-free data pricing schemes.

In *International Conference on Database Theory*, 2017.

[GBSW10] Andrey Gubichev, Srikanta Bedathur, Stephan Seufert, and Gerhard Weikum.

Fast and accurate estimation of shortest paths in large graphs.

In *Proceedings of ACM International Conference on Information and Knowledge Management*, pages 499–508, 2010.

[KEW09] Gjergji Kasneci, Shady Elbassuoni, and Gerhard Weikum.

Ming: mining informative entity relationship subgraphs.

In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 1653–1656, 2009.

References II

[KUB+12] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu.

Querymarket demonstration: Pricing for online data markets.

Proc. of the VLDB Endowment, 5(12):1962–1965, 2012.

[KUB+15] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu.

Query-based data pricing.

Journal of the ACM, 62(5):43, 2015.

[LK14] Bing-Rong Lin and Daniel Kifer.

On arbitrage-free pricing for general data queries.

Proceedings of the VLDB Endowment, 7(9):757–768, 2014.

[TF06] Hanghang Tong and Christos Faloutsos.

Center-piece subgraphs: problem definition and fast solutions.

In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 404–413, 2006.

References III

[TFK07] Hanghang Tong, Christos Faloutsos, and Yehuda Koren.

Fast direction-aware proximity for graph mining.

In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 747–756, 2007.

[U+16] Prasang Upadhyaya et al.

Price-optimal querying with data apis.

Proceedings of the VLDB Endowment, 2016.

[Vaz13] Vijay V Vazirani.

Approximation algorithms.

Springer Science & Business Media, 2013.

[YPS09] Xiaoyan Yang, Cecilia M Procopiuc, and Divesh Srivastava.

Summarizing relational databases.

Proceedings of the VLDB Endowment, 2(1):634–645, 2009.

References IV

[YPS11] Xiaoyan Yang, Cecilia M Procopiuc, and Divesh Srivastava.

Summary graphs for relational database schemas.

Proceedings of the VLDB Endowment, 2011.

[ZEPS13] Meihui Zhang, Hazem Elmeleegy, Cecilia M Procopiuc, and Divesh Srivastava.

Reverse engineering complex join queries.

In *Proceedings of the ACM International Conference on Management of Data*, pages 809–820, 2013.