UNIVERSITÀ DEGLI STUDI
DI GENOVA

dibris

Joint
Conference
2013

March, 18-22
Genoa, Italy

16th EDBT
International Conference on
Extending Database Technology

16th ICDT
International Conference on
Database Theory

Table of Content

1. Welcome

EDBT/ICDT 2013's Organizing Committee is pleased to welcome you to Genoa, for the joint edition of the 16th International Conference on Extending Database Technology and the 16th International Conference on Database Theory. Since they were first held in Italy, the conferences have traveled across Europe and are now returning to Italy. The conference venue is Palazzo Ducale, the Doge historical residence, in the heart of Genoa. In such a remarkable context, we trust you will find the program interesting and enjoyable.

The main reason for a conference success is its excellent technical program. The conference offers a well-integrated collage of events, including outstanding invited speakers, carefully refereed technical papers, selected tutorials, demonstrations, topic-focused workshops. This year's EDBT Program Chair is Norman Paton and ICDT Program Chair is Wang Chiew Tan. Kai-Uwe Sattler is responsible for workshops, Paolo Atzeni for tutorials, Piero Fraternali for demostrations and Malu Castellanos for industry & application sessions. Proceedings have been prepared by Anastasios Gounaris. With all of them it has been a great pleasure to cooperate.

We are delighted to announce outstanding keynote talks by Daniel Abadi, Jan van den Buscche, Luc Segoufin, C Mohan and an invited ICDT lecture by Yehoshua Sagiv. Besides them, we wish to thank all those who are organizing workshops, giving tutorials, presenting papers and demonstrations, offering posters for their significant efforts, all of which contribute to the richness of the conference program.

EDBT/ICDT is being organized by the Department of Computer Science, Bioengineering, Robotics and System Engineering of the University of Genoa. It is being held with the patronage of the Genova Municipality and under the aegis of EDBT Endowment and ICDT Council. We thank EDBT Endowment president, Marc Scholl, and ICDT Council chair, Thomas Schwentick, for their help and advices. The conference would not have been possible without the financial support of our generous sponsors. Special thanks go to SAP, Google, HP, IBM Research, Coop Liguria.

Welcome to Genoa and enjoy EDBT/ICDT 2013!

Giovanna Guerrini and Barbara Catania

on behalf of the Organizing Committee

UNIVERSITÀ DEGLI STUDI
DI GENOVA

dibris

We thank our sponsors

Gold Sponsors

SAP®

Silver Sponsors

Google™

hp®

IBM Research

Bronze Sponsors

coop

Liguria

We thank our supporters

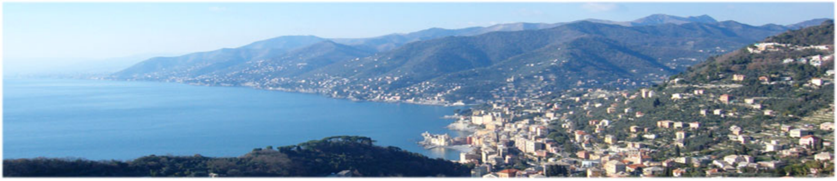
2. General Information

The city of Genova

Genova (referred in English as Genoa), the main city of Liguria, stretches along the bay of the same name from Voltri to the west as far as Nervi to the east, while the hinterland area takes in the lower parts of the Polcevera and Bisagno Valleys.

The original nucleus of the city, which already existed in pre-Roman times, developed around the Mandraccio wharf area and on Castello Hill, which overlooks it. In the ninth century, the Genoese built the first town walls and laid the foundations for the development of shipping and sea-trading, which would eventually make the Republic of Genova a Mediterranean sea power and create a dominion stretching across the entire region of Liguria. From the nineteenth century onwards, the great city port was flanked by large industrial areas. The old town district is one of the largest in Europe, and hosts some remarkable artistic and architectural treasures, including the Palazzi dei Rolli, fifty or so homes of the aristocracy entered on the UNESCO World Heritage List.

In addition to offering a wealth of cultural attractions, Genova is a fascinating destination for tourists, with its scenic vantage points, sea promenades, aristocratic villas and of course the Riviera to the east and west, both easy to reach: Porto Venere and Le Cinque Terre (UNESCO World Heritage Sites), Tigullio, Portofino and Camogli to the east and Alassio, Sanremo, Bordighera to the west.



Liguria

Liguria is a narrow strip of land, enclosed between the sea and the Alps and the Apennines mountains, it is a winding arched extension from Ventimiglia to La Spezia and is one of the smallest regions in Italy.

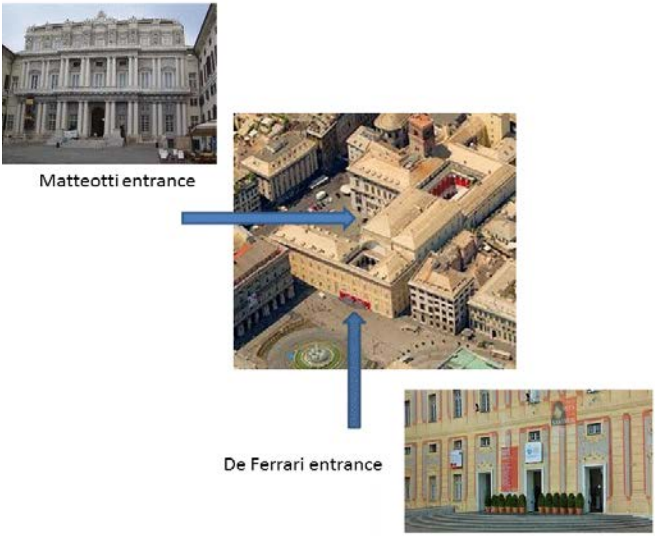
It is limited in size, but not in the variety of its vegetation and wildlife which is amongst the most diversified and interesting in Italy. The coast-line, which is geographically divided between the Western Riviera and the Eastern Riviera at the sides of the important center of Genova, from the scenic point of view is characterized by an alternating series of magnificent high coast-lines and flat, sandy coast-lines, whilst in the interior the steep hills meet up with the Apennines peaks.

Liguria has an abundance of natural beauty and the various names given to it such as “Paradise Gulf”, “Siren bay”, “Bay of silence”, “Bay of fairy tales”, “Sea's echo” are all a testimony to the magnificent beauty of these marine landscapes.

The host of hotels and seaside facilities ensure that the tourist can enjoy the very best kind of holiday. As an alternative to a morning spent on the beach there is the possibility of taking a trip into the hills, that are within easy reach as they sweep right down to the coast. There are numerous small villages, which often boast ruined castles that bear testimony to former glories of noble families. They are strewn around the interland, and provide a peaceful authentic setting away from the crowds, amongst the friendly hard-working local people. The ring of hills, lying immediately beyond the coast, together with the beneficial influx of the sea, account for the mild climate the whole year round (with average winter temperatures of 7-10°C and summer temperatures of 25-28°C) which makes for a pleasant stay even in the heart of winter.

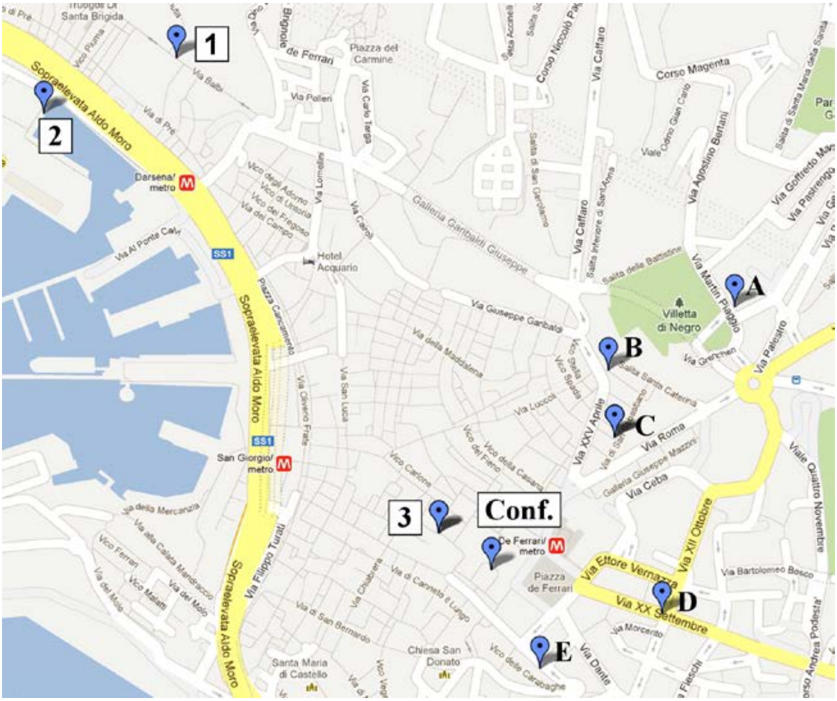
Conference Information

Conference venue



The conference will take place in the Doge's Palace (Italian: Palazzo Ducale), a historical building in Genova. Once the home of the Doges of Genova, it is now a museum and a center for cultural events and arts exhibitions. It was restored in 1992, in occasion of the celebrations of Christopher Columbus and the 500th anniversary of the discovery of America and hosted the G8 Summit in 2001. It is situated in the heart of the city, with two different entrances and facades, the main one on Piazza Matteotti, and the second one on Piazza De Ferrari. On the main floor, the so called Piano Nobile, are the frescoed halls of the Maggior and Minor Consiglio. On the basement floor is the picturesque space of the Munizioniere, the old munition depot. Situated in the city center, Palazzo Ducale is optimally connected through public transportation to the Genova Brignole (10 minutes by walking) and Genova Principe (15 minutes by walking) railway stations as well as to the Cristoforo Colombo airport (20 minutes by bus from the airport to the Brignole Station).

Area map



1. Welcome Reception - University Rectorate
2. Museum Visit and Social Dinner – Galata Maritime Museum
3. Sap-Sponsored Student Lunch – Museo Diocesano

A. NH Plaza
B. Best Western Metropoli
C. Best Western City
D. Bristol Palace
E. Colombo

Registration Desk

Registration desk will be positioned on Floor -1, just outside Munizioniere entrance on Monday 18 and Friday 22. It will be positioned on Floor 1, in the Foyer next to Maggior Consiglio from Tuesday 19 to Thursday 21.

Conference Secretariat

Conference secretariat (registration agency for receipts, on site registration payments) will be open at the desk: all day starting from 8.30 on Monday 18 and Tuesday 19; only in the morning 9-12.30 from Wednesday 20.

Staff and Student Volunteers

Staff members are wearing green profiled badges. Student volunteers are wearing white T-shirts.

Wireless Access

Three different WLANs will be available: (i) one in Maggior Consiglio and Loggiato Maggiore (Floor 1); (ii) one in Sala Camino (Floor 1): (iii) one in Munizioniere (Floor -1). Instruction on their use will be given during registration.

Information Board

Near the conference desk there will be a message and information board where you can leave messages and where the conference staff will post messages for participants and current information.

Luggage Facilities

Luggage can be stored in the wardrobe room at the left of the Foyer (Piano Nobile) on Tuesday, Wednesdayand Thursday. They can be stored in the Munizioniere on Monday and Friday. We are unable to accept any responsibility for loss or damage.

Catering Arrangements

Coffee breaks will be daily at **10:30-11:00** and at **15.30-16.00.** They will be served: on Monday 18 at *"Le Cisterne del Ducale"*, (Floor -2); from Tuesday 19 to Thursday 21 in Loggiato Maggiore (Floor 1); on Friday 22 in the Munizioniere.

Lunch will be served daily **12:30-14:00** at *"Le Cisterne del Ducale"* (Floor -2).

Other facilities

Conference participants can visit the Mirò Exibition (Floor 1) at the reduced price of Euro 8 and the Mc Curry Exibition (Floor -1) at the reduced price of Euro 4. Towers and prisons can be visited for free. Just show up your badge. They can also visit the Nazario Sauro submarine at the price of Euro 5.

Venue Maps

The conference is held at Floor 1 (Piano Nobile) and Floor -1 (Munizioniere) of Palazzo Ducale. Lunches will be served at Floor -2 (Le Cisterne del Ducale).

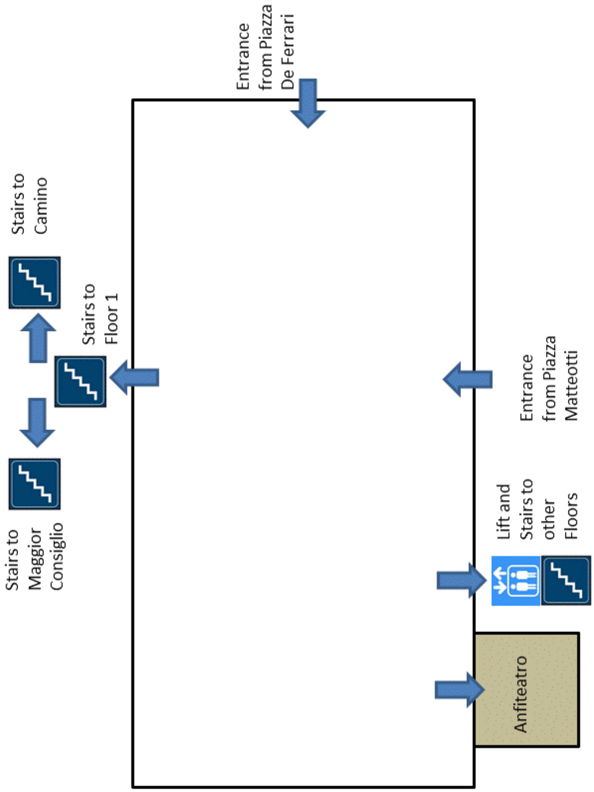
Some additional rooms at Floor 0 and Floor -2 will be used for workshops only on Friday 22.

Room Summary

The rooms used during the conference are summarized in the following table.

Room Name	Floor	Used for	Used on
Maggior Consiglio	1	Sessions	Tue, Wed, Thu
Camino	1	Sessions PAIS	Tue, Wed, Thu Fri
Liguria	1	Sessions LWDM	Tue, Wed Fri
Munizioniere 0	-1	Sessions	Mon, Tue, Wed, Thu
Munizioniere 1	-1	Demos GraphQ	Tue, Wed Fri
Munizioniere 2	- 1	Demos BIGProv	Tue, Wed Fri
Munizioniere 3	-1	Posters PhD Workshop	Wed Fri
Anfiteatro	0	EnDM	Fri
Cisterne	-2	SSSS/J	Fri

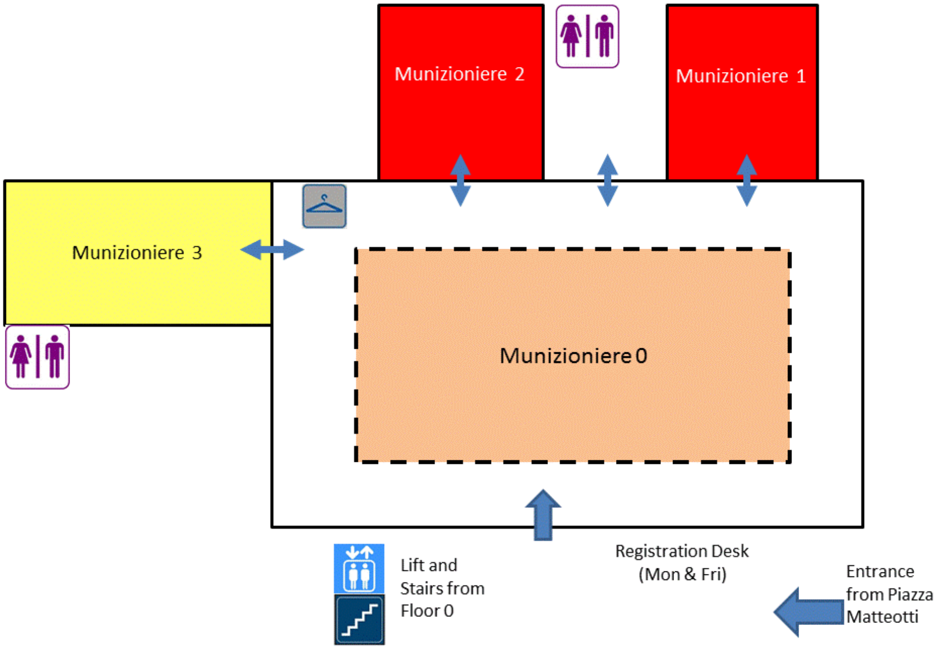
Floor Maps



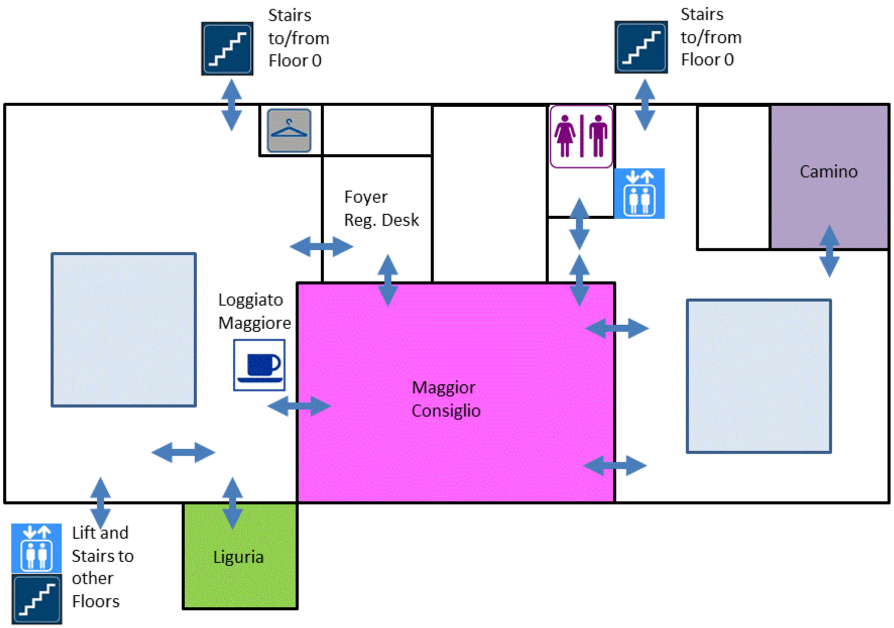
Entrance from Piazza De Ferrari
Stairs to Camino
Stairs to Floor 1
Stairs to Maggior Consiglio
Entrance from Piazza Matteotti
Lift and Stairs to other Floors
Anfiteatro

Floor 0

Floor -1


Munizioniere 2
Munizioniere 1
Munizioniere 3
Munizioniere 0
Lift and
Stairs from
Floor 0
Registration Desk
(Mon & Fri)
Entrance
from Piazza
Matteotti

Floor 1


Stairs
to/from
Floor 0
Stairs
to/from
Floor 0
Camino
Foyer
Reg. Desk
Loggiato
Maggiore
Maggior
Consiglio
Lift and
Stairs to
other
Floors
Liguria

Social Events

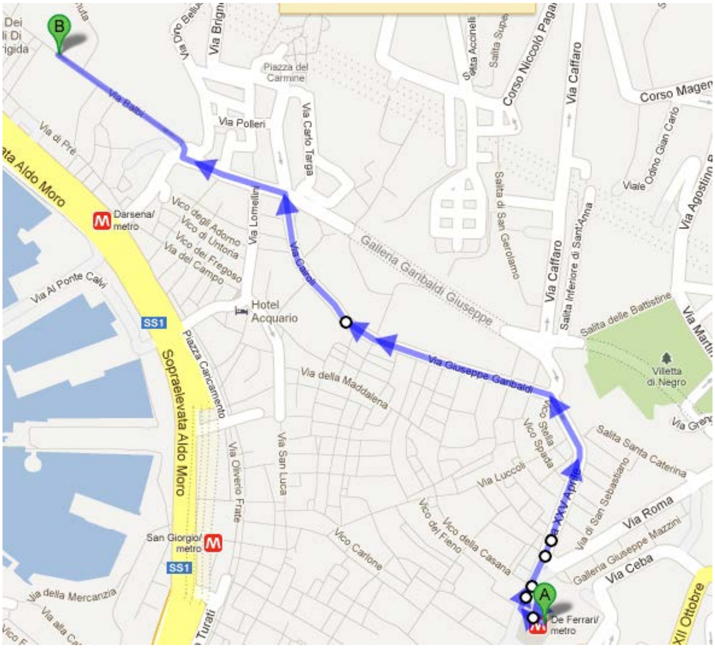
Important: Please always wear your badge since it works as a ticket for all the provided social events.

Tuesday 19th March, 19:00: Welcome Party

It will be served at the Aula Magna of the University Rectorate in via Balbi, where some University buildings are part of the Rolli system, a set of mid 17-th century palaces which are included in the UNESCO world heritage list. Snacks and drinks will be served.

The University Rectorate can be reached from the conference site by foot (10-minutes walk). It is a very nice walk through the main streets of Genova. Alternatively, you may take buses number 20, 30, or 35 in Piazza de Ferrari (Carlo Felice side) and get off at the Via Balbi-Università bus stop.

- **Address:** Via Balbi, 5 - 16126 Genova
- **Meeting Point:** Palazzo Ducale entrance on Piazza de Ferrari, 18:30


B
A
Via Balbi
Via di Prè
Via Polleri
Piazza del Carmine
Via Carlo Targa
Via Brignole
Via Lomellini
Via Cairoli
Vico degli Adorno
Vico di Untoria
Vico dei Fregoso
Via del Campo
Darsena/ metro
Via Al Ponte Calvi
Hotel Acquario
SS1
Sopraelevata Aldo Moro
Piazza Caricamento
Via della Maddalena
Galleria Garibaldi Giuseppe
Via Giuseppe Garibaldi
Salita di San Gerolamo
Corso Niccolò Paganini
Via Caffaro
Salita Inferiore di Sant'Anna
Corso Magenta
Viale Odino Gian Carlo
Via Agostino
Salita delle Battistine
Villetta di Negro
Vico Stella
Vico Spada
Via Luccoli
Via XXV Aprile
Salita Santa Caterina
Via di San Sebastiano
Via Roma
Galleria Giuseppe Mazzini
Via Ceba
XII Ottobre
Vico del Fieno
Vico della Casana
Vico Carlone
Via San Luca
Via Oliverio Frate
San Giorgio/ metro
Via della Mercanzia
Turati
De Ferrari/ metro

Wednesday 20th March, 18:30: Visit to Galata Maritime Museum

Visit to the Galata Museo del Mare (MUMA), the largest Maritime Museum in Italy, standing for the quality and innovation of its reconstructions. The Galata "impresses" through its ability to immerse visitors in accurate reconstructions of great scientific and formal quality, aimed at achieving the best representation of an era, typology of ship and the kind of on board life.

The visit to the museum is included in the conference registration and is not guided.

The Museum can be reached from the conference site by foot (15-minutes walk). It is a nice walk through via San Lorenzo, running along Genoa's Cathedral, and then the old port area, with the Bigo and the Aquarium. Alternatively, you may take the metro at Piazza De Ferrari and get off at Darsena.

The museum route progresses through twenty-three large rooms which are spread over four floors (Ground floor, Floor 1, Floor 2, Floor 3), plus the panoramic Mirador terrace, and are dedicated to the permanent exhibition and themed sections.

■ **Address:** Calata de Mari, 1 - 16126 Genoa

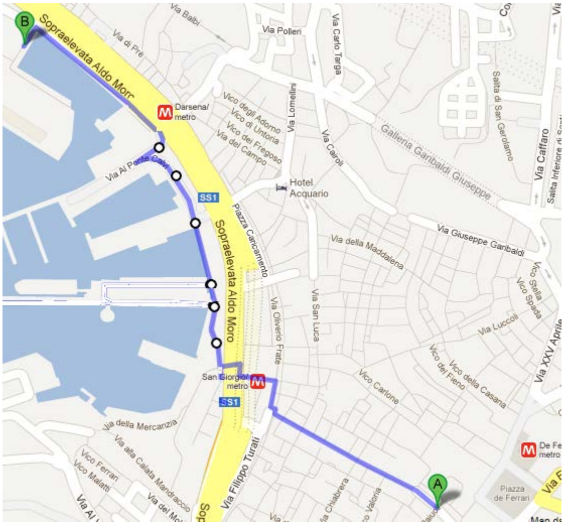
■ **Meeting Point:** Palazzo Ducale entrance on Piazza de Ferrari, 18:00

Wednesday 20th March, from 19:30: Social Dinner at the Galata Maritime Museum

The social dinner will be served inside the museum, at Floor 4, with a superb view of the old port area.

The aperitif will be served starting from 19:30, in Terrazza da Verrazzano. The dinner will be served from 20:30 in Sala Vespucci.

After dinner, an open bar and a musical entertainment by the Loneliest Vocal Ensemble will be provided in Terrazza da Verrazzano.



Wednesday 20th March, 12:30: SAP-sponsored lunch for students

Master and PhD students are invited by SAP to join a seated lunch at Museo Diocesano - Chiostro di San Lorenzo, a wonderful location just outside Palazzo Ducale.

■ **Address:** Via Tommaso Reggio, 20 r - 16123 Genova

■ **Meeting Point:** Just outside Le Cisterne, 12:30

Practical Information

Money exchange, currency

Euro is the official currency in Italy. Money exchange is available in most of banks and financial institutions. There are plenty of cash dispensers in Genova. Major international credit cards are widely but not always accepted in Italy (check in advance in small restaurants), and are not commonly used for small amounts.

Banks are open from Monday to Friday until 1.00/1.30 pm, with flexible opening hours in the afternoon.

Electricity supply

Electricity in Italy, as in the rest of Europe, comes out of the wall socket at 220 volts alternating at a 50 cycles per second. The sockets are for plugs with two or three round pins in a row (CEI 23-16/VII and CEE 7/4 German style Schuko).

Tipping

Service is usually included in restaurants, hotels and taxi.

Public Transport

Genova has a fully developed public transport network, consisting of buses, underground (“metro”) and rail (FFSS).

Metro. The metro is quite short with the following stops:

Brin ↔ Dinegro ↔ Principe (train station) ↔ Darsena, San Giorgio/Piazza Caricamento ↔ Sarzano/Sant'Agostino ↔ Piazza De Ferrari ↔ Brignole (train station).

You can assume that the buses and the Metro are running without a predefined schedule in the center of town during the day, but they are less frequently during the evening, and on weekends. The metro stops to run at 21.00 in the evening.

The price of the ticket is 1.50 Euro, each ticket must be stamped on the bus at the beginning of the journey, and then for 100 minutes you can travel on any bus, on the metro and on urban trains. You can buy a booklet of 10 tickets (called “carnet”) for 14 Euro. Several people traveling together, if they have the cover of the booklet, may also use tickets of the same carnet. Tickets may be bought from tobacconist shops and newspaper agents.

You can plan your itinerary in town using the public transportations (buses, metro and local trains) through the trip planner available at the address:

http://www.amt.genova.it/tripmanager/

Trains. Genova has two main train stations: Principe and Brignole, serving both long-range trains and local trains. The urban trains connect the suburbs of Genova along the sea (from Nervi at East to Voltri at West), and the two Rivieras (East and West). Long-range trains connect Genova with Rome and the south of Italy, with Milan, Venice and Munich, Turin and Paris.

Available trains between two localities in Italy can be found using the application available at the address:

https://www.trenitalia.com

Taxis. In Genova some taxis are waiting for a ride in various places. You can find a taxi near the train stations and in Piazza De Ferrari. You can also call a taxi by phone using this number: +39 010 5966

The price is given by what the taximeter is showing plus a fixed amount (5 Euro) and some extras (e.g., for the night, on Sunday, number of people, number of luggage).

Parking Facilities

Parking near the conference sites is possible only at payment parking places in *Piazza Piccapietra* and *Piazza Dante*.

Cristoforo Colombo Airport Shuttle

VOLABUS is the shuttle bus service departing from Cristoforo Colombo airport to Brignole and Principe Train Stations, crossing the city center, in 30 minutes at the cost of 6 euro.

TIMETABLE
BRIGNOLE RAILWAY STATION - AIRPORT TERMINAL
05.20 - 05.50 - 06.30 - 07.10 - 08.10 - 09.10 - 09.55 - 10.50 - 11.10 - 11.40 - 12.25 - 12.50 - 13.15 - 14.00 - 14.45 - 15.30 - 16.20 - 17.10 - 17.35 - 18.15 - 19.00 - 19.55 - 20.40 - 21.15 - 22.10
AIRPORT TERMINAL - BRIGNOLE RAILWAY STATION
06.00 - 06.30 - 07.15 - 08.15 - 09.00 - 10.05 - 10.45 - 11.35 - 12.05 - 12.30 - 13.15 - 13.40 - 14.00 - 14.45 - 15.30 - 16.25 - 17.25 - 18.10 - 18.45 - 19.20 - 19.55 - 20.35 - 21.35 - 22.10 - 23.30

Closing Days

On Monday most Museums, some attractions in the Old Port, the majority of restaurants and hair salons are closed. Some other shops may be closed on Monday, especially in the morning. Food shops close on Wednesday afternoons, except for supermarkets.

Shops' Opening Times

Many shops still close from 12.30 pm to 3.30 pm, however, in the city center many of them and all large stores stay open all-day.

International calls

Dial 00 + country code + area code + phone number.

The international code number from abroad is +39 followed by the number code of the person you are calling, comprehensive of the "zero" of region.

Emergency calls

In case of any emergency, you can call: 118 if you require an ambulance, 113 if you require the police, 115 if you require the fire brigade. Alternatively, you can dial 112 to reach the "carabinieri". These are special emergency numbers that can be called from any mobile or fixed-line phone.

Smoking

Smoking is not allowed within the conference venue, you can smoke outside of the main building. Furthermore, public indoor establishments in Italy - including train stations, restaurants - ban smoking.

Eating in Genova

Suggestions offered by "La Cuciniera Moderna" (food blog), all the places are in the center of Genova at a walking distance from the conference site, but Maniman that offers a nice view over Genova from its garden.

Restaurants

- **Antica Hostaria Pacetti** (traditional food)
  Via Borgo Incrociati 22r, Genova
  +39 010 8392848
  info@ostariapacetti.com /
  www.ostariapacetti.com
- **Antica Osteria di Vico Palla** (fish and local dishes)
  Vico Palla 15, GENOVA
  +39 010 2466575 / acap29@libero.it
  www.anticaosteriavicopalla.com
- **Antica Vetreria del Molo** (near the old harbor, quite difficult to find. But, it deserves the effort for its hot sandw iches, small selection of traditional dishes, and impressive offer of foreign beers)
  Vico Chiuso della Gelsa - 16128 – Genova
  010 2468700
- **Bicu** (a micro brew ery in the heart of the old harbour offering also food)
  Porto Antico - Magazzini del Cotone - Modulo 4, Genova
  +39 010 2534051 / eventi@bicu.it /
  www.bicu.it/
- **Birrificio Exultate** (an artigianal brew ery, serving their ow n products and foreign beers w ith hot sandw itches or large pizzas)
  Piazza Lavagna - 16123 – Genova
  010 2468724
- **Braxe** (resturant for meat lovers)
  Corso A. Podesta 51r, Genova
  + 39 010 098 55 18
  www.braxe.it / info@braxe.it
- **Eataly Genova** (in the heart of the old harbour the Genova branch of Eataly)
  Edificio Millo Porto Antico, Calata Cattaneo 15, Genova
  +39 010 8698721 /
  www.genova.eataly.it
- **Osteria La Forchetta Curiosa** (Imaginative food from Italian and French tradition. Don't miss a w alking nearby and you w ill see many monuments from the mediaeval time)
  Piazza Renato Negri - 16123 – Genova
  010 2511289
- **Le mani in pasta** (Not only pasta, but so many different kinds of pasta that they w ould be enough)
  Via del Molo, 45 - 16128 – Genova
  010 255646, 349 5356064
- **Ombre Rosse** (Traditional food w ith a tw ist and some dish inspired to Italian regional cuisines)
  Vico degli Indoratori, 22 - 16123 – Genova
  010 2757608
- **Taverna di Colombo** (Near via Garibaldi, traditional food at a reasonable price)
  Vico della Scienza - 16123 – Genova
  010 2462447
- **Trattoria Rina** (A traditional and expensive restaurant in the old tow n.

Particularly renown for the freshness of its fish and the perfect execution of regional recipes)
Mura delle Grazie, 3 - 16128 – Genova
010 2466475

- **I Tre Merli Antica Cantina** (it is an “enoteca”, a place specialized in wine, thus a lot of different good wines are available)
  Vico Dietro il Coro della Maddalena, 26 Genova
  +39 010 2474095
  anticacantina@itremerli.it / www.itremerli.it
- **I Tre Merli Ristorante Porto Antico** (wonderful location in the heart of the old harbour)
  Edificio Millo, Calata Cattaneo 17 (Area Porto Antico di Genova) Genova
  +39 010 2464416
  portoantico@itremerli.it / www.itremerli.it
- **Le Cantine Squarciafico** (in the old town)
  Piazza Invrea 3r, 16123 Genova
  +39 010 2470823
  squarciafico@libero.it / squarciafico.it
- **Enoteca Migone** (Attached to an *enoteca*, a prestigious wine shop, serves traditional dishes associated with an interesting selection of wines)
  Piazza San Matteo - 16123 – Genova
  010 2473282
- **Maniman** (with a nice view over Genova, you will need transportation to get there, as walking from Hotel Bristol would take about half an hour and half of the distance is quite steep)
  Salita San Nicolò 35, Genova
  +39 010211438
- **Maxela Genova 1** (specialized in meat, *Maxela* in local dialect means butcher)
  Vico Inferiore del Ferro 9, Genova
  +39 010 2474209 /
  genova@maxela.it
  www.maxela.it/ristorante_genova.htm
- **Osteria della Piazza** (near via XX Settembre, this place offers traditional food and excellent pizza)
  Piazza Colombo, Genova, Italia
  +390105760308
  http://www.osteriadellapiazzagenova.com/
- **Osteria Sopra Il Mare** (small place with typical food). Vico Cicala, 5, Genova
  +39 010 275 8188
- **Sapori di Genova** (Near Palazzo Ducale. Tradi-tional food from trofie with pesto, to tipical fried food, from stuffed pasta to vegetable pies…)
  Salita Pollaiuoli, 17 - 16123 – Genova
  010 4037622
- **Sa Pesta** (traditional food)
  Via Dei Giustiniani, 16R, Genova
  +39 010 2468336
  info@sapesta.it / www.sapesta.it
- **Trattoria Gianna** (fish and local dishes)
  Vico delle Camelie 26, 16128, Genova
  +39 010 2468659
  info@trattoriadagianna.com
  www.trattoriadagianna.com
- **Trattoria Rosmarino** (small restaurant with creative dishes based on local products)
  Salita del Fondaco 30, 16100 Genova
  +39 010 2510475 /
  www.trattoriarosmarino.it
  info@trattoriarosmarino.it

Pizza parlors (pizzerie)

- **Osteria della Piazza** (near via XX Settembre, this place offers both traditional food and pizza)
  Piazza Colombo, Genova, Italia
  +390105760308
  http://www.osteriadellapiazzagenova.com/
- **Regina Margherita** (pizza Napoli style)
  Piazza della Vittoria 89/103, Genova
  +39 010 5955753
  info@regina-margherita.com
  www.regina-margherita.it
- **Sciuscia & Sciorbi** (pizza with thin crust)
  Via 25 Aprile, 32 Rosso, Genova
  +39 348 83 91 924 (mobile)
  info@sciusciaesciorbi.it
  www.sciusciaesciorbi.it/
- **Zena Zuena** (pizza, farinata, focaccia and local food in an informal setting, also take away)
  There are two locations of this resturant
  Via Cesarea 84-86, Genova
  +39 010 530199

Fast Food

In Genova there are "ancient" fast food places (they were around from the nineteenth century) specialized in pies made with many different vegetables, and o lot of fried food. Usually they serve food to take away but there may be a few seats for immediate consumption. They are named "farinotti". A couple good addresses for places of this kind are the following

- **Antica Sciamadda**
  Via San Giorgio 14, Genova
  +39 010 246 8516
- **Antica Friggitoria Carega**
  Via Sottoripa, 113r, Genova, Italia
  +39 010 2470617
- **Ostaja San Vincenzo** (near Brignole station, cheap and traditional. Try *farinata* as appetizer or as full meal)
  Via San Vincenzo 64r, 16121 Genova, Italia
  010.565765

*If you are looking for a sandwich we suggest two places offering a huge choices of sandwiches with a very large variety of **filling:***

- **Panino d'autore**
  Via XX Settembre, 68 r, 16121 Genova,
  +39 010 561625
- **Gran Ristoro**
  Via Sottoripa, 27 r, Genova
  0102473127

Drinking Places

Cocktails in Genova are in most place served with many samples of food, from *focaccia*, to vegetable pies, to a bit of pasta, some cured meat, fresh vegetables and sauces…amount and variety depends on the fantasy of the place owner. They may cost up to 10 euro each, but are in most cases a good solution for a light dinner.

- **Banano Tsunami** (with a large terrace on the sea in the old harbour)
  Piazza Delle Feste in the old harbour, Genova
  flavors.me/bananotsunami
- **Il Barbarossa**
  Piano di Sant'Andrea, 21/23 r. - 16123 - Genova
  010 2465097
- **Caffé degli Specchi** (old cafè with the original furniture)
  Salita Pollaiuoli 43, Genova
  +39 0 10 246 8193
- **Capitan Baliano** (near Palazzo Ducale)
  Piazza Giacomo Matteotti 11r, Genova
  +39 010 265299
- **Le Corbusier**
  Via di San Donato, 36 - 16123 - Genova
  328 6967446, 349 5783405
- **Taggiou** (In dialect: chopping table. A tiny place built in an old *salumeria* (cured meat and cheese shop). It is specialized in wines served with assortments of cold food)
  Vico Superiore del Ferro, 8 - 16123 – Genova
  010 2759225

3. Program

Program at a Glance

Monday, March 18, 2013	
8:30-	**Registration** [Desk at Floor – 1]
Room	**Munizioniere0**
8:45-9:00	**ICDT Opening Session**
9:00-10:30	**ICDT Keynote 1: Jan van den Bussche**
10:30-11:00	***Coffee Break*** *[Cisterne]*
11:00-12:30	**ICDT Research 1**
12:30-14:00	***Lunch Break*** *[Cisterne]*
14:00-15:30	**ICDT Research 2**
15:30-16:00	***Coffee Break*** *[Cisterne]*
16:00-17:30	**ICDT Research 3**

Tuesday, March 19, 2013

8:30-	**Registration** **[Desk at Floor 1]**				
Rooms	**Maggior Consiglio**	**Munizioniere0**	**Camino**	**Liguria**	**Munizioniere1 +** **Munizioniere2**
9:00-	**EDBT Opening Session** [Maggior Consiglio]				
-10:30	**EDBT Keynote 1: Daniel Abadi** [Maggior Consiglio]				
10:30-11:00	***Coffee Break*** *[Loggiato Maggiore]*				
11:00-12:30	**EDBT Research 1**	**ICDT Research 4**	**EDBT Research 2**	**Tutorial 1**	
12:30-14:00	***Lunch Break*** *[Cisterne]*				
14:00-15:30	**EDBT Research 3**	**ICDT Research 5**	**EDBT Research 4**		**Demo Session 1-1**
15:30-16:00	***Coffee Break*** *[Loggiato Maggiore]*				
16:00- 17:30/18.00	**EDBT Research 5**	**ICDT Research 6**	**EDBT Research 6**		**Demo Session 2-1**
19:00-	**Welcome Reception**				

Wednesday, March 20, 2013

	Registration					
Rooms	**Maggior Consiglio**	**Munizioniere0**	**Camino**	**Liguria**	**Munizioniere1 + Munizioniere2**	**Munizioniere3**
9:00-10:30	**ICDT Keynote 2: Luc Segoufin** [Maggior Consiglio]					
10:30-11:00	***Coffee Break*** *[Loggiato Maggiore]*					
11:00-12:30	**EDBT Research 7**	**Industry & Applications 1**	**ICDT Research 7**	**Tutorial 2**		
12:30-14:00	***Lunch Break*** *[Cisterne]* **SAP-Sponsored Lunch for Students**					
14:00-15:30	**EDBT Research 8**	**Industry & Applications 2**	**ICDT Research 8**		**Demo Session 1-2**	Poster Session EDBT + PhD Workshop
15:30-16:00	***Coffee Break*** *[Loggiato Maggiore]*					
16:00-17:30	**EDBT Research 9**	**Industry & Applications 3**	**ICDT Research 9**		**Demo Session 2-2**	
18:30-	**Visit to the Galata Museum**					
19:30-	**Banquet**					

Thursday, March 21, 2013

	Registration		
Rooms	**Maggior Consiglio**	**Munizioniere0**	**Camino**
9:00-10:30	**EDBT Keynote 2: C. Mohan [Maggior Consiglio]**		
10:30-11:00	***Coffee Break*** *[Loggiato Maggiore]*		
11:00-12:30	**EDBT Research 10**	**Tutorial 3**	**Industry & Applications 4**
12:30-14:00	***Lunch Break*** *[Cisterne]*		
14:00-15:30	**EDBT Research 11**	**Tutorial 3**	**EDBT Research 12**
15:30-16:00	***Coffee Break*** *[Loggiato Maggiore]*		
16:00-17:30	**EDBT Research 13**	**EDBT Research 15**	**EDBT Research 14**

Friday, March 22, 2013

Rooms	Munizioniere 1	Munizioniere 2	Munizioniere 3	Camino	Liguria	Cisterne	Anfiteatro
Workshops	GraphQ	BigProv	PhD	PAIS	LWDM	SSSS/J	EnDM
8:45-9:00		Opening	Opening	Opening			
9:00-10:30	Opening Keynote	Session 1	Session 1	Keynote Session 1	Opening Flash Session Keynote	Opening and Session 1	Opening Session 1
10:30-11:00	*Coffee Break [Munizioniere]*						
11:00-12:30	Session 1	Session 2	Session 2	Session 2	Session 1	Session 2	Session 2
12:30-14:00	*Lunch Break [Cisterne]*						
14:00-15:30	Session 2	Session 3	Session 3	Panel	Session 2	Session 3 and Closing	
15:30-16:00	*Coffee Break [Munizioniere]*						
16:00-17.30	Closing	Session 4 and Closing	Closing	Session 3	BeWeb Session Panel		
				Closing	Closing		

Detailed Program

Monday, March 18

8:45-9:00	ICDT Opening Session		

9:00-10:30	Keynote		Jan van den Bussche
Room:	Munizioniere0	Chair:	Peter Buneman

The DNA Query Language DNAQL

11:00-12:30	ICDT Research Session 1		**Award Papers**
Room:	Munizioniere0	Chair:	Wang-Chiew Tan

Test-of-Time Award: *"Data Exchange: Semantics and Query Answering"*:
Ronald Fagin, Phokion Kolaitis, Renee Miller, and Lucian Popa

Best Paper Award: *"A Theory of Pricing Private Data"*:
Chao Li, Daniel Li, Gerome Miklau and Dan Suciu

14:00-15:30	ICDT Research Session 2		**Semi-structured Data and XML**
Room:	Munizioniere0	Chair:	Balder ten Cate

Fast Learning of Restricted Regular Expressions and DTDs:
Dominik D. Freydenberger and Timo Kötzing

Which DTDs are streaming bounded repairable?:
Pierre Bourhis, Gabriele Puppis and Cristian Riveros

XML compression via DAGs:
Sebastian Maneth, Markus Lohrey and Eric Noeth

16:00-17:30	ICDT Research Session 3		**Query Processing and Optimization**
Room:	Munizioniere0	Chair:	Frank Neven

Structural Tractability of Counting of Solutions to Conjunctive Queries:
Arnaud Durand and Stefan Mengel

Recursive queries on trees and data trees:
Serge Abiteboul, Pierre Bourhis, Anca Muscholl and Zhilin Wu

On optimum left-to-right strategies for active context-free games:
Henrik Björklund, Martin Schuster, Thomas Schwentick and Joscha Kulbatzki

Tuesday, March 19

9:00-	EDBT Opening Session		

-10:30	Keynote		Daniel Abadi
Room:	Maggior Consiglio	Chair:	Barbara Catania

Invisible Loading: Access-Driven Data Transfer from Raw Files into Database Systems

11:00-12:30	ICDT Research Session 4		**Graph Databases**
Room:	Munizioniere0	Chair:	Gerome Miklau

Walk Logic as a framework for path query languages on graph databases:
Jelle Hellings, Bart Kuijpers, Jan Van Den Bussche and Xiaowang Zhang

Querying Graph Databases with XPath:
Leonid Libkin, Wim Martens and Domagoj Vrgoc

Definability problems for graph query languages:
Timos Antonopoulos, Frank Neven and Frédéric Servais

11:00-12:30	EDBT Research Session 1		**High Performance Query Processing**
Room:	Maggior Consiglio	Chair:	Stefan Manegold

From A to E: Analyzing TPC's OLTP Benchmarks – The obsolete, the ubiquitous, the unexplored:
Pinar Tozun, Ippokratis Pandis, Cansu Kaynak, Djordje Jevdjic and Anastasia Ailamaki

Query-Aware Compression of Join Results:
Christopher Mullins, Lipyeow Lim and Christian Lang

Web Data Indexing in the Cloud: Efficiency and Cost Reductions:
Jesús Camacho-Rodríguez, Dario Colazzo and Ioana Manolescu

11:00-12:30	EDBT Research Session 2		**Multi-tenant Databases**
Room:	Camino	Chair:	Alfredo Cuzzocrea

ProRea - Live Database Migration for Multi-tenant RDBMS with Snapshot Isolation:
Oliver Schiller, Nazario Cipriani and Bernhard Mitschang

SWAT: A Lightweight Load Balancing Method for Multitenant Databases:
Hyun Moon, Hakan Hacigumus, Yun Chi and Wang-Pin Hsiung

CloudOptimizer: Multi-tenancy for I/O-Bound OLAP Workloads:
Hatem Mahmoud, Hyun Moon, Yun Chi, Divyakant Agrawal and Amr El-Abbadi, Hakan Hacigumus

11:00-12:30	Tutorial Session 1		
Room:	Liguria		

Trust and Reputation in and Across Virtual Communities
Nurit Gal-Oz (Ben-Gurion University and Sapir Academic College, Israel)
Ehud Gudes (Ben-Gurion University, Israel)

14:00-15:30	ICDT Research Session 5		**ICDT Invited Lecture**
Room:	Munizioniere0	Chair:	Maurizio Lenzerini

A Personal Perspective on Keyword Search over Data Graphs:
Yehoshua Sagiv (Hebrew University of Jerusalem)

14:00-15:30	EDBT Research Session 3		**MapReduce**
Room:	Maggior Consiglio	Chair:	Ioana Manolescu

Eagle-Eyed Elephant: Split-Oriented Indexing in Hadoop:
Mohamed Eltabakh, Fatma Ozcan, Yannis Sismanis, Peter Haas, Hamid Pirahesh and Jan Vondrak

Computing n-Gram Statistics in MapReduce:
Klaus Berberich and Srikanta Bedathur

Processing Multi-way Spatial Joins On Map-Reduce:
Himanshu Gupta, Bhupesh Chawda, Sumit Negi, Tanveer Faruquie, L. V. Subramaniam and Mukesh Mohania

14:00-15:30	EDBT Research Session 4		**Extending Database Technology**
Room:	Camino	Chair:	Bernhard Mitschang

Rapid Experimentation for Testing and Tuning a Production Database Deployment
Nedyalko Borisov and Shivnath Babu

Towards Context-Aware Search and Analysis on Social Media Data
Leon Derczynski, Bin Yang and Christian Jensen

Proactive Natural Language Search Engine: Tapping into Structured Data on the Web
Wensheng Wu

Anomaly Management using Complex Event Processing
Bastian Hoßbach and Bernhard Seeger

14:00-15:30	EDBT Demo Session 1-1		
Room:	Munizioniere1 + Munizioniere2		

iPark: Identifying Parking Spaces from Trajectories:
Bin Yang, Nicolas Fantini and Christian S. Jensen

Limosa: A System for Geographic User Interest Analysis in Twitter:
Jan Vosecky, Di Jiang and Wilfred Ng

Accelerating Spatial Range Queries:
Alexandros Stougiannis, Thomas Heinis, Farhan Tauheed and Anastasia Ailamaki

An Efficient Layout Method for a Large Collection of Geographic Data Entries:
Sarana Nutanong, Marco Adelfio and Hanan Samet

In the Mood4: Recommendation by Examples:
Rubi Boim and Tova Milo

YmalDB: A Result-Driven Recommendation System for Databases:
Marina Drosou and Evaggelia Pitoura

Hive Open Research Network Platform:
Jung Hyun Kim, Xilun Chen, K. Selcuk Candan and Maria Luisa Sapino

16:00-17:30	ICDT Research Session 6		**Provenance and Annotations**
Room:	Munizioniere0	Chair:	James Cheney

Algebraic Structures for Capturing the Provenance of SPARQL Queries:
Floris Geerts, Grigoris Karvounarakis, Vassilis Christophides and Irini Fundulaki

A Propagation Model for Provenance Views of Public/Private Workflows:
Susan Davidson, Tova Milo and Sudeepa Roy

Annotations are Relative:
Peter Buneman, Egor V. Kostylev and Stijn Vansummeren

16:00-18:00	EDBT Research Session 5		**Privacy**
Room:	Maggior Consiglio	Chair:	Ehud Gudes

Compromising Privacy in Precise Query Protocols:
Jonathan Dautrich and Chinya Ravishankar

Efficient Privacy-Aware Record Integration:
Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham and Bradley Malin

Updating Outsourced Anatomized Private Databases:
Ahmet Erhan Nergiz, Chris Clifton and Qutaibah Malluhi

Efficient and Accurate Strategies for Differentially-Private Sliding Window Queries:
Jianneng Cao, Qian Xiao, Gabriel Ghinita, Ninghui Li, Elisa Bertino, and Kian-Lee Tan

16:00-18:00	EDBT Research Session 6		**Potpourri**
Room:	Camino	Chair:	Ulf Leser

An Automatic Physical Design Tool for Clustered Column-Stores:
Alexander Rasin and Stan Zdonik

Mining Frequent Serial Episodes Over Uncertain Sequence Data:
Li Wan, Ling Chen and Chengqi Zhang

Efficient processing of containment queries on nested sets:
Ahmed Ibrahim and George H. L. Fletcher

Inferential Time-Decaying Bloom Filters:
Jonathan Dautrich and Chinya Ravishankar

16:00-17:30	EDBT Demo Session 2-1		
Room:	Munizioniere1 + Munizioniere2		

CrowdSeed: Query Processing on Microblogs:
Zhou Zhao, Wilfred Ng and Zhijun Zhang

Tuning in Action:
Dennis Shasha and Wei Cao

PostgreSQL Anomalous Query Detector:
Bilal Shebaro, Asmaa Sallam and Elisa Bertino

Processing XML Queries and Updates on Map/Reduce Clusters:
Nicole Bidoit, Dario Colazzo, Noor Malla, Maurizio Nolé, Carlo Sartiani and Federico Ulliana

CISC: Clustered Image Search by Conceptualization:
Kaiqi Zhao, Enxun Wei, Qingyu Sui, Kenny Zhu and Eric Lo

MinExp-Card: Limiting Data Collection Using a Smart Card:
Nicolas Anciaux, Walid Bezza, Benjamin Nguyen and Michalis Vazirgiannis

PrivComp: A Privacy-aware Data Service Composition System:
Mahmoud Barhamgi, Djamal Benslimane and Youssef Amghar

ProQua: A System for Evaluating Logic-Based Scoring Functions on Uncertain Relational Data:
Sebastian Lehrack, Sascha Saretz and Christian Winkel

ProvenanceCurious: A Tool to Infer Data Provenance from Scripts:
Mohammad Rezwanul Huq, Peter M.G. Apers and Andreas Wombacher

Wednesday, March 20

9:00-10:30	Keynote		Luc Segoufin
Room:	Maggior Consiglio	Chair:	Thomas Schwentick

Enumerating with Constant Delay the Answers to a Query

11:00-12:30	ICDT Research Session 7		**Data Exchange and Query Answering**
Room:	Camino	Chair:	Benny Kimelfeld

Schema Mappings and Data Exchange for Graph Databases:
Pablo Barceló, Jorge Pérez and Juan L. Reutter

Containment of Pattern-Based Queries over Data Trees:
Claire David, Amelie Gheerbrant, Leonid Libkin and Wim Martens

Access Patterns and Integrity Constraints Revisited:
Vince Barany, Michael Benedikt and Pierre Bourhis

11:00-12:30	EDBT Research Session 7		**Sensors**
Room:	Maggior Consiglio	Chair:	Qing Zhang

Utility-driven Data Acquisition in Participatory Sensing:
Mehdi Riahi, Thanasis Papaioannou, Karl Aberer and Immanuel Trummer

An RFID and Particle Filter-Based Indoor Spatial Query Evaluation System:
Jiao Yu, Wei-Shinn Ku, Min-Te Sun and Hua Lu

A Safe Zone Based Approach for Monitoring Moving Skyline Queries:
Muhammad Cheema, Xuemin Lin, Wenjie Zhang and Ying Zhang

11:00-12:30	EDBT Industry & Applications Session 1		**Systems and Tools**
Room:	Munizioniere0		

Temporal Query Processing in Teradata:
Mohammed Al-Kateb, Ahmad Ghazal, Alain Crolotte, Ramesh Bhashyam, Jaiprakash Chimanchode and Sai Pavan Pakala

Near Real-Time Analytics with IBM DB2 Analytics Accelerator:
Daniel Martin, Oliver Koeth, Iliyana Ivanova and Johannes Kern

AppSleuth: a Tool for Database Tuning at the Application Level:
Wei Cao and Dennis Shasha

11:00-12:30	Tutorial Session 2		
Room:	Liguria		

The W3C PROV Family of Specifications for Modelling Provenance
Paolo Missier (Newcastle University, UK)
Khalid Belhajjame (University of Manchester, UK)
James Cheney (University of Edinburgh, UK)

14:00-15:30	ICDT Research Session 8		**Ranking Query Answers**
Room:	Camino	Chair:	Michael Benedikt

Using the Crowd for Top-k and Group-by Queries:
Susan Davidson, Sanjeev Khanna, Tova Milo and Sudeepa Roy

Certain and Possible XPath Answers:
Sara Cohen and Yaacov Y. Weiss

Extracting Minimum-Weight Tree Patterns from a Schema with Neighborhood Constraints:
Benny Kimelfeld and Yehoshua Sagiv

14:00-15:30	EDBT Research Session 8		**Graph Querying**
Room:	Maggior Consiglio	Chair:	George Fletcher

Compressed Feature-based Filtering and Verification Approach for Subgraph Search:
Karam Gouda and Mosab Hassaan

Efficient Query Answering against Dynamic RDF Databases:
François Goasdoué, Ioana Manolescu and Alexandra Roatis

Efficient Breadth-First Search on Large Graphs with Skewed Degree Distributions:
Haichuan Shang and Masaru Kitsuregawa

14:00-15:30	EDBT Industry & Applications Session 2		**Big Data**
Room:	Munizioniere0		

Cost Exploration of Data Sharings in the Cloud :
Samer Al-Kiswany, Hakan Hacigumus, Ziyang Liu and Jagan Sankaranarayanan

A Performance Comparison of Parallel DBMSs and MapReduce on Large-Scale Text Analytics:
Fei Chen and Meichun Hsu

Sparkler: Supporting Large-Scale Matrix Factorization:
Sandeep Tata, Yannis Sismannis and Boduo Li

14:00-15:30	EDBT Demo Session 1-2		
Room:	Munizioniere1 + Munizioniere2		

iPark: Identifying Parking Spaces from Trajectories:
Bin Yang, Nicolas Fantini and Christian S. Jensen

Limosa: A System for Geographic User Interest Analysis in Twitter:
Jan Vosecky, Di Jiang and Wilfred Ng

Accelerating Spatial Range Queries:
Alexandros Stougiannis, Thomas Heinis, Farhan Tauheed and Anastasia Ailamaki

An Efficient Layout Method for a Large Collection of Geographic Data Entries:
Sarana Nutanong, Marco Adelfio and Hanan Samet

In the Mood4: Recommendation by Examples:
Rubi Boim and Tova Milo

YmalDB: A Result-Driven Recommendation System for Databases:
Marina Drosou and Evaggelia Pitoura

Hive Open Research Network Platform:
Jung Hyun Kim, Xilun Chen, K. Selcuk Candan and Maria Luisa Sapino

16:00-17:30	ICDT Research Session 9		**Privacy**
Room:	Camino	Chair:	Sebastian Maneth

On Optimal Differentially Private Mechanisms for Count-Range Queries:
Chen Zeng, Jin-Yi Cai, Pinyan Lu and Jeffrey Naughton

Optimal Error of Query Sets under the Differentially-private Matrix Mechanism:
Chao Li and Gerome Miklau

Private Predicate Sums on Decayed Streams:
Jean Bolot, Nadia Fawaz, S. Muthukrishnan, Aleksandar Nikolov, Nina Taft

16:00-17:30	EDBT Research Session 9		**Social Networks and Semantic Querying**
Room:	Maggior Consiglio	Chair:	Klaus Berberich

CINEMA: Conformity-Aware Greedy Algorithm for Influence Maximization in Online Social Networks:
Hui Li, Sourav S Bhowmick, and Aixin Sun

Pollux: Towards Scalable Distributed Real-time Search on Microblogs:
Liwei Lin, Xiaohui Yu and Nick Koudas

Semantic Query By Example:
Lipyeow Lim, Haixun Wang and Min Wang

16:00-17:30	EDBT Industry & Applications Session 3		**Applications**
Room:	Munizioniere0		

Choosing the Right Crowd: Expert Finding in Social Networks:
Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, Giuliano Vesci

Real-Time Wildfire Monitoring Using Scientific Database and Linked Data Technologies:
Manolis Koubarakis, Charalambos Kontoes, Stefan Manegold, Manos Karpathiotakis, Kostis Kyzirakos, Konstantina Bereta, George Garbis, Charalampos Nikolaou, Dimitrios Michail, Ioannis Papoutsis, Themistoklis Herekakis, Milena Ivanova, Ying Zhang, Holger Pirk, Martin Kersten, Kallirroi Dogani, Stella Giannakopoulou and Panayiotis Smeros

Efficient Multifaceted Screening of Job Applicants:
Sameep Mehta, Rakesh Pimplikar, Amit Singh, Lav Varshney, Karthik Viswesariah

16:00-17:30	EDBT Demo Session 2-2		
Room:	Munizioniere1 + Munizioniere2		

CrowdSeed: Query Processing on Microblogs:
Zhou Zhao, Wilfred Ng and Zhijun Zhang

Tuning in Action:
Dennis Shasha and Wei Cao

PostgreSQL Anomalous Query Detector:
Bilal Shebaro, Asmaa Sallam and Elisa Bertino

Processing XML Queries and Updates on Map/Reduce Clusters:
Nicole Bidoit, Dario Colazzo, Noor Malla, Maurizio Nolé, Carlo Sartiani and Federico Ulliana

CISC: Clustered Image Search by Conceptualization:
Kaiqi Zhao, Enxun Wei, Qingyu Sui, Kenny Zhu and Eric Lo

MinExp-Card: Limiting Data Collection Using a Smart Card:
Nicolas Anciaux, Walid Bezza, Benjamin Nguyen and Michalis Vazirgiannis

PrivComp: A Privacy-aware Data Service Composition System:
Mahmoud Barhamgi, Djamal Benslimane and Youssef Amghar

ProQua: A System for Evaluating Logic-Based Scoring Functions on Uncertain Relational Data:
Sebastian Lehrack, Sascha Saretz and Christian Winkel

ProvenanceCurious: A Tool to Infer Data Provenance from Scripts:
Mohammad Rezwanul Huq, Peter M.G. Apers and Andreas Wombacher

Thursday, March 21

9:00-10:30	Keynote		C. Mohan
Room:	Maggior Consiglio	Chair:	Norman Paton

History Repeat Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla

11:00-12:30	EDBT Research Session 10		**Search and Textual Data**
Room:	Maggior Consiglio	Chair:	Ziyang Liu

Scalable Top-K Spatial Keyword Search:
Dongxiang Zhang, Kian-Lee Tan and Anthony K. H. Tung

Panorama: A Semantic-Aware Application Search Framework:
Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung and Wilfred Ng

Selectivity Estimation for Hybrid Queries over Text-Rich Databases:
Andreas Wagner, Veli Bicer and Thanh Tran

11:00-12:30	EDBT Industry & Applications Session 4		**Potpourri**
Room:	Camino		

EXLEngine: executable schema mappings for statistical data processing :
Paolo Atzeni, Bellomarini Luigi and Bugiotti Francesca

HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm:
Stefan Heule, Marc Nunkesser and Alexander Hall

Entity Discovery and Annotation in Tables:
Gianluca Quercini and Chantal Reynaud

11:00-12:30	Tutorial Session 3		
Room:	Munizioniere0		

Schema Mapping and Data Examples
Balder ten Cate (UC Santa Cruz)
Phokion G. Kolaitis (UC Santa Cruz and IBM Research - Almaden)
Wang-Chiew Tan (UC Santa Cruz)

14:00-15:30	EDBT Research Session 11		**Skyline**
Room:	Maggior Consiglio	Chair:	Matteo Magnani

Skyline Probability over Uncertain Preferences:
Qing Zhang, Pengjie Ye, Xuemin Lin and Ying Zhang

SkyDiver: A Framework for Skyline Diversification:
George Valkanas, Apostolos N. Papadopoulos and Dimitrios Gunopulos

Subspace Global Skyline Query Processing:
Mei Bai, Junchang Xin and Guoren Wang

14:00-15:30	EDBT Research Session 12		**Database as a Service**
Room:	Camino	Chair:	Murat Kantarcioglu

SWORD: Scalable Workload-Aware Data Placement for Transactional Workloads:
Abdul Quamar, K.Ashwin Kumar and Amol Deshpande

PMAX: Tenant Placement in Multitenant Databases for Profit Maximization:
Ziyang Liu ,Hakan Hacigumus, Hyun Moon, Yun Chi and Wang-Pin Hsiung

Elastic Online Analytical Processing on RAMCloud:
Christian Tinnefeld, Donald Kossmann, Martin Grund, Joos-Hendrik Boese, Frank Renkes, Vishal Sikka and Hasso Plattner

14:00-15:30	Tutorial Session 3		
Room:	Munizioniere0		

Schema Mapping and Data Examples
Balder ten Cate (UC Santa Cruz)
Phokion G. Kolaitis (UC Santa Cruz and IBM Research - Almaden)
Wang-Chiew Tan (UC Santa Cruz)

16:00-17:30	EDBT Research Session 13		**Preference Queries**
Room:	Maggior Consiglio	Chair:	Mohamed Y. Eltabakh

Skyline Queries in Crowd-Enabled Databases:
Christoph Lofi, Kinda El Maarry and Wolf-Tilo Balke

From stars to galaxies: skyline queries on aggregate data :
Matteo Magnani and Ira Assent

Efficient Top-k Query Answering using Cached Views:
Min Xie, Laks Lakshmanan and Peter Wood

16:00-17:30	EDBT Research Session 14		**Stream Query Processing**
Room:	Camino	Chair:	Torben Bach Pedersen

Enhanced Stream Processing in a DBMS Kernel:
Erietta Liarou, Stratos Idreos, Stefan Manegold and Martin Kersten

Probabilistic Inference of Object Identifications for Event Stream Analytics:
Di Wang, Elke Rundensteiner, Han Wang and Richard Ellison

High Performance Complex Event Processing using Continuous Sliding Views:
Medhabi Ray, Elke Rundensteiner, Mo Liu, Chetan Gupta, Song Wang and Ismail Ari

16:00-17:30	EDBT Research Session 15		**Data Integration**
Room:	Munizioniere0	Chair:	Melanie Herschel

Data Exchange with Arithmetic Operations:
BalderTen Cate, Phokion Kolaitis and Walied Othman

HIL: A High-Level Scripting Language for Entity Integration:
Mauricio Hernandez, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa and Ryan Wisnesky

Optimizing Query Rewriting in Ontology-Based Data Access:
Floriana Di Pinto, Domenico Lembo, Maurizio Lenzerini, Riccardo Mancini, Antonella Poggi, Riccardo Rosati, Marco Ruzzi and Domenico Fabio Savo

4. Keynotes

Date: Monday, March 18

Time: 9:00-10:30

Room:
Munizioniere0

Speaker:
Jan van den Bussche

Title:
The DNA query language DNAQL

Abstract:

This talk presents an overview of our work on databases in DNA performed over the past four years, joint with my student Joris Gillis and postdoc Robert Brijder. Our goal is to better understand, at a theoretical level, the database aspects of DNA computing. The talk will be self-contained and will begin with an introduction to DNA computing. We then introduce a graph-based data model of so-called sticker DNA complexes, suitable for the representation and manipulation of structured data in DNA. We also define DNAQL, a restricted programming language over sticker DNA complexes. DNAQL stands to general DNA computing as the standard relational algebra for relational databases stands to general-purpose conventional computing. We show how DNA program can be statically typechecked. Thus, nonterminating reactions, as well as other things that could go wrong during DNA manipulation, can be avoided. We also investigate the expressive power of DNAQL and show how it compares to the relational algebra.

Date: Tuesday, March 19

Time: 9:00-10:30

Room:
Maggior Consiglio

Speaker:
Daniel Abadi

Title:
Invisible Loading: Access-Driven Data Transfer from Raw Files into Database Systems

Abstract:

HadoopDB began as a research effort in 2008 to transform Hadoop --- a batch-oriented scalable system designed for processing unstructured data --- into a full-fledged parallel database system that can achieve real-time (interactive) query responses across both structured and unstructured data. In 2010 it was commercialized by Hadapt, a start-up that was formed to accelerate the engineering of the HadoopDB ideas, and to harden the codebase for deployment in real-world, mission-critical applications. In this talk I will give an overview of HadoopDB, and how it combines ideas from the Hadoop and database system communities. I will then describe some research challenges that have emerged as HadoopDB increasingly gets deployed in the real world. Many of these challenges involve loading data into structured storage. Although this loading of data can greatly accelerate query execution times, the upfront cost of this load is antithetical to the Hadoop premise that data need not be organized, cleaned, and pre-processed before being available for query processing. Therefore, we will discuss two approaches to reducing these costs: (1) an invisible loading technique where data is incrementally loaded into structured storage over time, based on users' patterns of data access and (2) a queue-based locality scheduling technique that, when data had been loaded in a heterogeneous manner across the nodes in a cluster, improves upon Hadoop's greedy scheduler and more efficiently assigns tasks to nodes that have the data stored locally.

Authors: Azza Abouzied, Daniel J. Abadi, Avi Silberschatz

Date: Wednesday, March 20	**Time:** 9:00-10:30
Room: Maggior Consiglio	**Speaker:** Luc Segoufin

Title:
Enumerating with constant delay the answers to a query

Abstract:

We survey recent results about enumerating with constant delay the answers to a query over a database. More precisely, we focus on the case when enumeration can be achieved with a preprocessing running in time linear in the size of the database, followed by an enumeration process outputting the answers one by one with constant time between any consecutive outputs. We survey classes of databases and classes of queries for which this is possible. We also mention related problems such as computing the number of answers or sampling the set of answers.

Date: Thursday, March 21 **Time:** 9:00-10:30

Room:
Maggior Consiglio

Speaker:
C. Mohan

Title:
History Repeats Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla

Abstract:

In this talk, I describe some of the recent developments in the database management area, in particular the NoSQL phenomenon and the hoopla associated with it. The goal is not to do an exhaustive survey of NoSQL systems. The aim is to do a broad brush analysis of what these developments mean - the good and the bad aspects! Based on my more than three decades of database systems work in the research and product arenas, I will outline what are many of the pitfalls to avoid since there is currently a mad rush to develop and adopt a plethora of NoSQL systems in a segment of the IT population, including the research community. In rushing to develop these systems to overcome some of the shortcomings of the relational systems, many good principles of the latter, which go beyond the relational model and the SQL language, have been left by the wayside. Now many of the features that were initially discarded as unnecessary in the NoSQL systems are being brought in, but unfortunately in ad hoc ways. Hopefully, the lessons learnt over three decades with relational and other systems would not go to waste and we wouldn't let history repeat itself with respect to simple minded approaches leading to enormous pain later on for developers as well as users of the NoSQL systems!

ICDT Invited Lecture

Date: Tuesday, March 19 **Time:** 14:00-15:30

Room:
Munizioniere0

Speaker:
Yehoshua Sagiv

Title:
A Personal Perspective on Keyword Search over Data Graphs

Abstract:

Theoretical and practical issues pertaining to keyword search over data graphs are discussed. A formal model and algorithms for enumerating answers (by operating directly on the data graph) are described. Various aspects of a system are explained, including the object-connector-property data model, how it is used to construct a data graph from an XML document, how to deal with redundancies in the source data, what are duplicate answers, implementation and GUI. An approach to ranking that combines textual relevance with semantic considerations is described. It is argued that search over data graphs is inherently a two-dimensional process, where the goal is not just to ?nd particular content but also to collect information on how the desired data may be semantically connected.

5. Abstracts

ICDT Research Sessions

Monday, March 18			
11:00-12:30	ICDT Research Session 1		**Award Papers**
Room:	Munizioniere0	Chair:	Wang-Chiew Tan

Test-of-Time Award: Data Exchange: Semantics and Query Answering

Ronald Fagin, Phokion Kolaitis, Renee Miller, and Lucian Popa

Data exchange is the problem of taking data structured under a source schema and creating an instance of a target schema that reflects the source data as accurately as possible. In this paper, we address foundational and algorithmic issues related to the semantics of data exchange and to query answering in the context of data exchange. These issues arise because, given a source instance, there may be many target instances that satisfy the constraints of the data exchange problem. We give an algebraic specification that selects, among all solutions to the data exchange problem, a special class of solutions that we call *universal*. A universal solution has no more and no less data than required for data exchange and it represents the entire space of possible solutions. We then identify fairly general, and practical, conditions that guarantee the existence of a universal solution and yield algorithms to compute a canonical universal solution efficiently. We adopt the notion of "certain answers" in indefinite databases for the semantics for query answering in data exchange. We investigate the computational complexity of computing the certain answers in this context and also study the problem of computing the certain answers of target queries by simply evaluating them on a canonical universal solution.

Best Paper Award: A Theory of Pricing Private Data

Chao Li, Daniel Li, Gerome Miklau and Dan Suciu

Personal data has value to both its owner and to institutions who would like to analyze it. Privacy mechanisms protect the owner's data while releasing to analysts noisy versions of aggregate query results. But such strict protections of individual's data have not yet found wide use in practice. Instead, Internet companies, for example, commonly provide free services in return for valuable sensitive information from users, which they exploit and sometimes sell to third parties. As the awareness of the value of the personal data increases, so has the drive to compensate the end user for her private information. The idea of monetizing private data can improve over the narrower view of hiding private data, since it empowers individuals to control their data through financial means. In this paper we propose a theoretical framework for assigning prices to noisy query answers, as a function of their accuracy, and for dividing the price amongst data owners who deserve compensation for their loss of privacy. Our framework adopts and extends key principles from both differential privacy and query pricing in data markets. We identify essential properties of the price function and micro-payments, and characterize valid solutions.

Monday, March 18			
14:00-15:30	ICDT Research Session 2		**Semi-structured Data and XML**
Room:	Munizioniere0	Chair:	Balder ten Cate

Fast Learning of Restricted Regular Expressions and DTDs

Dominik D. Freydenberger and Timo Kötzing

We study the problem of generalizing from a finite sample to an infinite language taken from a predefined language class. The two language classes we consider are subsets of the regular languages and have significance in the specification of XML documents (the classes corresponding to so called chain regular expressions, CHAREs, and to single occurrence regular expressions, SOREs). The previous literature gave a number of algorithms for generalizing to SOREs providing a trade off between speed and quality of the solution. Furthermore, a fast but non-optimal algorithm for generalizing to CHAREs is known. For each of the two language classes we give an efficient algorithm returning a minimal generalization from the given finite sample to an element of the fixed language class; such generalizations are called descriptive. In this sense, both our algorithms are optimal.

Which DTDs are streaming bounded repairable?

Pierre Bourhis, Gabriele Puppis and Cristian Riveros

Integrity constraint management concerns both checking whether data is valid and taking action to restore correctness when invalid data is discovered. In XML the notion of valid data can be captured by schema languages such as Document Type Definitions (DTDs) and more generally XML schemas. DTDs have the property that constraint checking can be done in streaming fashion. In this paper we consider when the corresponding action to restore validity -- repair -- can be done in streaming fashion. We formalize this as the problem of determining, given a DTD, whether or not a streaming procedure exists that transforms an input document so as to satisfy the DTD, using a number of edits independent of the document. We show that this problem is decidable. In fact, we show the decidability of a more general problem, allowing a more general class of schemas than DTDs, and requiring a repair procedure that works only for documents that are already known to satisfy another class of constraints. The decision procedure relies on a new analysis of the structure of DTDs, reducing to a novel notion of game played on pushdown systems associated with the schemas.

XML compression via DAGs

Sebastian Maneth, Markus Lohrey and Eric Noeth

Unranked trees can be represented using their minimal dag (directed acyclic graph). For XML this achieves high compression ratios due to the repetitive mark up. Unranked trees are often represented through first child/next sibling (fcns) encoded binary trees. We study the difference in size (= number of edges) of minimal dag versus minimal dag of the fcns encoded binary tree. One main finding is that the size of the dag of the binary tree can never be smaller than the square root of the size of the minimal dag, and that there are examples that match this bound. We introduce a new combined structure, the hybrid dag, which is guaranteed to be smaller than (or equal in size to) both dags. Interestingly, we find through experiments that last child/previous sibling encodings are much better for XML compression via dags, than fcns encodings. This is because optional elements are more likely to appear towards the end of child sequences.

Monday, March 18			
16:00-17:30	ICDT Research Session 3		**Query Processing and Optimization**
Room:	Munizioniere0	Chair:	Frank Neven

Structural Tractability of Counting of Solutions to Conjunctive Queries

Arnaud Durand and Stefan Mengel

In this paper we explore the problem of counting solutions to conjunctive queries. We consider a parameter called the quantified star size of a formula φ which measures how the free variables are spread in φ. We show that for conjunctive queries that admit nice decomposition properties (such as being of bounded treewidth or generalized hypertree width) bounded quantified star size exactly characterizes the classes of queries for which counting the number of solutions is tractable. This also allows us to fully characterize the conjunctive queries for which counting the solutions is tractable in the case of bounded arity. To illustrate the applicability of our results, we also show that computing the quantified star size of a formula is possible in time $n^{O(k)}$ for queries of generalized hypertree width *k*. Furthermore, quantified star size is even fixed parameter tractable parameterized by some less general width measures, while it is $\W{1}$-hard for generalized hypertree width and thus unlikely to be fixed parameter tractable. We finally show how to compute an approximation of quantified star size even in polynomial time where the approximation ratio depends on the width of the input.

Recursive queries on trees and data trees

Serge Abiteboul, Pierre Bourhis, Anca Muscholl and Zhilin Wu

The analysis of datalog programs over relational structures has been studied in depth, most notably the problem of containment. The analysis problems that have been considered were shown to be undecidable with the exception of (i) containment of arbitrary programs in nonrecursive ones, (ii) containment of monadic programs, and (iii) emptiness. In this paper, we are concerned with a much less studied problem, the analysis of datalog programs over data trees. We show that the analysis of datalog programs is more complex for data trees than for arbitrary structures. In particular, we prove that the three aforementioned problems are undecidable for data trees. In practice, the data trees (e.g., XML trees) are often of bounded depth. We prove that all three problems are decidable over bounded depth data trees. Another contribution of the paper is the study of a new form of automata, namely pattern automata. Pattern automata are essentially equivalent to linear datalog programs. They provide a useful alternative viewpoint to the class of functions that can be captured by linear datalog programs. In particular, we use them to show that the emptiness is decidable, over data trees, for linear monadic datalog programs with data value inequalities.

On optimum left-to-right strategies for active context-free games

Henrik Björklund, Martin Schuster, Thomas Schwentick and Joscha Kulbatzki

We consider *context-free games*, which are two-player games on strings from finite alphabets with one player trying to rewrite the input string to match a target specification. These games have been investigated in the context of exchanging Active XML (AXML) data [AbiteboulMiloBenjelloun05, MuschollSchwentickSegoufin06]. While the rewriting problem is undecidable in general, we show that it is decidable whether all safely rewritable strings can be safely rewritten in a *left-to-right* manner, a problem that was previously considered in [AbiteboulMiloBenjelloun05]. We also investigate the complexity of this problem.

Tuesday, March 19			
11:00-12:30	ICDT Research Session 4		**Graph Databases**
Room:	Munizioniere0	Chair:	Gerome Miklau

Walk Logic as a framework for path query languages on graph databases

Jelle Hellings, Bart Kuijpers, Jan Van Den Bussche and Xiaowang Zhang

Motivated by the current interest in languages for expressing path queries to graph databases, this paper proposes to investigate Walk Logic (WL): the extension of first-order logic on finite graphs with the possibility to explicitly quantify over walks. WL can serve as a unifying framework for path query languages. To support this claim, WL is compared in expressive power with various established query languages for graphs, such as first-order logic extended with reachability; the monadic second-order logic of graphs; hybrid computation tree logic; and regular path queries. WL also serves as a framework to investigate the following natural questions: Is quantifying over walks more powerful than quantifying over paths (walks without repeating nodes) only? Is quantifying over infinite walks more powerful than quantifying over finite walks only? WL model checking is decidable, but determining the precise complexity remains an open problem.

Querying Graph Databases with XPath

Leonid Libkin, Wim Martens and Domagoj Vrgoc

XPath plays a prominent role as an XML navigational language due to several factors, including its abilityto express queries of interest, its close connection to yardstick database query languages (e.g., first-order logic), and the low complexity of query evaluation for many fragments. Another common database model - graph databases - also requires a heavy use of navigation in queries; yet it largely adopts a different approach to querying, relying on reachabilitypatterns expressed with regular constraints. Our goal here is to investigate the behavior and applicabilityof XPath-like languages for querying graph databases, concentrating on their expressiveness and complexity of query evaluation. We are particularly interested in a model of graph data that combines navigation through graphs with querying data held in the nodes, such as, for example, in a social network scenario. As navigational languages, we use analogs of core and regular XPath and augment them with various tests on data values. We relate these languages to first-order logic, its transitive closure extensions, and finite-variable fragments thereof, proving several capture results. In addition, we describe their relative expressive power. We then show that they behave very well computationally: they have a low-degree polynomial combined complexity, which becomes linear for several fragments. Furthermore, we introduce new types of tests for XPath languages that let them capture first-order logic with data comparisons and prove that the low complexity bounds continue to apply to such extended languages.

Definability problems for graph query languages

Timos Antonopoulos, Frank Neven and Frédéric Servais

Given a graph, a relation on its nodes, and a query language Q of interest, we study the Q-definability problem which amounts to deciding whether there exists a query in Q defining precisely the given relation over the given graph. Previous research has identified the complexity of FO- and CQ-definability. In this paper, we consider the definability problem for regular paths and conjunctive regular path queries (CRPQs) over labelled graphs.

Tuesday, March 19			
14:00-15:30	ICDT Research Session 5		**ICDT Invited Lecture**
Room:	Munizioniere0	Chair:	Maurizio Lenzerini

Tuesday, March 19			
16:00-17:30	ICDT Research Session 6		**Provenance and Annotations**
Room:	Munizioniere0	Chair:	James Cheney

Algebraic Structures for Capturing the Provenance of SPARQL Queries

Floris Geerts, Grigoris Karvounarakis, Vassilis Christophides and Irini Fundulaki

We show that the evaluation of SPARQL algebra queries on various notions of annotated RDF graphs can be seen as particular cases of the evaluation of these queries on RDF graphs annotated with elements of so-called spm-semirings. Spm-semirings extend semirings, used for positive relational algebra queries on annotated relational data, with a new operator to capture the semantics of the non-monotone SPARQL operator OPTIONAL. Furthermore, spm-semiring-based annotations ensure that desired SPARQL query equivalences hold when querying annotated RDF. In addition to introducing spm-semirings, we study their properties and provide an alternative characterization of these structures in terms of semirings with an embedded boolean algebra (or seba-structure for short). This characterization provides a way of constructing spm-semirings and of identifying a universal object in the class of spm-semirings. Finally, we show that the universal object provides a concise provenance representation and can be used to evaluate SPARQL queries on arbitrary spm-semiring annotated RDF graphs.

A Propagation Model for Provenance Views of Public/Private Workflows

Susan Davidson, Tova Milo and Sudeepa Roy

We study the problem of concealing functionality of a proprietary or private module when provenance information is shown over repeated executions of a workflow which contains both `public' and `private' modules. Our approach is to use `provenance views' to hide carefully chosen subsets of data over all executions of the workflow to ensure Γ-privacy: for each private module and each input x, the module's output f(x) is indistinguishable from Γ*-1* other possible values given the visible data in the workflow executions. We show that Γ-privacy cannot be achieved simply by combining solutions for individual private modules; data hiding must also be `propagated' through public modules. We then examine how much additional data must be hidden and when it is safe to stop propagating data hiding. The answer depends strongly on the workflow topology as well as the behavior of public modules on the visible data. In particular, for a class of workflows (which include the common tree and chain workflows), taking private solutions for each private module, augmented with a `public closure' that is `upstream-downstream safe', ensures Γ-privacy. We define these notions formally and show that the restrictions are necessary. We also study the related optimization problems of minimizing the amount of hidden data.

Annotations are Relative

Peter Buneman, Egor V. Kostylev and Stijn Vansummeren

Most attempts to provide a formal description of annotation describe a two-level structure in which annotation is superimposed on, and separate from, the data. In this paper we examine the possibility of a hierarchical model of annotation in which an annotated structure may itself be annotated. There are already practical examples of this: threads in e-mail and newsgroups allow the imposition of one comment on another, chains of belief can be seen as hierarchies of belief annotations; and annotations such as valid time and belief can be freely imposed on each other. In this paper we study the structure and querying of annotation hierarchies. First, we introduce a term model for annotations and in order to express the fact that an annotation may apply two or more data values with some shared structure we provide a simple schema for annotation hierarchies. We then look at how queries can be applied to such hierarchies; in particular we ask the usual question of how annotations should propagate through queries. We take the view that the query together with schema describes a level in the hierarchy: everything below this level is treated as data to which the query should be applied; everything above it is annotation which should, according to certain rules, be propagated with the query. We also examine the representation of annotation hierarchies in conventional relational structures and ask the question what it means -- in a rule based system -- for annotations to be superimposed in the sense that adding an annotation should not cause changes to the underlying data.

Wednesday, March 20			
11:00-12:30	ICDT Research Session 7		**Data Exchange and Query Answering**
Room:	Camino	Chair:	Benny Kimelfeld

Schema Mappings and Data Exchange for Graph Databases

Pablo Barceló, Jorge Pérez and Juan L. Reutter

Data exchange and schema mapping management have received little attention so far in the graph database scenario, and tools developed in this context for relational databases have significant drawbacks in the context of graph-structured data. In this paper we embark on the study of interoperability issues for graph databases, including schema mappings, data exchange and certain answers computation. We start by analyzing different possibilities for specifying mappings in graph databases. Our mapping languages are based on the most typical graph databases queries, ranging from regular path queries to conjunctions of nested regular expressions (NREs). They subsume all previously considered mapping languages, and let one express many data exchange scenarios in the graph database context. We study the problems of materializing solutions and query answering, in particular, the problem of computing universal representatives and certain answers for various classes of mappings. We show that both problems are difficult with respect to combined complexity, and that for the latter problem, even data complexity is high for some very simple mappings and queries. We then identify relevant classes of mappings and queries for which the problems of materializing solutions and query answering can be solved efficiently.

Containment of Pattern-Based Queries over Data Trees

Claire David, Amelie Gheerbrant, Leonid Libkin and Wim Martens

We study static analysis, in particular the containment problem, for analogs of conjunctive queries over XML documents. The problem has been studied for queries based on arbitrary patterns, not necessarily following the tree structure of documents. However, many applications force the syntactic shape of queries to be tree-like, as they are based on proper tree patterns. This renders previous results, crucially based on having non-tree-like features, inapplicable. Thus, we investigate static analysis of queries based on proper tree atterns. We go beyond simple navigational conjunctive queries in two ways: we look at unions and Boolean combinations of such queries as well and, crucially, all our queries handle data stored in documents, i.e., we deal with containment over data trees. We start by giving a general Pi^p_2 upper bound on the containment of conjunctive queries and Boolean combinations for patterns that involve all types of navigation through documents. We then show matching hardness for conjunctive queries with all navigation, or their Boolean combinations with the simplest form of navigation. After that we look at cases when containment can be witnessed by homomorphisms of analogs of tableaux. These include conjunctive queries and their unions over child and next-sibling axes; however, we show that not all cases of containment can be witnessed by homomorphisms. We look at extending tree patterns used in queries in three possible ways: with wildcard, with schema information, and with data-value comparisons. The first one is relatively harmless, the second one tends to increase complexity by an exponential, and the last one quickly leads to undecidability.

Access Patterns and Integrity Constraints Revisited

Vince Barany, Michael Benedikt and Pierre Bourhis

We consider which queries are answerable in the presence of access restrictions and integrity constraints, and which portions of the schema are accessible in the presence of access restrictions and constraints. Unlike prior work, we focus on integrity constraint languages that subsume inclusion dependencies. We also use a semantic definition of answerability: a query is answerable if the accessible information is sufficient to determine its truth value.We show that answerability is decidable for the class of guarded dependencies, which includes all inclusion dependencies, and also for constraints given in the guarded fragment of first-order logic. We also show that answerable queries have “query plans” in a restricted language. We give corresponding results for extractability of portions of the schema. Our results relate querying with limited access patterns, determinacy-vs-rewriting, and analysis of guarded constraints.

Wednesday, March 20			
14:00-15:30	ICDT Research Session 8		**Ranking Query Answers**
Room:	Camino	Chair:	Michael Benedikt

Using the Crowd for Top-k and Group-by Queries

Susan Davidson, Sanjeev Khanna, Tova Milo and Sudeepa Roy

Group-by and top-k are fundamental constructs in database queries. However, the criteria used for grouping and ordering certain types of data - such as unlabeled photos clustered by the same person ordered by age - are difficult to evaluate by machines. In contrast, these tasks are easy for humans to evaluate and are therefore natural candidates for being crowd-sourced. We study the problem of evaluating top-k and group-by queries using the crowd to answer either `type' or `value' questions. Given two data elements, the answer to a type question is `yes' if the elements have the same type and therefore belong to the same group or cluster; the answer to a value question orders the two data elements. The assumption here is that there is an underlying ground truth, but that the answers returned by the crowd may sometimes be erroneous. We formalize the problems of top-k and group-by in the crowd-sourced setting, and give efficient algorithms that are guaranteed to achieve good results with high probability. We analyze the crowd-sourced cost of these algorithms in terms of the total number of type and value questions, and show that they are essentially the best possible. We also show that fewer questions are needed when values and types are correlated, or when the error model is one in which the error decreases as the distance between the two elements in the sorted order increases.

Certain and Possible XPath Answers

Sara Cohen and Yaacov Y. Weiss

Formulating an XPath query over an XML document is a difficult chore for a non-expert user. This paper introduces a novel approach to ease the querying process. Instead of specifying a query, the user simply marks positive examples X+ of nodes that t her information need. She may also mark negative examples X- of undesirable nodes. A deductive method, to suggest additional nodes that will interest the user, is developed in this paper. To be precise, a node y is a certain answer if every query returning all positive examples X+, and not returning any negative example from X-, must also return y. Similarly, y is a possible answer if there exists a query returning X+ and y, while not returning any node in X-. Thus, y is likely to be of interest to the user if y is a certain answer, and unlikely to be of interest if y is not even a possible answer. The complexity of finding certain and possible answers, with respect to various classes of XPath, is studied. It is shown that for a wide variety of XPath queries (including child and descendant axes, wildcards, branching and attribute constraints), certain and possible answers can be found efficiently, provided that X+ and X- are of bounded size. To prove this result a novel algorithm is developed.

Extracting Minimum-Weight Tree Patterns from a Schema with Neighborhood Constraints

Benny Kimelfeld and Yehoshua Sagiv

The task of formulating queries is greatly facilitated when they can be generated automatically from some given data values, schema concepts or both (e.g., names of particular entities and XML tags). This automation is the basis of various database applications, such as keyword search and interactive query formulation. Usually, automatic query generation is realized by finding a set of small tree patterns that contain some given labels. More formally, the computational problem at hand is to find top-k patterns, that is, k minimum-weight tree patterns that contain a given bag of labels, conform to the schema, and are non-redundant. A plethora of systems and research papers include a component that deals with this problem. This paper presents an algorithm for this problem, with complexity guarantees, that allows nontrivial schema constraints and, hence, avoids generating patterns that cannot be instantiated. Specifically, this paper shows that for schemas with certain types of neighborhood constraints, the problem is fixed-parameter tractable (FPT), the parameter being the size of the given bag of labels. As machinery, an adaptation of Lawler-Murty's procedure is developed. This adaptation reduces a top-k problem, over an infinite space of solutions, to a prefix-constrained optimization problem. It is shown how to cast the problem of top-k patterns in this adaptation. A solution is developed for the corresponding prefix-constrained optimization problem, and it uses an algorithm for finding a (single) minimum-weight tree pattern. This algorithm generalizes an earlier work by handling leaf constraints (i.e., which labels may, must or should not be leaves). It all boils down to a reduction showing that, under a language for neighborhood constraints, finding top-k patterns is FPT if a certain variant of exact cover is FPT.

Wednesday, March 20			
16:00-17:30	ICDT Research Session 9		**Privacy**
Room:	Camino	Chair:	Sebastian Maneth

On Optimal Differentially Private Mechanisms for Count-Range Queries

Chen Zeng, Jin-Yi Cai, Pinyan Lu and Jeffrey Naughton

While there is a large and growing body of literature on differentially private mechanisms for answering various classes of queries, to the best of our knowledge “count-range” queries have not been studied. These are a natural class of queries that ask “is the number of rows in a relation satisfying a given predicate between two integers Θ_1 and Θ_2?” Such queries can be viewed as a simple form of SQL “having” queries. We begin by developing a provably optimal differentially private mechansim for count-range queries for a single consumer. For count queries (in contrast to count-range queries), Ghosh et al.[Ghosh:2009] have provided a differentially private mechanism that simultaneously maximizes utility for multiple consumers. This raises the question of whether such a mechanism exists for count-range queries. We prove that the answer is no. However, perhaps surprisingly, we prove that such a mechanism does exist for “threshold” queries, which are simply count-range queries for which either $\Theta_1 = 0$ or $\Theta_2 = +\infty$. Furthermore, we prove that this mechanism is a two-approximation for general count-range queries.

Optimal Error of Query Sets under the Differentially-private Matrix Mechanism

Chao Li and Gerome Miklau

A common goal of privacy research is to release synthetic data that satisfies a formal privacy guarantee and can be used by an analyst in place of the original data. To achieve reasonable accuracy, a synthetic data set must be tuned to support a specified set of queries accurately, sacrificing fidelity for other queries. This work considers methods for producing synthetic data under differential privacy and investigates what makes a set of queries “easy” or “hard” to answer. We consider answering sets of linear counting queries using the matrix mechanism, a recent differentially-private mechanism that can reduce error by adding complex correlated noise adapted to a specified workload. Our main result is a novel lower bound on the minimum total error required to simultaneously release answers to a set of workload queries. The bound reveals that the hardness of a query workload is related to the spectral properties of the workload when it is represented in matrix form. The bound is most informative for (ε,δ)-differential privacy but also applies to ε -differential privacy.

Private Predicate Sums on Decayed Streams

Jean Bolot, Nadia Fawaz, S. Muthukrishnan, Aleksandar Nikolov and Nina Taft

In many monitoring applications, recent data is more important than distant data. How does this affect privacy of data analysis? We study a general class of data analyses - predicate sums - in this context. Formally, we study the problem of estimating predicate sums *privately*, for sliding windows and other decay models. While we require accuracy in analysis with respect to the decayed sums, we still want differential privacy for the entire past. This is challenging because window sums are not monotonic or even near-monotonic as the problems studied previously [dwork-continual]. Predicate sums is a fundamental problem that is a useful subroutine for many data analyses tasks. We present accurate $\eps$-differentially private algorithms for decayed sums. For window and exponential decay sums, our algorithms are accurate up to additive $1/\eps$ and polylog terms in the range of the computed function; for polynomial decay sums which are technically more challenging because partial solutions do not compose easily, our algorithms incur additional relative error. Our algorithm for polynomial decay sums generalizes to arbitrary decay sum functions. The algorithm crucially relies on our solution for the window sum problem as a subroutine. Further, we show lower bounds, tight within polylog factors and tight with respect to the dependence on the probability of error. Our results are obtained via a natural dyadic tree we maintain, but the crux is we treat the tree data structure in non-uniform manner. We also extend our study and consider the “dual” question of maintaining conventional running sums on the entire data thus far, but when privacy constraints expire with time. We define a new model of privacy with expiration and consider the problems of designing accurate running sum and linear map algorithms in this model. Now the goal is to design algorithms whose accuracy guarantees do not deteriorate with the size of the entire input, but rather scale with the size of the privacy window. We reduce running sum with a privacy window *W* to window sum without privacy expiration. We also characterize the accuracy of output perturbation for general linear maps with privacy window *W*.

EDBT Research Sessions

Tuesday, March 19			
11:00-12:30	EDBT Research Session 1		**High Performance Query Processing**
Room:	Maggior Consiglio	Chair:	Stefan Manegold

From A to E: Analyzing TPC's OLTP Benchmarks - The obsolete, the ubiquitous, the unexplored

Pinar Tozun, Ippokratis Pandis, Cansu Kaynak, Djordje Jevdjic and Anastasia Ailamaki

Introduced in 2007, TPC-E is the most recently standardized OLTP benchmark by TPC. Even though TPC-E has been around already for six years, it has not gained the popularity of its predecessor TPC-C: all the published results for TPC-E use a single database vendor's product. TPCE is significantly different than its predecessors. Some of its distinguishing characteristics are the non-uniform input creation, longer-running and more complicated transactions, more difficult partitioning etc. These factors slow down the adoption of TPC-E. In turn, there is little knowledge in the community about how TPC-E behaves micro-architecturally and within the database engine. To shed light on TPC-E, we implement it on top of a scalable open-source database engine, Shore-MT, and perform a workload characterization study, comparing it with the previous, much better known OLTP benchmarks of TPC: TPC-B and TPC-C. In parallel, we study the evolution of the OLTP benchmarks throughout the decades. Our results demonstrate that TPC-E exhibits similar micro-architectural behavior to TPC-B and TPC-C, even though it incurs less stall time and higher instructions per cycle. On the other hand, within the database engine it suffers more from logical lock contention. Therefore, we argue that, on the hardware side, TPC-E needs less aggressive processors. Whereas on the software side it can benefit from designs based on intra-transaction parallelism, logical partitioning, and optimistic concurrency control to minimize the effects of lock contention without introducing distributed transactions.

Query-Aware Compression of Join Results

Christopher M. Mullins, Lipyeow Lim and Christian A. Lang

Client-server database query processing has become an important paradigm in many data processing applications today. In cloud-based data services, for example, queries over structured data are sent to cloud-based servers for processing and the results relayed back to the client devices. Network bandwidth between client devices and cloud-based servers is often a limited resource and the use of data compression to reduce the amount of query result data transmitted would not only conserve bandwidth but also help with battery lifetime in the case of mobile client devices. For query result compression, current data compression methods do not exploit redundancy information that can be inferred from the query structure itself for greater compression. In this paper we propose a novel query-aware compression method for compressing query results sent from database servers to client applications. Our method is based on two key ideas. We exploit redundancy information obtained from the query plan and possibly from the database schema to achieve better compression than standard non-query aware compressors. We use a collection of memory-limited dictionaries to encode attribute values in a lightweight and efficient manner. We evaluated our method empirically using the TPC-H benchmark show that this technique is effective especially when used in conjunction with standard compressors. Our results show that compression ratios of up to twice that of gzip are possible.

Web Data Indexing in the Cloud: Efficiency and Cost Reductions

Jesús Camacho-Rodríguez, Dario Colazzo and Ioana Manolescu

An increasing part of the world's data is either shared through the Web or directly produced through and for Web platforms, in particular using structured formats like XML or JSON. Cloud platforms are interesting candidates to handle large data repositories, due to their elastic scaling properties. Popular commercial clouds provide a variety of sub-systems and primitives for storing data in specific formats (files, key-value pairs etc.) as well as dedicated sub-systems for running and coordinating execution within the cloud. We propose an architecture for warehousing large-scale Web data, in particular XML, in a commercial cloud platform, specifically, Amazon Web Services. Since cloud users support monetary costs directly connected to their consumption of cloud resources, we focus on indexing content in the cloud. We study the applicability of several indexing strategies, and show that they lead not only to reducing query evaluation time, but also, importantly, to reducing the monetary costs associated with the exploitation of the cloud-based warehouse. Our architecture can be easily adapted to similar cloud-based complex data warehousing settings, carrying over the benefits of access path selection in the cloud.

Tuesday, March 19			
11:00-12:30	EDBT Research Session 2		**Multi-tenant Databases**
Room:	Camino	Chair:	Alfredo Cuzzocrea

Live Database Migration for Multi-tenant RDBMS with Snapshot Isolation

Oliver Schiller, Nazario Cipriani and Bernhard Mitschang

The consolidation of multiple tenants onto a single RDBMS instance turned out to be benefical with respect to resource utilization and scalability. The consolidation implies that multiple tenants share the physical resources available for the RDBMS instance. If the available resources tend to get insufficient to meet the SLAs agreed with the tenants, migration of a tenant's database from one RDBMS instance to another is compelling. Highly available services demand for live migration techniques that come with minimal service interruption and low performance impact. This paper meets the demand for live migration techniques by contributing ProRea. ProRea is a live database migration approach designed for multi-tenant RDBMS that run OLTP workloads under snapshot isolation. ProRea extends concepts of existing live database migration approaches to accomplish minimal service interruption, high efficiency and very low migration overhead. Measurements of a prototypical ProRea implementation underpin its good performance.

SWAT: A Lightweight Load Balancing Method for Multitenant Databases

Hyun Moon, Hakan Hacigumus, Yun Chi and Wang-Pin Hsiung

Multitenant databases achieve cost efficiency through the consolidation of multiple small tenants. However, performance isolation is an inherent problem in multitenant databases due to resource sharing among the tenants. That is, a bursty workload from a co-located tenant, i.e., a noisy neighbor, may affect the performance of the other tenants sharing the same system resources. We address this issue by using a load balancing method that is based on database replica swap. Unlike the traditional data migration-based load balancing, replica swapbased load balancing does not incur data movement, which makes it highly resource- and time-efficient. We propose a novel method of choosing which tenants should be subject to swaps. Our experimental results show that swap-based load balancing effectively reduces the number of SLA violations, which is the main performance metric we choose.

CloudOptimizer: Multi-tenancy for I/O-Bound OLAP Workloads

Hatem Mahmoud, Hyun Moon, Yun Chi, Hakan Hacigumus, Divyakant Agrawal and Amr El-Abbadi

Consolidation of multiple databases on the same server allows service providers to save significant resources because many production database servers are often under-utilized. Recent research investigates the problem of minimizing the number of servers required to host a set of tenants when the working sets of tenants are kept in main memory (e.g., in-memory OLAP workloads, or OLTP workloads), thus the memory assigned to each tenant, as well as the I/O bandwidth and CPU time, are all dictated by the working set size of the tenant. Other research investigates the reverse problem when the number of servers is fixed, but the amount of resources allocated to different tenants on the same server needs to be configured to optimize a cost function. In this paper we investigate the problem when neither the number of servers nor the amount of resources allocated to each tenant are fixed. This problem arises when consolidating OLAP workloads of tenants whose service-level agreements (SLAs) allow for queries to be answered from disk. We study the trade-off between the amount of memory and the I/O bandwidth assigned to OLAP workloads, and develop a principled approach for allocating resources to tenants in a manner that minimizes the total number of servers required to host all tenants while satisfying the SLA of each tenant. We then explain how we modified InnoDB, the storage engine of MySQL, to be able to change the amount of resources allocated to each tenant at runtime, so as to account for fluctuations in workloads. Finally, we evaluate our approach experimentally using the TPC-H benchmark to demonstrate its effectiveness and accuracy.

Tuesday, March 19			
14:00-15:30	EDBT Research Session 3		**MapReduce**
Room:	Maggior Consiglio	Chair:	Ioana Manolescu

Eagle-Eyed Elephant: Split-Oriented Indexing in Hadoop

Mohamed Eltabakh, Fatma Ozcan, Yannis Sismanis, Peter Haas, Hamid Pirahesh and Jan Vondrak

An increasingly important analytics scenario for Hadoop involves multiple (often ad hoc) grouping and aggregation queries with selection predicates over a slowly-changing dataset. These queries are typically expressed via high-level query languages such as Jaql, Pig, and Hive, and are used either directly for business-intelligence applications or to prepare the data for statistical model building and machine learning. In such scenarios it has been increasingly recognized that, as in classical databases, techniques for avoiding access to irrelevant data can dramatically improve query performance. Prior work on Hadoop, however, has simply ported classical techniques to the MapReduce setting, focusing on record-level indexing and key-based partition elimination. Unfortunately, record-level indexing only slightly improves overall query performance, because it does not minimize the number of mapper ``waves", which is determined by the number of processed splits. Moreover, key-based partitioning requires data reorganization, which is usually impractical in Hadoop settings. We therefore need to re-envision how data access mechanisms are defined and implemented. To this end, we introduce the Eagle-Eyed Elephant (E3) framework for boosting the efficiency of query processing in Hadoop by avoiding accesses of data splits that are irrelevant to the query at hand. Using novel techniques involving inverted indexes over splits, domain segmentation, materialized views, and adaptive

caching, E3 avoids accessing irrelevant splits even in the face of evolving workloads and data. Our experiments show that E3 can achieve up to 20x cost savings with small to moderate storage overheads.

Computing n-Gram Statistics in MapReduce

Klaus Berberich and Srikanta Bedathur

Statistics about n-grams (i.e., sequences of contiguous words or other tokens in text documents or other string data) are an important building block in information retrieval and natural language processing. In this work, we study how *n*-gram statistics, optionally restricted by a maximum n-gram length and minimum collection frequency, can be computed efficiently harnessing MapReduce for distributed data processing. We describe different algorithms, ranging from an extension of word counting, via methods based on the Apriori principle, to a novel method Suffix-\sigma that relies on sorting and aggregating suffixes. We examine possible extensions of our method to support the notions of maximality/closedness and to perform aggregations beyond occurrence counting. Assuming Hadoop as a concrete MapReduce implementation, we provide insights on an efficient implementation of the methods. Extensive experiments on The New York Times Annotated Corpus and ClueWeb09 expose the relative benefits and trade-offs of the methods.

Processing Multi-way Spatial Joins On Map-Reduce

Himanshu Gupta, Bhupesh Chawda, Sumit Negi, Tanveer Faruquie, Lv Subramaniam and Mukesh Mohania

In this paper we investigate the problem of processing multi-way spatial joins on map-reduce platform. We look at two common spatial predicates - *overlap* and *range*. We address these two classes of join queries, discuss the challenges and outline novel approaches for executing these queries on a map-reduce framework. We then discuss how we can process join queries involving both *overlap* and *range* predicates. Specifically we present a *Controlled-Replicate* framework using which we design the approaches presented in this paper. The *Controlled-Replicate* framework is carefully engineered to minimize the communication among cluster nodes. Through experimental evaluations we discuss the complexity of the problem under

investigation, details of *Controlled-Replicate* framework and demonstrate that the proposed approaches comfortably outperform naive approaches.

Tuesday, March 19			
14:00-15:30	EDBT Research Session 4		**Extending Database Technology**
Room:	Camino	Chair:	Bernhard Mitschang

Rapid Experimentation for Testing and Tuning a Production Database Deployment

Nedyalko Borisov and Shivnath Babu

The need to perform testing and tuning of database instances with production-like workloads (W), configurations (C), data (D), and resources (R) arises routinely. The further W, C, D, and R used in testing and tuning deviate from what is observed on the production database instance, the lower is the trustworthiness of the testing and tuning tasks done. For example, it is common to hear about performance degradation observed after the production database is upgraded from one software version to another. A typical cause of this problem is that the W, C, D, or R used during upgrade testing differed in some way from that on the production database. Performing testing and tuning tasks in principled and automated ways is very important, especially since “spurred by innovations in cloud computing” the number of database instances that a database administrator (DBA) has to manage is growing rapidly. We present Flex, a platform for trustworthy testing and tuning of production database instances. Flex gives DBAs a declarative language, called Slang, to specify definitions and objectives regarding running experiments for testing and tuning. Flex’s orchestrator schedules and runs these experiments in an automated manner that meets the DBA-specified objectives. Flex has been fully prototyped. We present results from a comprehensive empirical evaluation that reveals the effectiveness of Flex on diverse problems such as upgrade testing, near-real-time testing to detect corruption of data, and server configuration tuning. We also report on our experiences taking some of the testing and tuning software described in the literature and porting them to run on the Flex platform.

Proactive Natural Language Search Engine: Tapping into Structured Data on the Web

Wensheng Wu

In this era of “big data”, a key challenge facing the database community is to help average users tap into the huge amounts of structured data on the Web. To address this challenge, we propose a novel proactive template-based engine for searching structured data on the Web using natural language. Departing from conventional search engines, the proposed engine organizes questions it can answer using templates and figures out ahead of time which sources can answer which templates and how. Then, at query time, the engine can simply match queries with the templates and retrieve answers using the pre-compiled evaluation plans. While attractive, building such an engine requires innovations in template creation, query evaluation, and system evolution. In this paper, we propose novel techniques to address these challenges.

Towards Context-Aware Search and Analysis on Social Media Data

Leon Derczynski, Christian Jensen and Bin Yang

Social media has changed the way we communicate. Social media data capture our social interactions and utterances in machine readable format. Searching and analysing massive and frequently updated social media data brings significant and diverse rewards across many different application domains, from politics and business to social science and epidemiology. A notable proportion of social media data comes with explicit or implicit spatial annotations, and almost all social media data has temporal metadata. In this paper, we view social media data as a constant stream of data points, each containing text with spatial and temporal contexts. We identify challenges relevant to each context, which we intend to contend with using context aware querying and analysis, specifically including longitudinal analyses on social media archives, spatial keyword search, local intent search, and spatio-temporal intent search. Finally, for each context, emerging application scenarios and further avenues for investigation are discussed.

Anomaly Management using Complex Event Processing

Bastian Hoßbach and Bernhard Seeger

During the last decade, complex event processing (CEP) has emerged as a technological foundation for many time-critical monitoring applications. CEP is powerful, effective, easy to use and low in costs at the same time. Common CEP applications are for example stock-market analysis, detection of fraudulent credit card use, traffic monitoring and consumption forecasting in power grids. Many application domains are still hard to target by CEP, because state of the art CEP technology is characterized by a static behavior and by a signature-based detection paradigm. In this paper, we motivate substantial improvements of CEP technology by making the behavior of the infrastructure dynamic and by switching the detection paradigm from signatures to anomalies. This leads to multiple changes in the infrastructure that raise interesting and challenging research questions. The resulting dynamic CEP infrastructure not only makes existing applications more powerful and easier to maintain but also enables novel application domains.

Tuesday, March 19			
16:00-18:00	EDBT Research Session 5		**Privacy**
Room:	Maggior Consiglio	Chair:	Ehud Gudes

Compromising Privacy in Precise Query Protocols

Jonathan Dautrich and Chinya Ravishankar

Privacy and security for outsourced databases are often provided by Precise Query Protocols (PQPs). In a PQP, records are individually encrypted by a client and stored on a server. The client issues encrypted queries, which are run under encryption at the server, and the server returns the exact set of encrypted tuples needed to satisfy the query. We propose a general attack against the privacy of all PQPs that support range queries, using query results to partially order encrypted records. Existing attacks that seek to order etuples are less powerful and depend on weaknesses specific to particular PQPs. Our novel algorithm identifies permissible positions (loci) for encrypted records by organizing range query results using PQ-trees. These results can then be used to infer attribute values of encrypted records. We propose equivocation and permutation entropy as privacy metrics, and give experimental results that show PQP privacy to be easily compromised by our attack.

Efficient Privacy-Aware Record Integration

Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham and Bradley Malin

The integration of information dispersed among multiple repositories is a crucial step for accurate data analysis in various domains. In support of this goal, it is critical to devise procedures for identifying similar records across distinct data sources. At the same time, to adhere to privacy regulations and policies, such procedures should protect the confidentiality of the individuals to whom the information corresponds. Various private record linkage (PRL) protocols have been proposed to achieve this goal, involving secure multi-party computation (SMC) and similarity preserving data transformation techniques. SMC methods provide secure and accurate solutions to the PRL problem, but are prohibitively expensive in practice, mainly due to excessive computational requirements. Data transformation techniques offer more practical solutions, but incur the cost of information leakage and false matches. In this paper, we introduce a novel model for practical PRL, which 1) affords controlled and limited information leakage, 2) avoids false matches resulting from data transformation. Initially, we partition the data sources into blocks to eliminate comparisons for records that are unlikely to match. Then, to identify matches, we apply an efficient SMC technique between the candidate record pairs. To enable efficiency and privacy, our model leaks a controlled amount of obfuscated data prior to the secure computations. Applied obfuscation relies on differential privacy which provides strong privacy guarantees against adversaries with arbitrary background knowledge. In addition, we illustrate the practical nature of our approach through an empirical analysis with data derived from public voter records.

Updating Outsourced Anatomized Private Databases

Ahmet Erhan Nergiz, Chris Clifton and Qutaibah Malluhi

We introduce operations to safely update an anatomized database. The result is a database where the view of the server satisfies standards such as kanonymity or l-diversity, but the client is able to query and modify the original data. By exposing data where possible, the server can perform value-added services such as data analysis not possible with fully encrypted data, while still being unable to violate privacy constraints. Update is a key challenge with this model; naive application of insertion and deletion operations reveals the actual data to the server. This paper shows how data can be safely inserted, deleted, and updated. The key idea is that data is inserted or updated into an encrypted temporary table until enough data is available to safely decrypt, and that sensitive information of deleted tuples is left behind to ensure privacy of both deleted and undeleted individuals. This approach is proven effective in maintaining the privacy constraint against an adversarial server. The paper also gives empirical results on how much data remains encrypted, and how much the quality of the server's anatomized view of the data changes for various update and delete rates.

Efficient and Accurate Strategies for Differentially-Private Sliding Window Queries

Jianneng Cao, Qian Xiao, Gabriel Ghinita, Ninghui Li, Elisa Bertino and Kian-Lee Tan

Regularly releasing the aggregate statistics about data streams in a privacy-preserving way not only serves valuable commercial and social purposes, but also protects the privacy of individuals. This problem has already been studied under differential privacy, but only for the case of a single continuous query that covers the entire time span, e.g., counting the number of tuples seen so far in the stream. However, most real-world applications are window-based, that is, they are interested in the statistical information about streaming data within a window, instead of the whole unbound stream. Furthermore, a Data Stream Management System (DSMS) may need to answer numerous correlated aggregated queries simultaneously, rather than a single one. To cope with these requirements, we study how to release differentially private answers for a set of sliding window aggregate queries. We propose two solutions, each consisting of query sampling and composition. We first selectively sample a subset of representative sliding window queries from the set of all the submitted ones. The representative queries are answered in a way satisfying differential privacy. For all the non-representative queries, none of them is answered directly. Then, given any query, whether it is a selected one or not in the sampling procedure, we compose its answer from the query results of those representatives. The experimental evaluation shows that our solutions are efficient and effective.

Tuesday, March 19			
16:00-18:00	EDBT Research Session 6		**Potpourri**
Room:	Camino	Chair:	Ulf Leser

An Automatic Physical Design Tool for a Column Store with Clustering

Alexander Rasin and Stan Zdonik

Good database design is typically a very difficult and costly process. As database systems get more complex and as the amount of data under management grows, the stakes increase accordingly. Past research has addressed this problem by trying to build tools that can automatically produce a good database design for a known workload. This includes things like automatic secondary index selection and automatic materialized view selection. While this work has produced important results, new specialized database architectures demand a rethinking of automated design tool algorithms. In this paper, we present results for an automatic design tool that is aimed at column-oriented DBMS's on OLAP workloads. In particular, we have chosen the Vertica DBMS, and have run our experiments on the publicly-available trial version. In this setting, the key problem is selecting proper sort orders and compression schemes for the columns as well as appropriate pre-join views. This paper describes our automatic design algorithms as well as the results of some experiments using it on realistic data sets.

Mining Frequent Serial Episodes Over Uncertain Sequence Data

Li Wan, Ling Chen and Chengqi Zhang

Data uncertainty has posed many unique challenges to nearly all types of data mining tasks, creating a need for uncertain data mining. In this paper, we focus on the particular task of mining probabilistic frequent serial episodes (P-FSEs) from uncertain sequence data, which applies to many real applications including sensor readings as well as customer purchase sequences. We first define the notion of P-FSEs, based on the frequentness probabilities of serial episodes under possible world semantics. To discover P-FSEs over an uncertain sequence, we propose: 1) an exact approach that computes the accurate frequentness probabilities of episodes; 2) an approximate approach that approximates the frequency of episodes using probability models; 3) an optimized approach that efficiently prunes a candidate episode by estimating an upper bound of its frequentness probability using approximation techniques. We conduct extensive experiments to evaluate the performance of the developed data mining algorithms. Our experimental results show that: 1) while existing research demonstrates that approximate approaches are orders of magnitudes faster than exact approaches, for P-FSE mining, the efficiency improvement of the approximate approach over the exact approach is marginal; 2) although it has been recognized that the normal distribution based approximation approach is fairly accurate when the data set is large enough, for P-FSE mining, the binomial distribution based approximation achieves higher accuracy when the the number of episode occurrences is limited; 3) the optimized approach clearly outperforms the other two approaches in terms of the runtime, and achieves very high accuracy

Efficient processing of containment queries on nested sets

Ahmed Ibrahim and George H. L. Fletcher

We study the problem of computing containment queries on sets which can have both atomic and set-valued objects as elements, i.e., nested sets. Containment is a fundamental query pattern with many basic applications. Our study of nested set containment is motivated by the ubiquity of nested data in practice, e.g., in XML and JSON data management, in business and scientific workflow management, and in web analytics. Furthermore, there are to our knowledge no known efficient solutions to computing containment queries on massive collections of nested sets. Our specific contributions in this paper are: (1) we introduce two novel algorithms for efficient evaluation of containment queries on massive collections of nested sets; (2) we study caching and filtering mechanisms to accelerate query processing in the algorithms; (3) we develop extensions to the algorithms to a) compute several related query types and b) accommodate natural variations of the semantics of containment; and, (4) we present analytic and empirical analyses which demonstrate that both algorithms are efficient and scalable.

Inferential Time-Decaying Bloom Filters

Jonathan Dautrich and Chinya Ravishankar

Time-Decaying Bloom Filters are efficient, probabilistic data structures used to answer queries on recently inserted items. As new items are inserted, memory of older items decays. Incorrect query responses incur penalties borne by the application using the filter. Most existing filters may only be tuned to static penalties, and they ignore Bayesian priors and information latent in the filter. We address these issues in an integrated way by converting existing filters into inferential filters. Inferential filters combine latent filter information with Bayesian priors to make query-specific optimal decisions. Our methods are applicable to any Bloom Filter, but we focus on developing inferential time-decaying filters, which support new query types and sliding window queries with varying error penalties. We develop the inferential version of the existing Timing Bloom Filter. Through experiments on real and synthetic datasets, we show that when penalties are query-specific and prior probabilities are known, the inferential Timing Bloom Filter reduces penalties for incorrect responses to sliding-window queries by up to 70%.

Wednesday, March 20			
11:00-12:30	EDBT Research Session 7		**Sensors**
Room:	Maggior Consiglio	Chair:	Qing Zhang

Utility-driven Data Acquisition in Participatory Sensing

Mehdi Riahi, Thanasis Papaioannou, Karl Aberer and Immanuel Trummer

Participatory sensing is becoming a popular data acquisition means for emerging applications. However, as data queries from these applications increase, the sustainability of this platform for multiple applications concurrently is at stake. In this paper, we consider the problem of efficient data acquisition in participatory sensing when queries of different types come from different applications. We effectively deal with the issues related to resource constraints, user privacy, data reliability, and uncontrolled mobility. We formulate the problem as multi-query optimization and propose efficient heuristics for its effective solution for the various query types and mixes. Based on simulations with real and artificial data traces, we found that our heuristic algorithms outperform baseline approaches in a multitude of considered settings.

An RFID and Particle Filter-Based Indoor Spatial Query Evaluation System

Jiao Yu, Wei-Shinn Ku, Min-Te Sun and Hua Lu

People spend a significant amount of time in indoor spaces (e.g., office buildings, subway systems, etc.) in their daily lives. Therefore, it is important to develop efficient indoor spatial query algorithms for supporting various location-based applications. However, indoor spaces differ from outdoor spaces because users have to follow the indoor floor plan for their movements. In addition, positioning in indoor environments is mainly based on sensing devices (e.g., RFID readers) rather than GPS devices. Consequently, we cannot apply existing spatial query evaluation techniques devised for outdoor environments for this new challenge. Because particle filters can be employed to estimate the state of a system that changes over time using a sequence of noisy measurements made on the system. In this research, we propose the particle filter-based location inference method as the basis for evaluating indoor spatial queries with noisy RFID raw data. Furthermore, two novel models, indoor walking graph model and anchor point indexing model, are created for tracking object locations in indoor environments. Based on the inference method and tracking models, we develop innovative indoor range and *k* nearest neighbor (*k*NN) query algorithms. We validate our solution through extensive simulations with real world parameters. Our experimental results show that the proposed algorithms can evaluate indoor spatial queries effectively and efficiently.

A Safe Zone Based Approach for Monitoring Moving Skyline Queries

Muhammad Cheema, Xuemin Lin, Wenjie Zhang and Ying Zhang

Given a set of criterions, an object o dominates another object o' if o is more preferable than o' according to every criterion. A skyline query returns every object that is not dominated by any other object. In this paper, we study the problem of continuously monitoring a moving skyline query where one of the criterions is the distance between the objects and the moving query. We propose a safe zone based approach to address the challenge of efficiently updating the results as the query moves. A safe zone is the area such that the results of a query remain unchanged as long as the query lies inside this area. Hence, the results are required to be updated only when the query leaves its safe zone. Although the main focus of this paper is to present the techniques for Euclidean distance metric, the proposed techniques are applicable to any metric distance (e.g., Manhattan distance, road network distance). We present several non-trivial optimizations and propose an efficient algorithm for safe zone construction. Our experiments demonstrate that the cost of our safe zone based approach is reasonably close to a lower bound cost and is three orders of magnitude lower than the cost of a naive algorithm.

Wednesday, March 20			
14:00-15:30	EDBT Research Session 8		**Graph Querying**
Room:	Maggior Consiglio	Chair:	George Fletcher

Compressed Feature-based Filtering and Verification Approach for Subgraph Search

Karam Gouda and Mosab Hassaan

Subgraph search in graph datasets is an important problem with numerous applications. Many feature-based indexing methods have been proposed for solving this problem. These methods have to index too many features or select some of them in order to get an index with good pruning capabilities. None of these directions can give an effective solution to all graph indexing issues. In this paper, we propose an efficient indexing approach which improves over current feature-based methods, neither by the costly feature selection nor by explicitly indexing a multitude of features. We achieve this by compressing multiple features into one feature with some neighborhood information encoded. Neighborhood is further used to prune unmatched feature occurrences between the query and data graphs, thus cutting down the search space of subgraph matching, which significantly reduce the verification cost. We implement the approach by exhaustively enumerating small paths as features. A novel path-at-a-time verification method that benefits from the occurrences pruning method is introduced. Via an extensive evaluation on both real and synthetic datasets, we show that our approach is effective and scalable, and outperforms state-of-the-art indexing methods.

Efficient Query Answering against Dynamic RDF Databases

François Goasdoué, Ioana Manolescu and Alexandra Roatis

A promising method for efficiently querying RDF data consists of translating SPARQL queries into efficient RDBMS-style operations. However, answering SPARQL queries requires handling RDF reasoning, which must be implemented outside the relational engines that do not support it. We introduce the database (DB) fragment of RDF, going beyond the expressive power of previously studied RDF fragments. We devise novel sound and complete techniques for answering Basic Graph Pattern (BGP) queries within the DB fragment of RDF, exploring the two established approaches for handling RDF semantics, namely reformulation and saturation. In particular, we focus on handling database updates within each approach and propose a method for incrementally maintaining the saturation; updates raise specific difficulties due to the rich RDF semantics. Our techniques are designed to be deployed on top of any RDBMS(-style) engine, and we experimentally study their performance trade-offs.

Efficient Breadth-First Search on Large Graphs with Skewed Degree Distributions

Haichuan Shang and Masaru Kitsuregawa

Many recent large-scale data intensive applications are increasingly demanding efficient graph databases. Distributed graph algorithms, as a core part of practical graph databases, have a wide range of important applications, but have been rarely studied in sufficient detail. These problems are challenging as real graphs are usually extremely large and the intrinsic character of graph data, lacking locality, causes unbalanced computation and communication workloads. In this paper, we explore distributed breadth-first search algorithms with regards to large-scale applications. We propose DPC (Degree-based Partitioning and Communication), a scalable and efficient distributed BFS algorithm which achieves high scalability and performance through novel balancing techniques between computation and communication. In experimental study, we compare our algorithm with two state-of-the-art algorithms under the Graph500 benchmark with a variety of settings. The result shows our algorithm significantly outperforms the existing algorithms under all the settings.

Wednesday, March 20			
16:00-17:30	EDBT Research Session 9		**Social Networks and Semantic Querying**
Room:	Maggior Consiglio	Chair:	Klaus Berberich

CINEMA: Conformity-Aware Greedy Algorithm for Influence Maximization in Online Social Networks

Hui Li, Sourav S Bhowmick and Aixin Sun

Influence maximization (IM) is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence. Despite the progress achieved by state-of-the-art greedy IM techniques, they suffer from two key limitations. Firstly, they are inefficient as they can take days to find seeds in real-world networks containing millions of nodes. Secondly, although extensive research in social psychology suggests that humans will readily conform to the wishes or beliefs of others, surprisingly, existing IM techniques are conformity-unaware. That is, they only utilize an individual's ability to influence another but ignores conformity (a person's inclination to be influenced) of the individuals. In this paper, we propose a novel conformity-aware cascade (C2) model which leverages on the interplay between influence and conformity in obtaining the influence probabilities of nodes from underlying data for estimating influence spreads. We propose a novel greedy algorithm called cinema that generates high quality seed set by exploiting this model. It first partitions the network into a set of non-overlapping subnetworks and for each of these subnetworks it computes the influence and conformity indices of nodes. Each subnetwork is then associated with a COG-sublist which stores the marginal gains of the nodes in the subnetwork in descending order. The node with maximum marginal gain in each COG-sublist is stored in a data structure called MAG-list. These structures are manipulated by cinema to efficiently find the seed set. A key feature of such partitioning-based strategy is that each node's influence computation and updates can be limited to the subnetwork it resides instead of the entire network. Our empirical study with real-world social networks demonstrates that cinema generates superior quality seed set compared to state-of-the-art IM approaches.

Pollux: Towards Scalable Distributed Real-time Search on Microblogs

Liwei Lin, Xiaohui Yu and Nick Koudas

The last few years have witnessed a meteoric rise of microblogging platforms, such as Twitter and Tumblr. The sheer volume of the microblog data and its highly dynamic nature present unique technical challenges for the platforms that provide search services. In particular, the search service must provide real-time response to queries, and continuously update the results as new microblogs are posted. Conventional approaches either cannot keep up with the high update rate, or cannot scale well to handle the large volume of data. We propose Pollux, a system that provides distributed realtime indexing and search service on microblogs. It adopts the distributed stream processing paradigm advocated by the recently developed platforms that are designed for real-time processing of large volume of data, such as Apache S4 and Twitter Storm. Although those open-source platforms have found successful applications in production environments, they lack some critical features required for real-time search. In particular: (1) they only implement partial fault tolerance, and do not provide lossless recovery in the event of a node failure, and (2) they do have have a facility for storing global data, which is necessary in efficiently ranking search results. Addressing those problems, Pollux extends current platforms in two important ways. First, we propose a failover strategy that can ensure high system availability and no data/state loss in the event of a node failure. Second, Pollux adds a global storage facility that supports convenient, efficient, and reliable data storage for shared data. We describe how to apply Pollux to the task of realtime search. We implement Pollux based on Apache S4, and show through extensive experiments on a Twitter dataset that the proposed solutions are effective, and Pollux can achieve excellent scalability.

Semantic Query By Example

Lipyeow Lim, Haixun Wang and Min Wang

Supporting semantic queries in relational databases is essential to many advanced applications. Recently, with the increasing use of ontology in various applications, the need for querying relational data togther with its related ontology has become more urgent. In this paper, we identify two fundamental challenges in this task. First, it is extremely difficult to express queries against graph structured ontology in the relational query language SQL, and second, in many cases where data and its related ontology are complicated, queries are usually not precise, that is, users often have only a vague notion, rather than a clear understanding and definition, of what they query for. To address the two challenges, we introduce a novel method that enables us to support semantic queries in relational databases with ease. Instead of endeavoring to incorporate ontology into relational form and create new language constructs to express such queries, we ask the user to provide a small number of examples that satisfy the query he has in mind. Using these examples as seeds, the system infers the exact query automatically, and the user is therefore shielded from the complexity of interfacing with the ontology. More specifically, our approach consists of three steps. In the first step, the user provides several examples that satisfy the query. In the second step, we use machine learning techniques to mine the semantics of the query from the given examples and related ontologies. Finally, we apply the query semantics on the data to generate the full query result. We also implement an optional active learning mechanism in the process so that we can find the accurate query semantics quickly. Our experiments validate the effectiveness of our approach.

Thursday, March 21			
11:00-12:30	EDBT Research Session 10		**Search and Textual Data**
Room:	Maggior Consiglio	Chair:	Ziyang Liu

Scalable Top-K Spatial Keyword Search

Dongxiang Zhang, Kian-Lee Tan and Anthony K. H. Tung

In this big data era, huge amounts of spatial documents have been generated everyday through various location based services. Top-*k* spatial keyword search is an important approach to exploring useful information from a spatial database. It retrieves *k* documents based on a ranking function that takes into account both textual relevance (similarity between the query and document keywords) and spatial relevance (distance between the query and document locations). Various hybrid indexes have been proposed in recent years which mainly combine the R-tree and the inverted index so that spatial pruning and textual pruning can be executed simultaneously. However, the rapid growth in data volume poses significant challenges to existing methods in terms of the index maintenance cost and query processing time. In this paper, we propose a scalable integrated inverted index, named I^3, which adopts the Quadtree structure to hierarchically partition the data space into cells. The basic unit of I^3 is the *keyword cell*, which captures the spatial locality of a keyword. Moreover, we design a new storage mechanism for efficient retrieval of keyword cell and preserve additional summary information to facilitate pruning. Experiments conducted on real spatial datasets (Twitter and Wikipedia) demonstrate the superiority of I^3 over existing schemes such as IR-tree and S2I in various aspects: it incurs shorter construction time to build the index, it has lower index storage cost, it is order of magnitude faster in updates, and it is highly scalable and answers top-*k* spatial keyword queries efficiently.

Panorama: A Semantic-Aware Application Search Framework

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung and Wilfred Ng

Third-party applications (or commonly referred to the apps) proliferate in the web and mobile platforms in recent years. The tremendous amount of available apps in app marketplaces suggests the necessity of designing effective app search engines. However, existing app search engines typically ignore the latent semantics in the app corpus and thus usually fail to provide high-quality app snippets and effective app rankings for the users. In this paper, we present a novel framework named Panorama to provide independent search results for Android apps with semantic awareness. We first propose the App Topic Model (ATM) to discover the latent semantics from the app corpus. Using the discovered semantics, we tackle two central challenges that are faced by current app search engines: (1) how to generate concise and informative snippets for apps and (2) how to rank the apps effectively with respect to search queries. To handle the first challenge, we propose several new metrics for measuring the quality of the sentences in app description and develop a greedy algorithm with a fixed probability of near-optimal performance for app snippet generation. To handle the second challenge, we propose a variety of new features for app ranking and also design a new type of hybrid index to support efficient Top-K query processing. We conduct extensive experiments on a large scale data collection of Android apps and build an app search engine prototype for human-based performance evaluation. The proposed framework demonstrates superior performance against several strong baselines with respect to different metrics.

Selectivity Estimation for Hybrid Queries over Text-Rich Data Graphs

Andreas Wagner, Veli Bicer and Thanh Tran

Many databases today are text-rich, comprising not only structured, but also textual data. Querying such databases involves predicates matching structured data combined with string predicates featuring textual constraints. Based on selectivity estimates for these predicates, query processing as well as other tasks that can be solved through such queries can be optimized. Existing work on selectivity estimation focuses either on string or on structured query predicates alone. Further, probabilistic models proposed to incorporate dependencies between predicates are focused on the relational setting. In this work, we propose a template-based probabilistic model, which enables selectivity estimation for general graph-structured data. Our probabilistic model allows dependencies between structured data and its text- rich parts to be captured. With this general probabilistic solution, BN+, selectivity estimations can be obtained for queries over text-rich graph-structured data, which may contain structured and string predicates (hybrid queries). In our experiments on real-world data, we show that capturing dependencies between structured and textual data in this way greatly improves the accuracy of selectivity estimates without compromising the efficiency.

Thursday, March 21			
14:00-15:30	EDBT Research Session 11		**Skyline**
Room:	Maggior Consiglio	Chair:	Matteo Magnani

Skyline Probability over Uncertain Preferences

Qing Zhang, Pengjie Ye, Xuemin Lin and Ying Zhang

Skyline analysis is a key in a wide spectrum of real applications involving multi-criteria optimal decision making. Recent years, a considerable amount of research has been contributed on efficient computation of skyline probabilities over uncertain environment. Most studies if not all, unanimously assume uncertainty lies only in attribute values. To the extent of our knowledge, only one study addresses the skyline probability computation problem in scenarios where uncertainty resides in attribute preferences, instead of values. However this study takes a problematic approach by assuming *independent object dominance*, which is not always true in uncertain preference scenarios. Actually even in uncertain value scenarios, this assumption has already been demonstrated as problematic. Motivated by this, we revisit the skyline probability computation over uncertain preferences in this paper. We first show that the problem of skyline probability computation over uncertain preferences is #P-complete. Then we propose efficient exact and approximate algorithms to tackle this problem. While the exact algorithm remains exponential in the worst case, our experiments demonstrate its efficiency in practice. The approximate algorithm can be applied on any data sets and can always achieves ε-approximation by the confidence (1-δ) with time complexity O($d\ n\ \frac{1}{\varepsilon^2}\frac{1}{\delta}$), where *n* is the number of objects and *d* is the dimensionality. The efficiency and effectiveness of our methods are verified by extensive experimental results on real and synthetic data sets.

SkyDiver: A Framework for Skyline Diversification

George Valkanas, Apostolos Papadopoulos, Dimitrios Gunopulos

Skyline queries have attracted considerable attention by the database community during the last decade, due to their applicability in a series of domains. However, most existing works tackle the problem from an efficiency standpoint, i.e., returning the skyline as quickly as possible. The user is then presented with the entire skyline set, which may be in several cases overwhelming, therefore requiring manual inspection to come up with the most informative data points. To overcome this shortcoming, we propose a novel approach in selecting the k most diverse skyline points, i.e., the ones that best capture the different aspects of both the skyline and the dataset they belong to. We present a novel formulation of diversification which, in contrast to previous proposals, is intuitive, because it is based solely on the domination relationships among points. Consequently, additional artificial distance measures (e.g., Lp norms) among skyline points are not required. We present efficient approaches in solving this problem and demonstrate the efficiency and effectiveness of our approach through an extensive experimental evaluation with both real-life and synthetic data sets.

Subspace Global Skyline Query Processing

Mei Bai, Junchang Xin and Guoren Wang

Global skyline, as an important variant of skyline, has been widely applied in multiple criteria decision making, business planning and data mining, while there are no previous studies on the global skyline query in the subspace. Hence in this paper we propose subspace global skyline (SGS) query, which is concerned about global skyline in ad hoc subspace. Firstly, we propose an appropriate index structure RB-tree to rapidly find the initial scan positions of query. Secondly, by making analysis of basic properties of SGS, we propose a single SGS algorithm based on RB-tree (SSRB) to compute SGS points. Then an optimized single SGS algorithm based on RB-tree (OSSRB) is proposed, which can reduce the scan space and improve the computation efficiency in contrast to SSRB. Next, by sharing the scan space of different queries, a multiple SGS algorithm based on RB-tree (MSRB) is proposed to compute multiple SGS (MSGS). Finally, the performances of our proposed algorithms are verified through a large number of simulation experiments

Thursday, March 21			
14:00-15:30	EDBT Research Session 12		**Database as a Service**
Room:	Camino	Chair:	Murat Kantarcioglu

SWORD: Scalable Workload-Aware Data Placement for Transactional Workloads

Abdul Quamar, K.Ashwin Kumar, Amol Deshpande

In this paper, we address the problem of transparently scaling out transactional (OLTP) workloads on relational databases, to support *database-as-a-service* in cloud computing environment. The primary challenges in supporting such workloads include choosing how to *partition* the data across a large number of machines, minimizing the number of *distributed transactions*, providing high data *availability*, and tolerating *failures* gracefully. Capturing and modeling the transactional workload over a period of time, and then exploiting that information for data placement and replication has been shown to provide significant benefits in performance, both in terms of transaction latencies and overall throughput. However, such workload-aware data placement approaches can incur very high overheads, and further, may perform worse than naive approaches if the workload changes. In this work, we propose SWORD, a scalable workload-aware data partitioning and placement approach for OLTP workloads, that incorporates a suite of novel techniques to significantly reduce the overheads incurred both during the initial placement, and during query execution at runtime. We model the workload as a hypergraph over the data items, and propose using a *hypergraph compression* technique to reduce the overheads of partitioning. To deal with workload changes, we propose an incremental data repartitioning technique that modifies data placement in small steps without resorting to complete workload repartitioning. We have built a workload-aware *active replication* mechanism in SWORD to increase availability and enable load balancing. We propose the use of *fine-grained quorums* defined at the level of *groups of tuples* to control the cost of distributed updates, improve throughput, and provide adaptability to different workloads. To our knowledge, SWORD is the first system that uses fine-grained quorums in this context. The results of our experimental evaluation on SWORD deployed on an Amazon EC2 cluster show that our techniques result in orders-of-magnitude reductions in the partitioning and book-keeping overheads, and improve tolerance to failures and workload changes; we also show that choosing quorums based on the query access patterns enables us to better handle query workloads with different read and write access patterns.

PMAX: Tenant Placement in Multitenant Databases for Profit Maximization

Ziyang Liu, Hakan Hacigumus, Hyun Moon, Yun Chi and Wang-Pin Hsiung

There has been a great interest in exploiting the cloud as a platform for database as a service. As with other cloud-based services, database services may enjoy cost efficiency through consolidation: hosting multiple databases within a single physical server. Aggressive consolidation, however, may hurt the service quality, leading to SLA violation penalty, which in turn reduces the total business profit, called SLA profit. In this paper, we consider the problem of tenant placement in the cloud for SLA profit maximization, which, as will be shown in the paper, is strongly NP-hard. We propose SLA profit-aware solutions for database tenant placement based on our model for expected penalty computation for multitenant servers. Specifically, we present two approximation algorithms which have constant approximation ratios, and we further discuss improving the quality of tenant placement using a dynamic programming algorithm. Extensive experiments based on TPC-W workload verified the excellent performance of the proposed approaches.

Elastic Online Analytical Processing on RAMCloud

Christian Tinnefeld, Donald Kossmann, Martin Grund, Joos-Hendrik Boese, Frank Renkes, Vishal Sikka and Hasso Plattner

A shared-nothing architecture is state-of-the-art for deploying a distributed analytical in-memory database management system: it preserves the in-memory performance advantage by processing data locally on each node but is difficult to scale out. Modern switched fabric communication links such as InfiniBand narrow the performance gap between local and remote DRAM data access to a single order of magnitude. Based on these premises, we introduce a distributed in-memory database architecture that separates the query execution engine and data access: this enables a) the usage of a large-scale DRAM- based storage system such as Stanford's RAMCloud and b) the push-down of bandwidth-intensive database operators into the storage system. We address the resulting challenges such as finding the optimal operator execution strategy and partitioning scheme. We demonstrate that such an architecture delivers both: the elasticity of a shared-storage approach and the performance characteristics of operating on local DRAM.

Thursday, March 21			
16:00-17:30	EDBT Research Session 13		**Preference Queries**
Room:	Maggior Consiglio	Chair:	Mohamed Y. Eltabakh

Skyline Queries in Crowd-Enabled Databases

Christoph Lofi, Kinda El Maarry, Wolf-Tilo Balke

Skyline queries are a well-established technique for database query personalization and are widely acclaimed for their intuitive query formulation mechanisms. However, when operating on incomplete datasets, skylines queries are severely hampered and often have to resort to highly error-prone heuristics. Unfortunately, incomplete datasets are a frequent phenomenon, especially when datasets are generated automatically using various information extraction or information integration approaches. Here, the recent trend of crowd-enabled databases promises a powerful solution: during query execution, some database operators can be dynamically outsourced to human workers in exchange for monetary compensation, therefore enabling the elicitation of missing values during runtime. Unfortunately, this powerful feature heavily impacts query response times and (monetary) execution costs. In this paper, we present an innovative hybrid approach combining dynamic crowd-sourcing with heuristic techniques in order to overcome current limitations. We will show that by assessing the individual risk a tuple poses with respect to the overall result quality, crowd-sourcing efforts for eliciting missing values can be narrowly focused on only those tuples that may degenerate the expected quality most strongly. This leads to an algorithm for computing skyline sets on incomplete data with maximum result quality, while optimizing crowd-sourcing costs.

From stars to galaxies: skyline queries on aggregate data

Matteo Magnani and Ira Assent

The skyline operator extracts relevant records from multidimensional databases according to multiple criteria. This operator has received a lot of attention because of its ability to identify the best records in a database without requiring to specify complex parameters like the relative importance of each criterion. However, it has only been defined with respect to single records, while one fundamental functionality of the database query language is aggregation, enabling operations over sets of records. In this paper we introduce aggregate skylines, where the skyline works as a filtering predicate on sets of records. This operator can be used to express queries in the form: "return the best groups depending on the features of their elements", and thus provides a powerful combination of grouping and skyline functionality. We define semantics for aggregate skylines based on a sound theoretical framework and study its computational complexity. We propose efficient algorithms to implement this operator and test them on real and synthetic data, showing that they outperform a direct SQL implementation of up to two orders of magnitude.

Efficient Top-k Query Answering using Cached Views

Min Xie, Laks Lakshmanan and Peter Wood

Top-k query processing has recently received a significant amount of attention due to its wide application in information retrieval, multimedia search and recommendation generation. In this work, we consider the problem of how to efficiently answer a top-k query by using previously cached query results. While there has been some previous work on this problem, existing algorithms suffer from either limited scope or lack of scalability. In this paper, we propose two novel algorithms for handling this problem. The first algorithm LPTA$^+$ provides significantly improved efficiency compared to the state-of-the-art LPTA algorithm by reducing the number of expensive linear programming problems that need to be solved. The second algorithm we propose leverages a standard space partition-based index structure in order to avoid many of the drawbacks of LPTA-based algorithms, thereby further improving the efficiency of query processing. Through extensive experiments on various datasets, we demonstrate that our algorithms significantly outperform the state of the art.

Thursday, March 21			
16:00-17:30	EDBT Research Session 14		**Stream Query Processing**
Room:	Camino	Chair:	Torben Bach Pedersen

Enhanced Stream Processing in a DBMS Kernel

Erietta Liarou, Stratos Idreos, Stefan Manegold and Martin Kersten

Continuous query processing has emerged as a promising query processing paradigm with numerous applications. A recent development is the need to handle both streaming queries and typical one-time queries in the same application. For example, data warehousing can greatly benefit from the integration of stream semantics, i.e., on-line analysis of incoming data and combination with existing data. This is especially useful to provide low latency in data intensive analysis in big data warehouses that are augmented with new data on a daily basis. However, state-of-the-art database technology cannot handle streams efficiently due to their “continuous” nature. At the same time, state-of-the-art stream technology is purely focused on stream applications. The research efforts are mostly geared towards the creation of specialized stream management systems built with a different philosophy than a DBMS. The drawback of this approach is the limited opportunities to exploit successful past data processing technology, e.g., query optimization techniques. For this new problem we need to combine the best of both worlds. Here we take a completely different route by designing a stream engine on top of an existing relational database kernel. This includes reuse of both its storage/execution engine and its optimizer infrastructure. The major challenge then becomes the efficient support for specialized stream features. This paper, focuses on incremental window-based processing, arguably the most crucial stream-specific requirement. In order to maintain and reuse the generic storage and execution model of the DBMS, we elevate the problem at the query plan level. Proper optimizer rules, scheduling and intermediate result caching and reuse, allow us to modify the DBMS query plans for efficient incremental processing. We describe in detail the new approach and we demonstrate efficient performance even against specialized stream engines, especially when scalability becomes a crucial factor.

Probabilistic Inference of Object Identifications for Event Stream Analytics

Di Wang, Elke Rundensteiner, Han Wang and Richard Ellison

Recent years have witnessed the emergence of real-time object monitoring applications driven by the explosion of small inexpensive sensors. In many real-world applications, not all sensed events carry the identification of the object whose action they report on, so called "non-ID-ed" events. Reasons range from heterogeneous sensing devices to human's choosing to conceal their identifications. Such non-ID-ed events prevent us from performing object-based analytics, such as tracking, alerting and pattern matching. We propose a probabilistic inference framework, called FISS, to tackle this problem by inferring the missing object identification associated with an event. Specifically, as a foundation we design a time-varying graphic model to capture correspondences between sensed events and objects. Upon this formal model, we elaborate how to adapt the Forward-backward (FB) inference algorithm to continuously infer probabilistic identifications for non-ID-ed events. However, we demonstrate that FB is neither scalable nor efficient over event streams. To overcome this deficiency, we propose a suite of strategies for optimizing its performance, including the selective smoothing technique that significantly reduces the number of random variables that need to be smoothed, and the finish-flag mechanism that enables early termination of backward computations. Our experimental results, using large-volume streams of a real-world healthcare application, demonstrate the accuracy, efficiency, and scalability of FISS. Especially FISS achieves on average 15x higher throughput than our basic FB inference.

High Performance Complex Event Processing using Continuous Sliding Views

Medhabi Ray, Elke Rundensteiner, Mo Liu, Chetan Gupta, Song Wang and Ismail Ari

Complex Event Processing (CEP) has become a paradigm of choice for the development of monitoring and reactive applications in areas ranging from financial services to RFID-based inventory management. While novel powerful languages are being proposed to express complex nested event expressions, efficient techniques for computing this broad class of nested CEP queries are lacking. In this paper we propose the concept of "Continuous Sliding Views" over streams as a methodology for efficiently computing nested CEP queries by caching results of sub-expressions and sharing them across query invocations as the query expression slides over the stream. We design a family of continuous sliding view strategies, along with targeted algorithms for incrementally loading, purging and exploiting these views for nested CEP query optimization. The first strategy aggressively pre-computes continuous views over each sliding window while the second strategy instead applies a passive method of piggy-backing the capture of results as part of query processing. The latter is achieved by augmenting our continuous sliding view structures with temporal validity ranges. Lastly, we devise a strategy that trades-off the prior two methods by selecting a more balanced view granularity. Our experimental study using real-world stock trade data evaluates the performance of our proposed strategies for a diversity of query types and parameters. Our strategies not only out-perform the state-of-the-art nested CEP processing technique, but are also applicable in a wide range of scenarios where state-of-the-art query rewriting techniques are not applicable.

Thursday, March 21			
16:00-17:30	EDBT Research Session 15		**Data Integration**
Room:	Munizioniere0	Chair:	Melanie Herschel

Data Exchange with Arithmetic Operations

Balder Ten Cate, Phokion Kolaitis and Walied Othman

Data exchange is the problem of transforming data structured under a source schema into data structured under a target schema, taking into account structural relationships between the two schemas, which are described by a schema mapping. Existing schema-mapping languages lack the ability to express arithmetic operations, such as addition and multiplication, which naturally arise in data warehousing, ETL applications, and applications involving scientific data. We initiate the study of data exchange for arithmetic schema mappings, that is, schema mappings specified by source-to-target dependencies and target dependencies that may include arithmetic formulas interpreted over the algebraic real numbers (we restrict attention to algebraic real numbers to maintain finite presentability, and the ability to study questions of computational complexity). We show that, for arithmetic schema mappings without target dependencies, the existence-of-solutions problem can be solved in polynomial time, and, if a solution exists, then a universal solution (suitablydefined) exists and can be computed in polynomial time. In the case of arithmetic schema mappings with a weakly acyclic set of target dependencies, a universal solution may not exist, but a finite universal basis exists (if a solution exists) and can be computed in polynomial space. The existence-of-solutions problem turns out to be NP-hard, and solvable in PSPACE. In fact, we show it is \ER-complete, which means that it has the same complexity as the decision problem for the existential theory of the real numbers, or, equivalently, the problem of deciding whether or not a quantifier-free arithmetic formula has a solution over the real numbers. If we allow only linear arithmetic formulas in the schema mapping and in the query, interpreted over the rational numbers, then the existence-of-solutions problem is NP-complete. We obtain analogous complexity results for the data complexity of computing the certain answers of arithmetic conjunctive queries and linear arithmetic conjunctive queries.

HIL: A High-Level Scripting Language for Entity Integration

Mauricio Hernandez, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa and Ryan Wisnesky

We introduce HIL, a high-level scripting language for entity resolution and integration. HIL aims at providing the core logic for complex data processing flows that aggregate facts from large collections of structured or unstructured data into clean, unified entities. Such flows typically include many stages of processing that start from the outcome of information extraction and continue with entity resolution, mapping and fusion. A HIL program captures the overall integration flow through a combination of SQL-like rules that link, map, fuse and aggregate entities. A salient feature of HIL is the use of logical indexes in its data model to facilitate the modular construction and aggregation of complex entities. Another feature is the presence of a flexible, open type system that allows HIL to handle input data that is irregular, sparse or partially known. As a result, HIL can accurately express complex integration tasks, while still being high-level and focused on the logical entities (rather than the physical operations). Compilation algorithms translate the HIL specification into efficient run-time queries that can execute in parallel on Hadoop. We show how our framework is applied to a real-world integration of entities in the financial domain, based on public filings archived by the U.S. Securities and Exchange Commission (SEC). Furthermore, we apply HIL on a larger-scale integration scenario that performs fusion of data from hundreds of millions of Twitter messages into tens of millions of structured entities.

Optimizing Query Rewriting in Ontology-Based Data Access

Floriana Di Pinto, Domenico Lembo, Maurizio Lenzerini, Riccardo Mancini, Antonella Poggi, Riccardo Rosati, Marco Ruzzi and Domenico Fabio Savo

In ontology-based data access (OBDA), an ontology is connected to autonomous, and generally pre-existing, data repositories through mappings, so as to provide a high-level, conceptual view over such data. User queries are posed over the ontology, and answers are computed by reasoning both on the ontology and the mappings. Query answering in OBDA systems is typically performed through a query rewriting approach which is divided into two steps: (i) the query is rewritten with respect to the ontology (ontology rewriting of the query); (ii) the query thus obtained is then reformulated over the database schema using the mapping assertions (mapping rewriting of the query). In this paper we present a new approach to the optimization of query rewriting in OBDA. The key ideas of our approach are the usage of inclusions between mapping views and the usage of perfect mappings, which allow us to drastically lower the combinatorial explosion due to mapping rewriting. These ideas are formalized in PerfectMap, an algorithm for OBDA query rewriting. We have experimented PerfectMap in a real-world OBDA scenario: our experimental results clearly show that, in such a scenario, the optimizations of PerfectMap are crucial to effectively perform query answering.

EDBT Industry & Application Sessions

Wednesday, March 20			
11:00-12:30	EDBT Industry & Applications Session 1		**Systems and Tools**
Room:	Munizioniere0		

Temporal Query Processing in Teradata

Mohammed Al-Kateb, Ahmad Ghazal, Alain Crolotte, Ramesh Bhashyam, Jaiprakash Chimanchode and Sai Pavan Pakala

The importance of temporal data management is evident by the temporal features recently released in major commercial database systems. In Teradata, the temporal feature is based on the TSQL2 specification. In this paper, we present Teradata's implementation approach for temporal query processing. There are two common approaches to support temporal query processing in a database engine. One is through functional query rewrites to convert a temporal query to a semantically-equivalent non-temporal counterpart, mostly by adding time-based constraints. The other is a native support that implements temporal database operations such as scans and joins directly in the DBMS internals. These approaches have competing pros and cons. The rewrite approach is generally simpler to implement. But it adds a structural complexity to original query, which can pose a potential challenge to query optimizer and cause it to generate sub-optimal plans. A native support is expected to perform better. But it usually involves a higher cost of implementation, maintenance, and extension. We discuss why and describe how Teradata adopted the rewrite approach. In addition, we present an evaluation of our approach through a performance study conducted on a variation of the TPC-H benchmark with temporal tables and queries.

Near Real-Time Analytics with IBM DB2 Analytics Accelerator

Daniel Martin, Oliver Koeth, Iliyana Ivanova, Johannes Kern

The IBM DB2 Analytics Accelerator (IDAA) implements the vision of a universal relational DBMS that processes OLTP and analytical-type queries in a single system, but on two fundamentally different query engines. Based on heuristics in DB2 for z/OS, the DB2 optimizer decides if a query should be executed by “mainline” DB2 or if it is beneficial to offload it to the attached IBM DB2 Analytics Accelerator that operates on copies of the DB2 tables. In this paper, we introduce the “incremental update” functionality of IDAA that keeps these copy tables in sync by employing replication technology that monitors the DB2 transaction log and asynchronously applies the changes to IDAA. This enables near real-time analytics over online data, effectively marrying traditionally separated OLTP and data warehouse environments. With IDAA, analytic queries can access data that is constantly refreshed in contrast to traditional warehouses that are updated on a daily or even weekly basis. Without any changes to the applications and without the need to introduce cross-system ETL flows, an existing operational data store can be used for data warehousing as well. The analytic query performance provided by IDAA makes it possible to execute reports directly against the transactional schema, thus avoiding the need for costly design and maintenance of a separate reporting schema. Additionally, the Accelerator shields DB2 for z/OS as the transactional system from performance degradation caused by the analytical workload and the replication component synchronizes all data changes in near-real time. We present the architecture of the integrated replication component of IDAA and discuss design decisions that we made when combining the different technologies as well as performance characteristics of the resulting system.

AppSleuth: a Tool for Database Tuning at the Application Level

Wei Cao and Dennis Shasha

Excellent work ([1]-[6]) has shown that memory management and transaction concurrency levels can often be tuned automatically by the database management systems. Other excellent work ([7]]-[14]) has shown how to use the optimizer to do automatic physical design or to make the optimizer itself more self-adaptive ([15]-[17]). Our performance tuning experience across

various industries (finance, gaming, data warehouses, and travel) has shown that enormous additional tuning benefits (sometimes amounting to orders of magnitude) can come from reengineering application code and table design. The question is: can a tool help in this effort? We believe so. We present a tool called AppSleuth that parses application code and the tracing log for two popular database management systems in order to lead a competent tuner to the hot spots in an application. This paper discusses (i) representative application "delinquent design patterns", (ii) an application code parser to find them, (iii) a log parser to identify the patterns that are critical, and (iv) a display to give a global view of the issue. We present an extended sanitized case study from a real travel application to show the results of the tool at different stages of a tuning engagement, yielding a 300 fold improvement. This is the first tool of its kind that we know of.

Wednesday, March 20			
14:00-15:30	EDBT Industry & Applications Session 2		**Big Data**
Room:	Munizioniere0		

Cost Exploration of Data Sharings in the Cloud

Samer Al-Kiswany, Hakan Hacigumus, Ziyang Liu and Jagan Sankaranarayanan

Enabling data sharing among mobile apps hosted in the same cloud infrastructure can provide a competitive advantage to the mobile apps by giving them access to rich information as well as increasing the revenue to the cloud provider. We introduce a costing tool that allows application owners (i.e., consumers) and the cloud service provider to assess the cost of a desired data sharing. The costing tool enables the consumers to effectively explore the cost space by choosing between alternative configurations of varying data qualities, specified by the staleness and the accuracy of the data sharing. In other words, staleness and accuracy requirements on the data sharing are used as levers for controlling costs. These capabilities are implemented in a What-if analysis tool, which has been integrated with a large data-sharing platform. We conducted extensive experiments on the integrated platform with a sharing ecosystem created around Twitter data and show the effectiveness of the results produced by the What-if tool.

A Performance Comparison of Parallel DBMS and MapReduce on Large-Scale Text Analytics

Fei Chen and Meichun Hsu

Text analytics has become increasingly important with the rapid growth of text data. Particularly, *information extraction* (IE), which extracts structured data from text, has received significant attention. Unfortunately, IE is often computationallyintensive. To address this issue, MapReduce has been used for large scale IE. Recently, there are emerging efforts from both academia and industry on pushing IE inside DBMS. This leads to an interesting and important question: Given that both MapReduce and parallel DBMS are for large scale analytics, which platform is a better choice for large scale IE? In this paper, we propose a benchmark to systematically study the performance of both platforms for large scale IE tasks. The benchmark includes both statistical learning based and rule based IE programs, which have been extensively used in real-world IE tasks. We show how to express these programs on both platforms and conduct experiments on real-world datasets. Our results show that parallel DBMS is a viable alternative for large scale IE.

Sparkler: Supporting Large-Scale Matrix Factorization

Boduo Li, Sandeep Tata and Yannis Sismannis

Low-rank matrix factorization has recently been applied with great success on matrix completion problems for applications like recommendation systems, link predictions for social networks, and click prediction for web search. However, as this approach is applied to increasingly larger datasets, such as those encountered in web-scale recommender systems like Netflix and Pandora, the data management aspects quickly become a challenging roadblock to solving this problem at scale. In this paper, we introduce a system called Sparkler to solve such large instances of low rank matrix factorizations. Sparkler extends Spark, an existing platform for running parallel iterative algorithms on datasets that fit in the aggregate main memory of a cluster. Sparkler supports distributed stochastic gradient descent as an approach to solving the factorization problem -- an iterative technique that has been shown to perform very well in practice. We identify the shortfalls of Spark in solving large matrix factorization problems, especially when running on the cloud and solve this by introducing a novel abstraction called “Carousel

Maps" (CMs). CMs are well suited to storing large matrices in the aggregate memory of a cluster and can efficiently support the operations performed on them during distributed stochastic gradient descent. We describe the design, implementation, and the use of CMs in Sparkler programs. Through a variety of experiments, we demonstrate that Sparkler is faster than Spark by 4x to 21x, with bigger advantages for larger problems. Equally importantly, we show that this can be done without imposing any changes to the ease of programming. We argue that Sparkler provides a convenient and efficient platform for solving matrix factorization problems on very large datasets.

Wednesday, March 20			
16:00-17:30	EDBT Industry & Applications Session 3		**Applications**
Room:	Munizioniere0		

Choosing the Right Crowd: Expert Finding in Social Networks

Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri and Giuliano Vesci

Expert selection is an important aspect of many Web applications, e.g., when they aim at matching contents, tasks or advertisement based on user profiles, possibly retrieved from social networks. This paper focuses on selecting experts within the population of social networks, according to the information about the social activities of their users. We consider the following problem: given an expertise need (expressed for instance as a natural language query) and a set of social network members, who are the most knowledgeable people for addressing that need? We considers social networks both as a source of expertise information and as a route to reach expert users, and define models and methods for valuating people's expertise by considering their profiles and by tracing their activities in social networks. For matching queries to social resources, we use both text analysis and semantic annotation. An extensive set of experiments shows that the analysis of social activities, social relationships, and socially shared contents helps improving the effectiveness of an expert Finding system.

Real-Time Wildfire Monitoring Using Scientific Database and Linked Data Technologies

Manolis Koubarakis, Charalambos Kontoes, Stefan Manegold, Manos Karpathiotakis, Kostis Kyzirakos, Konstantina Bereta, George Garbis, Charalampos Nikolaou, Dimitrios Michail, Ioannis Papoutsis, Themistoklis Herekakis, Milena Ivanova, Ying Zhang, Holger Pirk, Martin Kersten, Kallirroi Dogani, Stella Giannakopoulou and Panayiotis Smeros

We present a real-time wildfire monitoring service that exploits satellite images and linked geospatial data to detect hotspots and monitor the evolution of fire fronts. The service makes heavy use of scientific database technologies (array databases, SciQL, data vaults) and linked data technologies (ontologies, linked geospatial data, stSPARQL) and is implemented on top of MonetDB and Strabon. The service is now operational at the National Observatory of Athens and has been used during the previous summer by emergency managers monitoring wildfires in Greece.

Efficient Multifaceted Screening of Job Applicants

Sameep Mehta, Rakesh Pimplikar, Amit Singh, Lav Varshney and Karthik Viswesariah

Built on top of human resources management databases within the enterprise, we present a decision support system for managing and optimizing screening activities during the hiring process in a large organization. The basic idea is to prioritize the efforts of human resource practitioners to focus on candidates that are likely of high quality, that are likely to accept a job offer if made one, and that are likely to remain with the organization for the long term. To do so, the system first individually ranks candidates along several dimensions using a keyword matching algorithm and several bipartite ranking algorithms with univariate loss trained on historical actions. Next, individual rankings are aggregated to derive a single list that is presented to the recruitment team through an interactive portal. The portal supports multiple filters that facilitate effective identification of candidates. We demonstrate the usefulness of our system on data collected from a large organization over several years with business value metrics, showing greater hiring yield with less interviews. Similarly, using historical pre-hire data we demonstrate accurate identification of candidates that will have quickly left the organization.

Thursday, March 21			
11:00-12:30	EDBT Industry & Applications Session 4		**Potpourri**
Room:	Camino		

EXLEngine: executable schema mappings for statistical data processing

Paolo Atzeni, Bellomarini Luigi and Bugiotti Francesca

Data processing is the core of any statistical information sys- tem. Statisticians would be interested in specifying trans- formations and manipulations of data at a high level, in terms of entities of statistical models such as time series. We illustrate here an experience at the Bank of Italy where (i) a language, EXL, has been defined for the declarative specification of statistical programs, (ii) an approach for the translation of EXL code into executables in various target systems has been developed, and (iii) a concrete implementation, EXLEngine, has been carried out. The approach leverages on schema mappings as an intermediate specification step, in order to facilitate the translation from EXL towards several target systems.

HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm

Stefan Heule, Marc Nunkesser and Alexander Hall

Cardinality estimation has a wide range of applications and is of particular importance in database systems. Various algorithms have been proposed in the past, and the HyperLogLog algorithm is one of them. In this paper, we present a series of improvements to this algorithm that reduce its memory requirements and significantly increase its accuracy for an important range of cardinalities. We have implemented our proposed algorithm for a system at Google and evaluated it empirically, comparing it to the original HyperLogLog algorithm. Like HyperLogLog, our improved algorithm parallelizes perfectly and computes the cardinality estimate in a single pass.

Entity Discovery and Annotation in Tables

Gianluca Quercini and Chantal Reynaud

The Web is rich of tables (e.g., HTML tables, spreadsheets, Google Fusion Tables) that host a considerable wealth of high-quality relational data. Unlike unstructured texts, tables usually favour the automatic extraction of data because of their regular structure and properties. The data extraction is usually complemented by the annotation of the table, which determines its semantics by identifying a type for each column, the relations between columns, if any, and the entities that occur in each cell. In this paper, we focus on the problem of discovering and annotating entities in tables. More specifically, we describe an algorithm that identifies the rows of a table that contain information on entities of specific types (e.g., restaurant, museum, theatre) derived from an ontology and determines the cells in which the names of those entities occur. We implemented this algorithm while developing a faceted browser over a repository of RDF data on points of interest of cities that we extracted from Google Fusion Tables. We claim that our algorithm complements the existing approaches, which annotate entities in a table based on a pre-compiled reference catalogue that lists the types of a finite set of entities; as a result, they are unable to discover and annotate entities that do not belong to the reference catalogue. Instead, we train our algorithm to look for information on previously unseen entities on the Web so as to annotate them with the correct type.

EDBT Demo Sessions

Tuesday, March 19 and Wednesday, March 20			
14:00-15:30	EDBT Demo Session 1		
Room:	Munizioniere1 + Munizioniere2		

iPark: Identifying Parking Spaces from Trajectories

Bin Yang, Nicolas Fantini and Christian S. Jensen

A wide variety of desktop and mobile Web applications involve geo-tagged content, e.g., photos and (micro-) blog postings. Such content, often called User Generated Geo-Content (UGGC), plays an increasingly important role in many applications. However, a great demand also exists for “core” UGGC where the geospatial aspect is not just a tag on other content, but *is* the primary content, e.g., a city street map with up-to-date road construction data. Along these lines, the iPark system aims to turn volumes of GPS data obtained from vehicles into information about the locations of parking spaces, thus enabling effective parking search applications. In particular, we demonstrate how iPark helps ordinary users annotate an existing digital map with two types of parking, on-street parking and parking zones, based on vehicular tracking data.

Limosa: A System for Geographic User Interest Analysis in Twitter

Jan Vosecky, Di Jiang and Wilfred Ng

In this demonstration, we present Limosa, an interactive system for visualization of geographic interests of users in Twitter. The system supports the modeling of comprehensive geographic characteristics of topics discussed in microblogs, both with respect to locations that postings originate from and also locations mentioned within the posting itself. Limosa then provides visualizations of geographic user interests, including the geographic scope of topics, terms, or the semantics associated with specific locations. Using a variety of recommendation strategies for exploration, we show that Limosa provides effective news and user recommendations.

Accelerating Spatial Range Queries

Alexandros Stougiannis, Thomas Heinis, Farhan Tauheed, Anastasia Ailamaki

Scientists in general and neuroscientists in particular increasingly use computational tools to build and simulate spatial models of the phenomena they are studying. The spatial models they build are becoming increasingly detailed and the spatial data involved in these simulations thus is unprecendent in size, making the efficient management of this data crucial. A specific problem in analyzing these spatial models of increasing detail is the scalable execution of range queries. State-of-the-art approaches like the R-Tree and related approaches do not perform well on today's models and will not scale for more detailed, future models. Their problem is that with the increasing level of detail of the model, also the overlap in the tree structure increases, ultimately slowing down query execution. In this demonstration, we will showcase FLAT, a new approach to efficiently execute spatial range queries on increasingly detailed (denser) models. FLAT's query execution strategy effectively allows to decouple the query execution time from the level of detail/density of the data set, thereby ensuring efficient query execution. At the core of the demonstration will thus be the visualization of the novel query execution strategy of FLAT and we will contrast it with a visualization of the query execution of an R-Tree.

An Efficient Layout Method for a Large Collection of Geographic Data Entries

Sarana Nutanong, Marco Adelfio and Hanan Samet

Many spatial applications require the ability to display locations of geographic data entries on an online map. For example, an online photo-sharing service may wish to display photos (as thumbnails) according to where they were taken. Since displaying geographic data entries as thumbnails or icons on a map requires some amount of space, displayed entries can overlap each other. As a result, we may wish to discard less popular or older entries (based on a given measure of importance) so that these more popular or newer entries become more distinct. A straightforward solution is to apply a spatial database extension such as PostGIS (i) to retrieve entries within a given display window; (ii) to discard entries in proximity of a more important one. In this paper, we demonstrate our method for efficiently selecting distinct entries from a large geographical point set. Specifically, our demonstration software presents a voting system built upon an ensemble of interrelated indexes, which is the main novelty of our query processing method. This allows us to efficiently determine the degree of distinctiveness of all entries within a query window using simple index traversal operations rather than expensive spatial operations. The effectiveness of our method in comparison to a traditional spatial query is shown by our experimental results using a real dataset of over 9 million locations. These experimental results show that our proposed method is capable of consistently producing subsecond response times, while the spatial query-based method takes more than 10 seconds on average in a low spatial selectivity setting.

In the Mood4: Recommendation by Examples

Rubi Boim and Tova Milo

Traditional recommender systems generate personalized recommendations based on a profile that they create for each user. We argue here that such profiles are often too coarse to capture the current user's state of mind and desire. For example, a serious user that usually prefers documentary features may, at the end of a long and tiring conference, be in the mood for a lighter entertaining movie, not captured by her usual profile. As communicating one's state of mind to a system in (key)words may be difficult, we present in this demo MOOD4 - a novel plug-in for recommender systems, which allows users to describe their current desire/mood through examples. MOOD4 utilizes the user's examples to refine the recommendations generated by a given recommender system, considering several, possibly competing, desired properties of the recommended items set (rating, diversity, coverage). The system uses a novel algorithm, based on a simple geometric representation of the items, that allows for efficient processing and the generation of suitable recommendations even in the absence of semantic information.

YmalDB: A Result-Driven Recommendation System for Databases

Marina Drosou and Evaggelia Pitoura

To assist users in database exploration, we present the YmalDB system, a database system enhanced with a recommendation functionality. Along with the results of each user query, YmalDB computes and presents to the users additional results, called Ymal (i.e., "You May Also Like") results, that are highly related with the results of their original query. Such results are computed using the most interesting sets of attribute values, called faSets, that appear either in the results of the original query or in the results of an appropriately expanded one. The interestingness of a faSet is based on its frequency both in the query result and in the database.

Hive Open Research Network Platform

Jung Hyun Kim, Xilun Chen, K. Selcuk Candan and Maria Luisa Sapino

Did you ever return back from a conference, having met a lot of interesting folks, listened to many inspiring talks, or having your presentation welcomed with a barrage of (of course, constructive!) questions, wishing if only you managed to take record of all these during the event? We are developing the Hive Open Research Network, a social platform for fostering scientific interactions and reducing friction in scientific exchanges and the underlying integrated services supporting content personalization, preview, and social/scientific recommendations. Hive is a conference centric, but cross-conference platform, where researchers can seed and expand their research networks, keep track of the technical research sessions they are attending, meet new colleagues, share their ideas, ask questions, give and receive comments, or simply keep and/or view records of interactions at a conference they have attended (or wanted to attend, but missed due to other commitments). In its core, Hive leverages dynamically evolving knowledge structures, including user connections, concept maps, co-authorship networks, content from papers and presentations, and contextual knowledge to create and to promote networks of peers. These peer networks support each other explicitly through direct communication or indirectly through collaborative filtering.Hive provides the following online integrated services: a) understanding the personal activity context through access patterns and analysis of user supplied content, b) context-aware resource discovery, including search, presentation, and exploration support within the scientific knowledge structures, and c) peer discovery, and peer driven resource and knowledge sharing and collaborative recommendations.

Tuesday, March 19 and Wednesday, March 20			
16:00-17:30	EDBT Demo Session 2		
Room:	Munizioniere1 + Munizioniere2		

CrowdSeed: Query Processing on Microblogs

Zhou Zhao, Wilfred Ng and Zhijun Zhang

Databases often offer poor answers with respect to judgemental queries such as asking the best among the movies shown in recent months. Processing such queries requires human input for providing missing information in order to clarify uncertainty or inconsistency in queries. Nowadays, it is common to see people seeking answers on micro-blogs through asking or sharing questions with their friends. This can be easily done via smart phones, which diffuse a question to a large number of users through message propagation in microblogs. This trend is important and known as CrowdSearch. Due to conflicting attitudes among crowds, the majority vote is employed as a crowd-wisdom aggregation schema. In this demo, we show the problem of minimizing the monetary cost of a crowdsourced query, given the specified expected accuracy of the aggregated answer?. We present CrowdSeed, a database system that automatically integrates human input for processing queries imposed on microblogs. We demonstrate the effectiveness and efficiency of our system using real world data, as well as presenting interesting results from a game called “Who is in the CrowdSeed?”. Demonstration video in http://www.youtube.com/watch?v=uik1d98ZS2w.

Tuning in Action

Wei Cao and Dennis Shasha

Imagine that your database has all the right indexes. Its buffer manager has been tuned to give a high hit ratio, the buffer fits in RAM, and the data is well distributed on disk. You're done, right? Well, no, because the application code might be poorly written. It might include delinquent design patterns. The demoed tuning tool AppSleuth will find those delinquent design patterns but it is the demo visitor's job to fix them. The demo scenario will consist of several "Test Your Skill" challenges with a tee-shirt as a prize. The scenarios will come from a transactional application, a Data Warehousing application, and an E-travel agency. For concreteness, we illustrate the functionality of AppSleuth on the E-travel agency here.

PostgreSQL Anomalous Query Detector

Bilal Shebaro, Asmaa Sallam, Ashish Kamra and Elisa Bertino

We propose to demonstrate the design, implementation, and the capabilities of an anomaly detection (AD) system integrated with a relational database management system (DBMS). Our AD system is trained by extracting relevant features from the parse-tree representation of the SQL commands, and then uses the DBMS roles as the classes for the bayesian classifier. In the detection phase, the maximum apriori probability role is chosen by the classifier which, if not matching the role associated with the SQL command, raises an alarm. We have implemented such system in the PostgreSQL DBMS, integrated with the statistics collection and the query processing mechanism of the DBMS. During the demonstration, our audience will be given the choice of training our system using either synthetic role-based SQL query traces based on probability sampling, or by entering their own set of training queries. In the subsequent detection mode, the audience can test the detection capabilities of the system by submitting arbitrary SQL commands. We will also allow the audience to generate arbitrary workloads to measure the overhead of the training phase and the detection phase of our AD mechanism on the performance of the DBMS.

Processing XML Queries and Updates on Map/Reduce Clusters

Nicole Bidoit, Dario Colazzo, Noor Malla, Maurizio Nolé, Carlo Sartiani and Federico Ulliana

In this demo we will showcase a research prototype for processing queries and updates on large XML documents. The prototype is based on the idea of statically and dynamically partitioning the input document, so to distribute the computing load among the machines of a Map/Reduce cluster. Attendees will be able to run predefined queries and updates on documents conforming to the XMark schema, as well as to submit their own queries and updates.

CISC: Clustered Image Search by Conceptualization

Kaiqi Zhao, Enxun Wei, Qingyu Sui, Kenny Zhu and Eric Lo

Clustering of images from search results can improve the user experience of image search. Most of the existing systems use both visual features and surrounding texts as signals for clustering while this paper demonstrates the use of an external knowledge base to make better sense out of the text signals in a prototype system called CISC. Once we understand the semantics of the text better, the result of the clustering is significantly improved. In addition to clustering the images by their semantic entities, our system can also conceptualize each image cluster into a set of concepts to represent the meaning of the cluster

MinExp-Card: Limiting Data Collection Using a Smart Card

Nicolas Anciaux, Walid Bezza, Benjamin Nguyen and Michalis Vazirgiannis

Online services such as social care, tax services, bank loans and many others, request individuals to fill in application forms with hundreds of private data items, in order to calibrate their offer. In practice, far too much data is requested, leading to over data disclosure. As shown in our previous works, avoiding this problem would (1) improve the privacy of the applicants and (2) decrease costs for service providers. We demonstrate here a prototype designed and implemented in partnership with the General Council of Yvelines District in France. The prototype targets forms used to calibrate social care for dependant people. To maintain the privacy of the decision process used to calibrate the social care, we propose a smartcard implementation. We will show that a 50% reduction of the items exposed in application forms can be achieved, explore the quality and scalability of our smartcard implementation, and demonstrate its scope.

PrivComp: A Privacy-aware Data Service Composition System

Mahmoud Barhamgi, Djamal Benslimane, Youssef Amghar, Nora Cuppens-Boulahia and Frederic Cuppens

In this demo paper, we present a new privacy preserving composition execution system. Our system allows to execute queries over multiple data services without revealing any extra information to any of the involved services. None of involved services (and their providers) is be able to infer any information about the data the other services provide beyond what is permitted

ProQua: A System for Evaluating Logic-Based Scoring Functions on Uncertain Relational Data

Sebastian Lehrack, Sascha Saretz and Christian Winkel

ProQua is an innovative probabilistic database system which enables the application of logic-based and weighted similarity conditions on uncertain relation data. In this demonstration paper we describe the interrelations among the main concepts, present an archaeological example scenario and sketch the software architecture of ProQua.

ProvenanceCurious: A Tool to Infer Data Provenance from Scripts

Mohammad Rezwanul Huq, Peter M.G. Apers and Andreas Wombacher

The increasing data volume and highly complex models used in different domains make it difficult to debug models in cases of anomalies. Data provenance provides scientists sufficient information to investigate their models. In this paper, we propose a tool which can infer fine-grained data provenance based on a given script. The tool is demonstrated using a hydrological model. The tool is also tested successfully handling other scripts in different contexts.

Tutorials

Tuesday, March 19			
11:00-12:30	Tutorial Session 1		
Room:	Liguria		

Trust and Reputation in and Across Virtual Communities

Authors: Nurit Gal-Oz (Ben-Gurion University and Sapir Academic College, Israel), Ehud Gudes (Ben-Gurion University, Israel)

Trust and Reputation have become key enablers of positive interaction experiences on the Web. These systems accumulate information regarding activities of people or peers in general, to infer their reputation in some context or within a virtual community. Reputation information improves the quality of interactions between peers and reduces the effect of fraudulent members. In this tutorial we motivate the use of trust and reputation systems and survey some of the important models introduced in the past decade. Among these models, we present our work on the knot model, which deals with communities of strangers. Special attention is given to the way existing models tackle attempts to attack reputation systems. In a dynamic world, a person or a service may be a member of multiple communities and valuable information can be gained by sharing reputation of members among communities. In the second part of the tutorial, we present the CCR model for sharing reputation across virtual communities and address major privacy concerns related to it. In the third part of our talk, we discuss the use of reputation systems in other contexts, such as domain reputation for fighting malware, and outline our research directions on this subject.

Nurit Gal-Oz is a post-doc at the Telekom Innovation Laboratories and the department of Computer Science, Ben-Gurion university. She is also a faculty member in the department of Computer Science at Sapir college, Israel. In her PhD thesis she developed the Knots and CCR reputation models and since then published numerous papers in the area. Her research interests include Trust and Reputation, Privacy, Data mining and Software engineering.

Ehud Gudes is a professor in the department of Computer Science, Ben-Gurion university, Israel. His research interests are in Databases, Data mining and Data security and he has published over 100 papers and edited several books in the above areas. Together with Nurit Galoz he has published several recent papers on Trust and Reputation and was a member of the program committee of several major conferences on this topic.

Wednesday, March 20			
11:00-12:30	Tutorial Session 2		
Room:	Liguria		

The W3C PROV family of specifications for modelling provenance metadata

Authors: Paolo Missier (Newcastle University, UK) Khalid Belhajjame (University of Manchester, UK) James Cheney (University of Edinburgh, UK)

Provenance, a form of structured metadata designed to record the origin or source of information, can be instrumental in deciding whether information is to be trusted, how it can be integrated with other diverse information sources, and how to establish attribution of information to authors throughout its history. The PROV set of specifications, produced by the World Wide Web Consortium (W3C), is designed to promote the publication of provenance information on the Web, and offers a basis for interoperability across diverse provenance management systems. The PROV provenance model is deliberately generic and domain-agnostic, but extension mechanisms are available and can be exploited for modelling specific domains. This tutorial provides an account of these specifications. Starting from intuitive and informal examples that present idiomatic provenance patterns, it progressively introduces the relational model of provenance along with the constraints model for validation of provenance documents, and concludes with example applications that show the extension points in use.

Paolo Missier is a Lecturer in Information and Knowledge Management at the School of Computing Science, Newcastle University, UK. His current research interests include models and architectures for the management and exploitation of data provenance, specifically extensions of e-infrastructures for scientific provenance. He is the PI of the EPSRCfunded project ”Trusted Dynamic Coalitions”, investigating provenance-based policies for information exchanges amongst partners in the presence of limited trust. Paolo contributes to two Provenance-focused Working Groups: he is a member of the W3C Working Group on Provenance on the Web, where he co-edited the PROV-DM as well as other specifications; and he is co-lead of the Provenance Working Group of the NSF-funded DataONE project. Paolo holds degrees from the University of Manchester, UK (PhD, 2007), University of Houston, Tx., USA (MS, 1993) and Università di Udine, Italy (BS, MS, 1990).

James Cheney holds a Royal Society University Research Fellowship at the University of Edinburgh, working on topics in programming languages and databases. He holds degrees from Carnegie Mellon University (BS, MS 1998) and Cornell University (PhD, 2004). His research on provenance focuses on the reproducibility and transparency needs of scientific data, particularly curated databases; in addition he helped start the USENIX workshop series on ”Theory and Practice of Provenance”, co-organized the first Dagstuhl Seminar on Principles of Provenance in 2012, serves on the W3C Provenance Working Group, and was lead editor of the PROVCONSTRAINTS specification [PROV-CONSTR].

Khalid Belhajjame is a research associate at the University of Manchester. His general research areas are information and knowledge management, where he has contributed to research proposals in the fields of data integration, knowledge engineering of semantic web services, and scientific workflows. He is an active member of the EU Wf4EVER project, the W3C provenance working group, and the DataONE scientific workflow and provenance working group.

Thursday, March 21			
11:00-12:30 14:00-15:30	Tutorial Session 3		
Room:	Munizioniere0		

Schema Mapping and Data Examples

Authors: Balder ten Cate (UC Santa Cruz) Phokion G. Kolaitis (UC Santa Cruz and IBM Research - Almaden) Wang-Chiew Tan (UC Santa Cruz)

A fundamental task in data integration and data exchange is the design of schema mappings, that is, high-level declarative specifications of the relationship between two database schemas. Several research prototypes and commercial systems have been developed to facilitate schema-mapping design; a common characteristic of these systems is that they produce a schema mapping based on attribute correspondences across schemas solicited from the user via a visual interface. This methodology, however, suffers from certain shortcomings. In the past few years, a fundamentally different methodology to designing and understanding schema mappings has emerged. This new methodology is based on the systematic use of data examples to derive, illustrate, and refine schema mappings. Example-driven schema-mapping design is currently an active area of research in which several different approaches towards using data examples in schema-mapping design have been explored. After a brief overview of the earlier methodology, this tutorial will provide a comprehensive overview of the different ways in which data examples can be used in schema-mapping design. In particular, it will cover the basic concepts, technical results, and prototype systems that have been developed in the past few years, as well as open problems and directions for further research in this area

Balder ten Cate is an Associate Adjunct Professor of Computer Science at the University of California Santa Cruz. He is also a Computer Scientist at LogicBlox Inc. His recent research focuses on schema mappings, data exchange, and database queries with guarded negation.

Phokion Kolaitis is a Distinguished Professor of Computer Science at the University of California Santa Cruz and a Research Staff Member at IBM Research - Almaden. His research interests include principles of database systems and logic in computer science.

Wang-Chiew Tan is an Associate Professor of Computer Science at the University of California Santa Cruz. Her research interests include data integration, data exchange, and data provenance.

6. Workshops

Joint EDBT/ICDT PhD Workshop

Room: Munizioniere 3

8:45 - 9:00	Opening	
9:00 - 10:30	Session 1	Chair: Bernhard Volz
9:00	Data Warehouse Testing **Neveen Elgamal**	
9:30	Mining Irregularities in Maritime Container Itineraries **Muriel Pellissier**	
10:00	Multidimensional Process Mining - A flexible analysis approach for health services research **Thomas Vogelgesang**	
10:30	Coffee Break	
11:00-12:30	Session 2	Chair: Marco Mesiti
11:00	**Keynote Talk: Boris Novikov** Can High Quality Research be Really Useful?	
12:00	Database Support for Processing Complex Aggregate Queries over Data Streams **Yuanzhen Ji**	

12:30	Lunch	
14:00 - 15:30	Session 3	Chair: Bernhard Volz
14:00	ProPub: A Declarative Approach To Customize, Analyze, and Query Workflow Provenance **Saumen Dey**	
14:30	Discovery Querying in Linked Open Data **Stefan Hagedorn**	
15:00	Towards efficient and practical solutions for ontology-based data management **Valerio Santarelli**	
15:30	Coffee Break	
16:00 - 16:30	Closing	Chair: Marco Mesiti and Bernhard Volz
16:00-16:15	Bestowal of Best Submission Award	
16:15-16:30	Farewell	

Third International Workshop on Linked Web Data Management (LWDM)

Room: Liguria

9:00-9:10	Opening
9:10-9:30	Flash Session
	Quick 2 minute teaser of each of the 9 accepted papers presented later in the day
9:30-10:30	Keynote Talk
	New tools for ontology querying on Linked Open Data **Andrea Calì**
10:30	Coffee Break
11:00-12:30	Session 1: Query Answering of Linked Data
11:00	Structure Inference for Linked Data Sources using Clustering **Klitos Christodoulou, Norman W. Paton and Alvaro A. A. Fernandes**
11:30	Linked Data Classification: a Feature-based Approach **Alfio Ferrara, Lorenzo Genta and Stefano Montanelli**
12:00	Semantic Question Answering System over Linked Data using Relational Patterns **Sherzod Hakimov, Hakan Tunc, Marlen Akimaliev and Erdogan Dogdu**
12:30	Lunch
14:00-15:30	Session 2: Engineering Linked Open Data

14:00	Assessing Linkset Quality for Complementing Third-Party Datasets **Riccardo Albertoni and Asunción Gómez Pérez**
14:25	Improving Geo-spatial Linked Data with the wisdom of the crowds **Roula Karam and Michele Melchiori**
14:50	LOVER: Support for Modeling Data Using Linked Open Vocabularies **Johann Schaible, Thomas Gottron, Stefan Scheglmann and Ansgar Scherp**
15:15	Community Evolution Detection in Time-Evolving Information Networks **Alfredo Cuzzocrea, Francesco Folino and Claudia Diamantini**
15:30	Coffee Break
16:00-16:45	Special Session: (4th International Workshop on Business intelligencE and the WEB - BEWEB 2013)
16:05	Supporting Range Queries in XML Keyword Search **Yong Zeng, Zhifeng Bao, Tok Wang Ling**
16:20	A flexible framework for context-aware recommendations in the Social Commerce domain **Davide Ronca, Armando Calvanese and Cosimo Birtolo**
16:45-17:30	Panel: How to query Linked Open Data?
17:30	**Closing**

PAIS Workshop

Room: Camino

8:50-9:00	Opening	
9:00-10:00	Keynote Talk	Chair: Traian Marius Truta
9:00	Digital Identity Protection - Concepts and Issues **Elisa Bertino**	
10:00-10:30	Session 1	Chair: Traian Marius Truta
10:00	A Privacy Preserving Model Bridging Data Provider and Collector Preferences (short paper) **Kambiz Ghazinour and Ken Barker**	
10:15	A Trustworthy Database for Privacy-Preserving Video Surveillance (short paper) **Antoni Martínez-Ballesté, Hatem A. Rashwan, Jordi Castellà-Roca and Domènec Puig**	
10:30	Coffee Break	
11:00-12:30	Session 2	Chair: Farshad Fotouhi
11:00	Secure Multiparty Aggregation with Differential Privacy: A Comparative Study **Slawomir Goryczka, Li Xiong and Vaidy Sunderam**	
11:25	Anonymizing Sequential Releases under Arbitrary Updates **Adeel Anjum and Guillaume Raschia**	
11:50	Privacy Framework: Indistinguishable Privacy **Jinfei Liu, Li Xiong and Jun Luo**	
12:15	Monitoring and Recommending Privacy Settings in Social Networks (short paper) **Kambiz Ghazinour, Stan Matwin and Marina Sokolova**	
12:30	Lunch	

14:00-15:30	Panel	Chair: Tal Soffer
14:00	Privacy Challenges Facing Future and Emerging Technologies **Tal Soffer, Aharon Hauptman, Niv Ahituv, Bukhard Aufferman, Claire Lobet-Maris, Nicolas Bach and Michael Birnhack**	
15:30	Coffee Break	
16:00-17:30	Session 3	Chair: Li Xiong
16:00	Efficient Tree Pattern Queries On Encrypted XML Documents **Jianneng Cao, Fang-Yu Rao, Mehmet Kuzu, Elisa Bertino and Murat Kantarcioglu**	
16:25	Design and Implementation of Privacy-Preserving Reconciliation Protocols **Georg Neugebauer, Lucas Brutschy, Ulrike Meyer and Susanne Wetzel**	
16:50	A Theoretical Model for Obfuscating Web Navigation Trails **Fida Dankar and Khaled El Emam**	
17:15	Biometric Access Control for e-Health Records in Pre-hospital Care (short paper) **José Díaz-Palacios, Víctor Romo-Aledo and Amir Chinaei**	
17:30 - 17:35	Closing	

Workshop on Querying Graph Structured Data (GraphQ)

Room: Munizioniere 1

9:15-9:30	Opening	
9:30-10:30	Keynote Talk	Chairs: Federica Mandreoli, Wilma Penzo
9:30	What we talk about when we talk about graphs **George H. L. Fletcher**	
10:30	Coffee Break	
11:00-12:30	Session 1	Chair: Federica Mandreoli
11:00	Towards query model integration: topology-aware, IR-inspired metrics for declarative graph querying **Luiz Gomes-Jr, Rodrigo Jensen, and André Santanchè**	
11:30	A Similarity Measure for Approximate Querying over RDF data Roberto De Virgilio, Antonio Maccioni, and Riccardo Torlone	
12:00	Dynamic Multi-version Ontology-based Personalization **Fabio Grandi**	
12:30	Lunch	
14:00-15:30	Session 2	Chair: Wilma Penzo
14:00	On Implementing Provenance-Aware Regular Path Queries with Relational Query Engines **Saumen Dey, Victor Cuevas-Vicenttin, Sven Kohler, and Bertram Ludaescher**	
14:30	Performance of Graph Query Languages - Comparison of Cypher, Gremlin and Native Access in Neo4j **Florian Holzschuher and René Peinl**	
15:00	Finding Nearest Neighbors in Road Networks: A Tree	

	Decomposition Method **Fang Wei-Kleiner**	
15:30	Coffee Break	
16:00-16:30	Closing	Chairs: Federica Mandreoli, Wilma Penzo
16:00-16:30	Conclusions and farewell	

Energy Data Management (EnDM)

Room: Anfiteatro

9:00-9:05	Opening	
9:05-10:25	Session 1	Chair: Torben Bach Pedersen
9:05	Industrial Keynote	
9:30	Symbolic Representation of Smart Meter Data: **Tri Kurniawan Wijaya, Julien Eberle and Karl Aberer**	
10:25	Coffee Break	
10:45-13:00	Session 2	Chair: Torben Bach Pedersen
10:45	Visualizing Complex Energy Planning Objects With Inherent Flexibilities: **Laurynas Siksnys and Dalia Kaulakiene**	
11:15	Towards Ontological Foundations of Knowledge related to the Emissions Trading System: **Umberto Ciorba, Antonio De Nicola, Stefano La Malfa, Tiziano Pignatelli, Vittorio Rosato, Maria Luisa Villani**	
11:35	Optimized Renewable Energy Forecasting in Local Distribution Networks: **Robert Ulbricht, Ulrike Fischer, Wolfgang Lehner and Hilko Donker**	
11:55	Towards the Automated Extraction of Flexibilities from Electricity Time Series: **Dalia Kaulakiene, Laurynas Siksnys and Yoann Pitarch**	
12:15	Research Challenges for Energy Data Management (panel) **Torben Bach Pedersen and Wolfgang Lehner (organizers) plus additional panelists**	
13:00	Lunch	

BigProv: Managing and Querying Provenance Data at Scale / ProvBench

Room: Munizioniere 2

8:45-9:00	Opening	
9:00-10:30	Session 1	Chair: Paolo Missier
9:00	**Keynote 1:** Use of Real-Life EHR Data for Clinical Research and Personalized Medicine **Dr. Joerg Kraenzlein, Medical Director, Big Data, iSOFT Health GmbH**	
9:45	Towards Design Support for Provenance Awareness: A Classification of Provenance Questions (regular paper) **Paraskevi Zerva, Steffen Zschaler, and Simon Miles**	
10:10	Enhancing and Abstracting Scientific Workflow Provenance for Data Publishing (short paper) **Pinar Alper, Khalid Belhajjame, Carole A. Goble, and Pinar Karagoz**	
10:30	Coffee Break	
11:00-12:30	Session 2	Chair:
11:00	Provenance from Log Files: a BigData Problem (regular paper) **Devarshi Ghoshal and Beth Plale**	
11:25	Capturing and Querying Workflow Runtime Provenance with PROV: a Practical Approach (short paper): **Flavio Costa, Vítor Silva, Daniel de Oliveira, Kary Ocaña, Eduardo Ogasawara, Jonas Dias, Marta Mattoso**	
11:45	ProvBench short presentations:	

	VisTrails Provenance Traces for Benchmarking **Fernando Chirigati, David Koop, Juliana Freire, and Claudio Silva** Provenance Traces of the Swift Parallel Scripting System **Luiz M. R. Gadelha Jr., Michael Wilde, Marta Mattoso, and Ian Foster** A Workflow PROV-Corpus based on Taverna and Wings **Khalid Belhajjame, Jun Zhao, Daniel Garijo, Aleix Garrido, Stian Soiland-Reyes, Pinar Alper, and Oscar Corcho** PROV-O Provenance Traces From Agent-based Social Simulation **Edoardo Pignotti, Gary Polhill, and Peter Edwards**	
12:30	Lunch	
14:00-15:30	Session 3	Chair:
14:00	**Keynote 2:** Foundations and applications of Data Provenance **Dr. Grigoris Karvounarakis, LogicBlox, USA**	
14:45	WebLab PROV : Computing fine-grained provenance links for XML artifacts (regular paper) **Bernd Amann, Camelia Constantin, Clément Caron, and Patrick Giroux**	
15:10	Using Provenance to Analyse Agent-based Simulations (short paper) **Edoardo Pignotti, Gary Polhill, and Peter Edwards**	
15:30	Coffee Break	
16:00 - 17:30	Session 4	Chair: Bertram Ludascher
16:00	Provenance for seismological processing pipelines in a distributed streaming workflow (short paper) **Alessandro Spinuso, James Cheney, and Malcolm Atkinson**	
16:20	ProvBench short presentations: Extracting PROV provenance traces from Wikipedia history pages, **Paolo Missier and Ziyu Chen** Provenance Traces from Chiron Parallel Workflow Engine, **Felipe Horta, Vítor Silva, Flavio Costa, Daniel de**	

	Oliveira, Kary Ocaña, Eduardo Ogasawara, Jonas Dias, and Marta Mattoso Provenance in Streamflow Forecasting**, Heiko Müller, Chris Peters, Yanfeng Shu, and Andrew Terhorst**
17:00	**Discussion and wrap up**

Scalable String Similarity Search/Join

Room: Cisterne

9:00-10:30	Session 1	Chair: Ulf Leser
9:00-09:30	Welcome **Ulf Leser**	
9:30-09:50	Efficient Edit Distance based String Similarity Search using Deletion Neighborhoods **Akhil Arora;Shashwat Mishra;Tejas Gandhi;Arnab Bhattacharya**	
09:50-10:10	Approximate String Matching by Position Restricted Alignment **Manish Patil;Xuanting Cai;Sharma V. Thankachan;Rahul Shah;David Foltz;Seung-Jong Park**	
10:10-10:30	Scalable string similarity search / join with approximate seeds and multiple backtracking **Enrico Siragusa;David Weese;Knut Reinert**	
10:30-11:00	Coffee Break	
11:00-12:20	Session 2	Chair: Ulf Leser
11:00-11:20	Efficient Parallel Partition-based Algorithms for Similarity Search and Join with Edit Distance Constraints **Yu Jiang;Dong Deng;Jiannan Wang;Guoliang Li;Jianhua Feng**	
11:20-11:40	WallBreaker - overcoming the wall effect in similarity search **Stefan Gerdjikov;Stoyan Mihov;Petar Mitankin;Klaus U. Schulz**	

11:40-12:00	Efficient algorithms for edit similarity queries **Jianbin Qin;Xiaoling Zhou;Wei Wang;Chuan Xiao**	
12:00-12:20	Cache-Aware Parallel Approximate String Search and Join Using BWT **Jiaying Wang;Xiaochun Yang;Bin Wang**	
12:30-14:00	Lunch	
14:00-15:00	Session 3	Chair: Ulf Leser
14:00-14:20	Evaluation of the competition **Sebastian Wandelt**	
14:20-15:00	Lessons learned, feedback, publication and further plans **Ulf Leser**	

7. Conference Organization

General Chair

Giovanna Guerrini, Università di Genova, Italy

Executive Chair

Barbara Catania, Università di Genova, Italy

EDBT Program Chair

Norman Paton, University of Manchester, UK

ICDT Program Chair

Wang Chiew Tan, IBM Almaden - Almaden and UC Santa Cruz, USA

Workshop Chair

Kai-Uwe Sattler, Technische Universität Ilmenau, Germany

Tutorial Chair

Paolo Atzeni, Università Roma Tre, Italy

Demo Chair

Piero Fraternali, Politecnico di Milano, Italy

Industrial Chair

Malu Castellanos, HP Labs, USA

Proceedings Chair

Anastasios Gounaris, Aristotle University of Thessaloniki, Greece

Sponsorship Chair

Giorgio Delzanno, Università di Genova, Italy

Publicity Chair

Marco Mesiti, Università di Milano, Italy

Website Chair

Federico Cavalieri, Università di Genova, Italy

Organization Support

Graphics

Daniela Peghini, Università di Genova, Italy

Conference Helpers

Università di Genova, Italy:

Alessandro Solimando, Beyza Yaman, Ulanbek Mambetakunov, Andrea Corradi, Davide De Tommaso, Laura Di Rocco, Paolo Farina, Angela Locoro, Alberto Minetti, Alessio Parma, Paola Podestà, Andrea Stocco, Sonia Volpe

Others: Federico Catania, Francesca Martignone, Mara Mazzarello

Sponsored by

SAP

Google

hp

IBM Research

coop
Liguria

Supported by

ti Liguria

TOMASONI
YACHTING & SPORT