EDBT ICDT

EDBT / ICDT 2014

JOINT CONFERENCE

24-28 March 2014 • Athens, Greece

EDBT: 17th International Conference on Extending Database Technology
ICDT: 17th International Conference on Database Theory

www.edbticdt2014.gr

COMMUNICATION SUPPORT

EKT

ORGANIZERS

FORTH

University of Crete

Gold Sponsors	Silver Sponsors	Bronze Sponsors
LIG cnrs LogicBlox	technicolor IBM Research Google	INTRASOFT INTERNATIONAL hp

1. Welcome

EDBT/ICDT 2014's Organizing Committee is pleased to welcome you to Athens for the joint edition of the 17th International Conference on Extending Database Technology and the 17th International Conference on Database Theory. Since ICDT 1997 held in Delphi and EDBT 2004 held in Heraklion, the Conference have traveled across Europe to return this year back in Greece. Athens is both an "ancient" and "modern" city, in which visitors can walk safely and enjoy the rich – almost 5,000 year old – history it has to offer. The city offers a lot of sightseeing, museums, shopping and nightlife. The Conference venue is the Royal Olympic Hotel located at the Athenian historical center called Plaka, only 2 minutes' walk from the new Athens Acropolis Museum. In such an exciting context, we trust you will find the program interesting and enjoyable.

The main factor determining the success of a Conference is the quality of its technical program, and this year's Conference offers a well-integrated collage of events, including outstanding invited speakers, carefully refereed technical papers, selected tutorials, demonstrations and topic-focused workshops. We are delighted to announce outstanding keynote talks by Serge Abiteboul (INRIA & Ecole Normale Supérieure de Cachan, France), Peter Boncz (CWI & VU University Amsterdam, Netherlans), Frank Neven (Hasselt University), Christopher Ré (Stanford University) and an invited ICDT lecture by Christian Bizer (University of Mannheim). Furthermore, we wish to thank all those who are organizing workshops, giving tutorials, presenting papers, demonstrations and posters for their significant efforts, all of which contribute to the richness of the Conference program.

This year's EDBT Program Chair is Sihem Amer-Yahia (CNRS - LIG, France) and ICDT Program Chair is Nicole Schweikardt (University of Frankfurt, Germany). Selcuk Candan (Arizona State University, USA) is responsible for workshops, Minos Garofalakis (Technical University of Crete, Greece) for tutorials, Stratos Idreos (Harvard University, USA) for demonstrations, and Anastasios Kementsietsidis (IBM Research, USA) for industry & application sessions. Proceedings have been prepared by Vincent Leroy (University of Grenoble & CNRS - LIG, France), sponsorships have been pursued by Dimitris Kotzinos (Université de Cergy-Pontoise, France & FORTH-ICS, Greece), publicity duties have been undertaken by Grigoris Karvounarakis (LogicBlox, USA), while the web site has been maintained by Yannis Stavrakas (IMIS - Athena RC, Greece) and all local support has been provided by Efi Papastavropoulou (Triaena Tours & Congress, Greece). I would finally like to thank Theodossia Bitzou (FORTH-ICS, Greece) for the graphic design of posters and Nikos Karakostas (Psimythi Ltd.), Christina Machaira and Panos Kassiaris (Myriki & Co) for the layout and print of the Conference brochure. They have all form a very efficient team for successfully organizing the Conference and cooperating with all of them has been a great pleasure.

The joint EDBT/ICDT 2014 Conference is being organized by the Department of Computer Science, University of Crete and the Institute of Computer Science of FORTH at Heraklion, Greece. It is being held under the aegis of EDBT Association and ICDT Council. We thank EDBT Association president, Christine Collet, and the ICDT Council chair, Thomas Schwentick, for their help and advice. Since ACM has changed their Open Access (OA) strategy significantly, this has led to a late change in the publication platform from ACM's AICPS series to OpenProceedings.org, a free, Open Access platform for high-profile Conference proceedings series. The EDBT and ICDT governing bodies continue the OA policy to their proceedings.

The Conference would not have been possible without the financial support of our sponsors. Special thanks go to LogicBlox, CNRS-LIG France (for supporting the student scholarships), Technicolor SA (for supporting the EDBT/ICDT Test of Time Awards), Google, IBM Research, HP (for supporting invited speakers), Intrasoft International (for supporting the ICDT best papers awards) and the National Documentation Centre (for live streaming support).

On behalf of the Conference Organization Committee, we would like to thank you all for your submissions to EDBT/ICDT 2014, your hard work serving on EDBT/ICDT 2014 committees, and for participating in the Conference. It is your ideas, interest and input that shaped the Conference and that continue to make it such a valuable venue for our community.

Welcome to Athens and enjoy EDBT/ICDT 2014!

Vassilis Christophides
(University of Crete, Greece, and
Technicolor R&I Center, Paris, France)
on behalf of the Organizing Committee

Organizers

University of Crete

FORTH
Institute of Computer Science

Gold Sponsors

LIG

cnrs

LogicBlox

Silver Sponsors

technicolor

IBM Research

Google

Bronze Sponsors

INTRASOFT INTERNATIONAL

hp

Supporters

EKT

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ

NATIONAL DOCUMENTATION CENTRE

General Information

The city of Athens

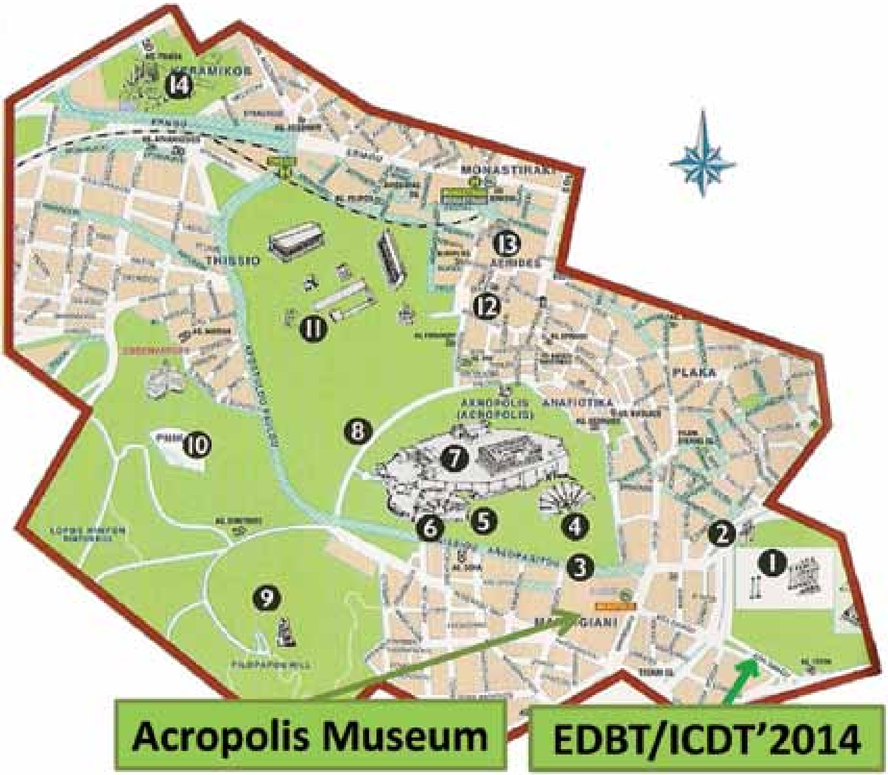
With an urban population of more than 4 million people, Athens is the capital of Greece and the 4th most populous capital in the E.U. Athens is known as the birth place of Democracy, the city with the world-renown “Acropolis” and “Parthenon”, the famous Theater of Herodes Atticus, the “marble stadium” where the first modern-time Olympic Games took place in 1896, as well as the home of Socrates, Plato and Pericles (Golden Age).

Besides its strong links with the past, Athens has many future prospects, especially after the creation of a new city center in the context of the ongoing project “Rethink Athens” (www.rethinkathens.org). It is a vibrant metropolis with great appeal and charm and with so many things to do and to see. It is among the most beautiful, hospitable and exciting cities in Europe, attracting thousands of visitors who enjoy the excellent dry climate, the sea, the nice food, the ancient treasures and a night life that seems to go on forever.

Walking Itineraries

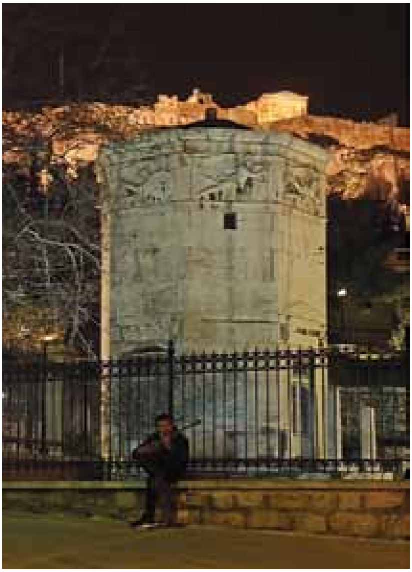
It is impossible to describe all the sites of interest and cultural opportunities offered by Athens. While attending the Conference you can enjoy one or more of the following walks in the historical center of Athens.

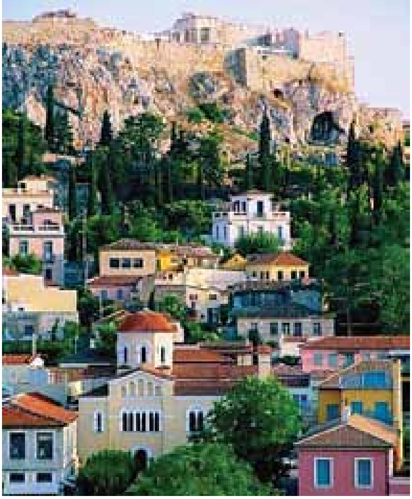
The Olympieion (1) is located in a 2 min walking distance from the meeting venue and includes the sanctuary (temple) of Olympian Zeus, Roman baths, classical houses together with a small section of the ancient city’s fortification wall. The construction of Olympieion started as early as 515 BC, but was eventually completed in 131 AC by the Roman Emperor Hadrian. In front of the Olympieion (1), on the northeastern perimeter of the temple, there is a magnificent arch called Hadrian’s Arch (2), which was built entirely of Pentelic marble in 131 AD by the Athenians, in honor of Emperor Hadrian. Dionyssiou Aeropagitou street (3) is one of the



most impressive pedestrian streets in this area. It is located in a 2 min walking distance from the meeting venue and offers a magnificent view of the Acropolis (7) and the Parthenon. It starts from the Hadrian's Arch (2), passes through the entrance of the Ancient theatre of Dionysos (4) (world's most ancient theatre), the Stoa of Eumenes (5) (that provided shelter to the theatergoers), the Odeion of Herodes Atticus (6) (build in 161 AC and still used today for theatrical plays, concerts etc, e.g. see Athens Festival), the Acropolis Museum, the entrance to Acropolis (7) itself, the National Observatory located in the highest point of Nymphon hill and many more sites of great archeological interest (Areios Pagos (8), Philopappou hill (9), the Pnyx (10) and the Ancient Agora (11)) to reach Thissio area. The Dionyssiou Areopagitou Street represents an ideal site for walking, relaxing and even jogging.

Thissio area, apart from a place that includes numerous sites of significant archeological interest (e.g. Temple of Hephestos, the monument of Eponymous Heroes, the Stoa of Attalus and more) is also a place with many small cafés and restaurants with a spectacular view to the Ancient Agora and the Acropolis. In the northeast of Thissio is the Roman Agora (12) (built between 19 and 11 BC) and Hadrian's library (13) (build in 132 AD). These two sites,





along with many others (e.g. Aerides), are located in the northern foot of the Acropolis, in a part of the Athenian historical center called Plaka. Next to Plaka, lies a distinctive "old" Athens area, with narrow streets and small buildings called Monastiraki.

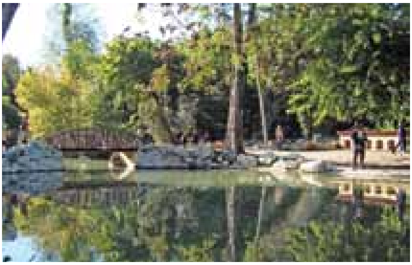
In both Plaka and Monastiraki, visitors can find a variety of small shops with shoes, cloths, old furniture, old books, jewelries, souvenirs, records etc. together with delicious small cafés and traditional restaurants. A visit to Plaka and Monastiraki is an unforgettable experience. North of Monastiraki, the Psyri area provides a colorful neoclassical touch to the historical Center of Athens. An area filled with art galleries, small bars, night clubs, shops, cafés and restaurants with both traditional Greek cuisine (taverns) and international modern cuisine. The historic neighborhood of Kerameikos, located in the northwestern part of Athens was initially the biggest and first cemetery of Attica and moreover a place of burial for citizens that were honored by the city of Athens. In recent years, Kerameikos has become one of the most popular places of Athens with night clubs, restaurants, bars etc.

The Panathenaikon Stadium (Kallimarmaro) is found in the northeastern part of the Olympieion (not visible in the Map above), only a 10 minutes' walk from the Conference venue, surrounded by Ardittos Hill, that provides an excellent wooded place for walking and jogging. Kallimarmaro is a spectacular construction, built entirely of white marble, although it was originally crafted from wood (330 BC). It used to be a place that hosted the athletic competitions during the Panathenian festival. The Panathenaiko Stadium was the venue for the first Olympic Games in the modern world history (1896).

Right across from the Panathenaiko Stadium is the National Garden of Athens that is open for visitors throughout the day, from dawn till sunset. It is a 160.000 m^2 area that contains hundreds of different kinds of plants, trees and bushes, while it also comprises an important natural life reserve in the city of Athens with many different species of birds, turtles etc that have found shelter there. Inside the National Garden that is connected to Zappeion, visitors can find a Botanical museum, a children's library, a playground and a café. It goes without saying that this is also an ideal place for walking, relaxing and also jogging.

Museums

The National Archaeological Museum (www.namuseum.gr) is the largest archaeological museum in Greece and one of the most important museums in the world devoted to ancient Greek art. Although its original purpose was to secure all the finds from the nineteenth century excavations in and around Athens, it gradually became the central National Archaeological Museum and was enriched with finds from all over Greece.



The museum is housed in an imposing neoclassical building near the juncture of Alexandras Avenue and Patission Avenue of the end of the nineteenth century, which was designed by L. Lange and remodeled by Ernst Ziller. The vast exhibition space-numerous galleries on each floor accounting for a total of 8,000 square metres- house five large permanent collections: The Prehistoric Collection including works of the great civilizations that developed in the Aegean from the sixth millennium BC to 1050 BC (Neolithic, Cycladic, Mycenaean), and finds from the prehistoric settlement at Thera. The Sculptures Collection, exhibiting the development of ancient Greek sculpture from the seventh to the fifth centuries BC with unique masterpieces. The Vase and Minor Objects Col-lection, with representa-

tive works of ancient Greek pottery from the eleventh century BC to the Roman period, also including the Stathatos Collection, a corpus of minor objects of all periods. The Metallurgy Collection, with many fundamental statues, figurines and minor objects as well as the only Egyptian and Near Eastern Antiquities Collection in Greece, with works dating from the pre-dynastic period (5000 BC) to the Roman conquest.

The museum is a five minutes' walk from Victoria station (Metro line 1) and a 10 minutes' walk from Omonia stop (Metro lines 1 & 2). The full admission fee is 7 Euro and the reduced is 3 Euro. A special ticket package (Full: 12 Euro, Reduced: 6 Euro) is available for visiting National Archaeological Museum along with the *Byzantine & Christian Museum*, the *Numismatic Museum* and the *Epigraphic Museum*.

Opening Hours:
Monday: 13:00 - 20:00
Tuesday - Sunday & Public Holidays: 09:00-16:00
The National Archaeological Museum will be closed on 25th March.

The Benaki Museum (www.benaki.gr) ranks among the major institutions that have enriched the material assets of the Greek state. It is also the oldest museum in Greece operating as a Foundation under Private Law. Through its extensive collections that cover several different cultural fields and its more general range of activities serving more than one social need, the Benaki Museum is perhaps the sole instance of a complex structure within the broader network of museum foundations in Greece. The Museum's purpose is to preserve collective memories and cultural heritage; and its holdings include distinctive artifacts from every stage of Greek and world artistic production, dating from prehistoric times to the present day. The Museum's main concern is to protect, conserve and showcase cultural heritage, while also seeking ways to create links between our contemporary society and this heritage and render it accessible to the research establishment. To this end, the Museum maintains close links with similar Institutions and prestigious domestic and foreign universities, actively developing interdisciplinary approaches able to locate and process thoroughly modern-day social and cultural challenges.

The Benaki Museum occupies one of the few neoclassical buildings that continue to resist the aesthetic deterioration of postwar Athens. It is located on an exceptionally favorable location in the historic center of the city, right across from the lush greenery of the

National Gardens and the garden of the Presidential Mansion, and in the same vicinity as related institutions such as the *Museum of Cycladic Art* and the *Byzantine & Christian Museum*. The museum spans over several buildings.

Main Building: Vassilisis Sofias and 1 Koumbari street
Open Hours
Wednesday, Friday: 9:00 - 17:00
Thursday, Saturday: 9.00 - 24.00
Sunday: 9:00 - 15:00
Full admission : € 7 Temporary Exhibition: € 5
Reduced admission : € 5 Temporary Exhibition : € 3

Pireos St. Annexe: 138 Pireos Street
Thursday, Sunday: 10:00 - 18:00
Friday, Saturday: 10:00 - 22:00
Full admission: € 4 to € 6
Reduced admission: € 2 to € 3

Museum of Islamic Art: 22 Ag. Asomaton & 12 Dipilou St (close to the grounds of the ancient Agora the Doric temple of Hephaestus, “Theseio”, and the Museum of the ancient Kerameikos necropolis -cemetery).
Thursday, Friday, Saturday, Sunday: 9:00 - 17:00
Full admission: € 7
Reduced admission: € 5
All museum building are closed on Monday and Tuesday while free admission is offered every Thursday.

The Herakleidon Museum of Art (www.herakleidon-art.gr) is private museum located in “Theseio” that opened its doors to friends of the fine arts in the summer of 2004. Its mission is to introduce visitors to the art of the exhibited artist, to show how the artist has evolved through the various important periods in his or her career, to explain the various techniques used by the artist to express his or her vision, and to help the visitors connect to both the artist’s life and work. To accomplish this, not only will works of the artist be on exhibit, but also preparatory sketches, drawings, photographs and personal items. Audiovisual aids provide complete and detailed information on the life of the artist, each phase of his or her work, and his or her techniques. In addition to the permanent collections, the museum “Herakleidon, Experience in Visual Arts” hosts exhibitions of Greek and foreign artists. The general admission is 6€, for students & seniors (over 65) is 4€ and for children up to 12 is free.

From the “Theseio” stop of Metro Line 1, after getting to street level, turn right and follow the pedestrian street Apostolou Pavlou (walk up towards the Acropolis). At Londos Restaurant-

Cafe turn right onto Herakleidon St. Keep to your right and follow the pedestrian street. Herakleidon, Experience in Visual Arts is number 16, just past the cafes on your right. Alternatively, after leaving the Acropolis, follow the pedestrian street Apostolou Pavlou down towards "Theseio". When you reach the Londos Cafe-Restaurant (on your left hand side), turn left onto Herakleidon St. Keep to your right and follow the pedestrian street to arrive at the number 16.

Open Hours
Friday: 13:00 - 21:00
Saturday: 11:00 - 19:00
Sunday: 11:00 - 19:00

The Museum of Cycladic Art (www.cycladic.gr) is dedicated to the study and promotion of ancient cultures of the Aegean and Cyprus, with special emphasis on Cycladic Art of the 3rd millennium BC. It was founded in 1986, to house the collection of Nicholas and Dolly Goulandris. Since then it has grown in size to accommodate new acquisitions, obtained either through direct purchases or through donations by important collectors and institutions.

Today, over 3.000 artefacts of Cycladic, Ancient Greek and Cypriot art are on display on four floors, in the galleries of the Museum, a living cultural space in the heart of Athens. One of the Museum wings that hosts temporary exhibitions is the Stathatos Mansion (1 Irodotou & Vas. Sofias Ave), one of the best examples of neoclassical architecture in Athens, work of the Bavarian architect Ernest Ziller. A glass-covered passage connects Stathatos Mansion to the Permanent Collections (main) building at 4 Neophytou Douka Str, Kolonaki. The standard admission is 7€, for students & seniors (over 65) is 3,5€, for kids and young persons under 18 is free while every Monday the entrance fee is 3,5€.

Opening hours
Monday - Wednesday - Friday -
Saturday: 10:00 - 17:00
Thursday: 10:00 - 20:00
Sunday: 11:00 - 17:00
Tuesday: closed

The Byzantine and Christian Museum (www.byzantinemuseum. gr), which is based in Athens, is one of Greece's national museums. Its areas of competency are centered on – but not limited to – religious artefacts of the Early Christian, Byzantine, Medieval, post-Byzantine and later periods which it exhibits, but also acquires, receives, preserves, conserves, records, documents, researches, studies, publishes and raises awareness of. The museum has over 25,000 artefacts in its possession. The artefacts date from between the 3rd and 20th century AD, and their provenance encompasses the entire Greek world, as well as regions in which Hellenism flourished. The size and range of the collections and value of the exhibits makes the Museum a veritable treasury of Byzantine & post-Byzantine art and culture.

The Villa Ilissia, which nowadays houses the Byzantine and Christian Museum, is one of the loveliest buildings erected in Athens during its early years as capital of the newly-founded Greek State. When Athens was officially declared the capital, in 1834, it was a town of some 7000 souls. Within two years, however, its population had doubled, as the new administrative authorities were installed here and many new inhabitants arrived from all over Greece and other countries too. Noteworthy among the incomers were Greek and Bavarian civil servants, European philhellenes and devotees of the East, Phanariots and other educated Greeks from abroad, veterans of the War of Independence, and community leaders from the provinces, merchants, entrepreneurs and bankers, together with ordinary folk from all parts of Greece, who flocked to Athens seeking employ or simply a better life. The café-bistro «ILISSIA» is located in the garden, in the shadow of the Villa where visitors can enjoy coffee, desert or a light meal. The general admission is 4€, for students & seniors (over 65) is 2€, for kids and young persons under 18 is free.

Opening hours
Tuesday-Sunday: 08:30-16:00 (temporary exhibitions & shop: 08:30-16:00). The Byzantine and Christian Museum will be closed on 25th March.

The Museum of city of Athens (www.athenscitymuseum.gr) Vouros - Eutaxias presents the modern history of Europe's oldest and most famous city, since it became the capital of the newly founded Hellenic State in 1834 under the first royal couple, Otto and Amalia. Your tour will also include many other facets of Athenian history, culture and

life. Several other collections and a series of typical late 19th century- early 20th century sitting and living rooms of the Athenian aristocracy complete the different aspects of the city offered by the Museum. It consists of two historic buildings connected by a bridge for the museum needs. The house Stamatiou Dekozis Vourou, Paparrigopoulou 7, one of the oldest buildings in the post-revolutionary Athens, built in 1833-4 to designs by German architects G. Lueders and J. Hoffman, on the northern outskirts of the destroyed by the war of independence, city of Athens. Along with the neighboring house Afthonidis, which was built at the same time, became the first and temporary residence of the first royal couple of modern Greece, known as the old palace. The general admission is 3€, for students, seniors (over 65) & groups is 2€ while for persons with special needs is free.

Opening hours
Tuesday-Friday: 09:00-16:00.
Saturday-Sunday: 10:00-15:00
The Museum of city of Athens will be closed on 25th March.

The National Museum of Contemporary Art (www.emst.gr) will be hosted in the historic building of Athens' Fix brewery and its openings are planned for March 2014. A remarkable nucleus of works of art by Greek and international artists has been already developed, composed of collections of selective rather than encyclopedic character, in order to promote advanced tendencies and critical explorations of the artistic present but also its historical depths which reach as far as the second half of the 20th century. The museum goal, both by exhibitions and collections, is to offer all the visitors, which remains an unreservedly democratic institution, the "other" dimension which in our time cannot be conceived outside transcultural and ecumenical patterns.

Conference Information

Conference venue

The Conference will take place in the Royal Olympic Hotel. After a complete renovation in 2009, Royal Olympic was transformed into an art hotel, very elegantly decorated and, more importantly, very well looked-after in every detail. One of the aspects given particular attention to was to create a very personal hotel that is as environmentally friendly as possible.

The Royal Olympic Hotel is one of the best five-star hotels in Athens. The hotel is located just in front of the famous Temple of Zeus and the National Gardens. The hotel is underneath the Acropolis and only 2 minutes' walk to the new Athens Acropolis Museum and other major archeological sites.

Athens Royal Olympic Hotel is located in 28-34 Athanasiou Diakou Street. The closest metro station is "Acropolis" (Line 2). From "Acropolis" Station the Hotel is 200m away, walking along Athanasiou Diakou Str.

Registration Desk

Registration desk will be positioned on Floor -1, at Olympia Foyer just after the hotel entrance on Monday 24 until Friday 28.

Conference Secretariat

Conference secretariat desk (registration agency for receipts, onsite registration payments) will be open at the Olympia foyer: on Monday 24 until Friday 28. 8:30-17:30

Staff

Staff members are wearing green profiled badges.

Wireless Access

The wireless network will be freely available by all the Hotel areas

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 Hotel reception. 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 at the Olympia Foyer.
Light Lunch will be served daily 12:30-14:00 at the Olympia Foyer

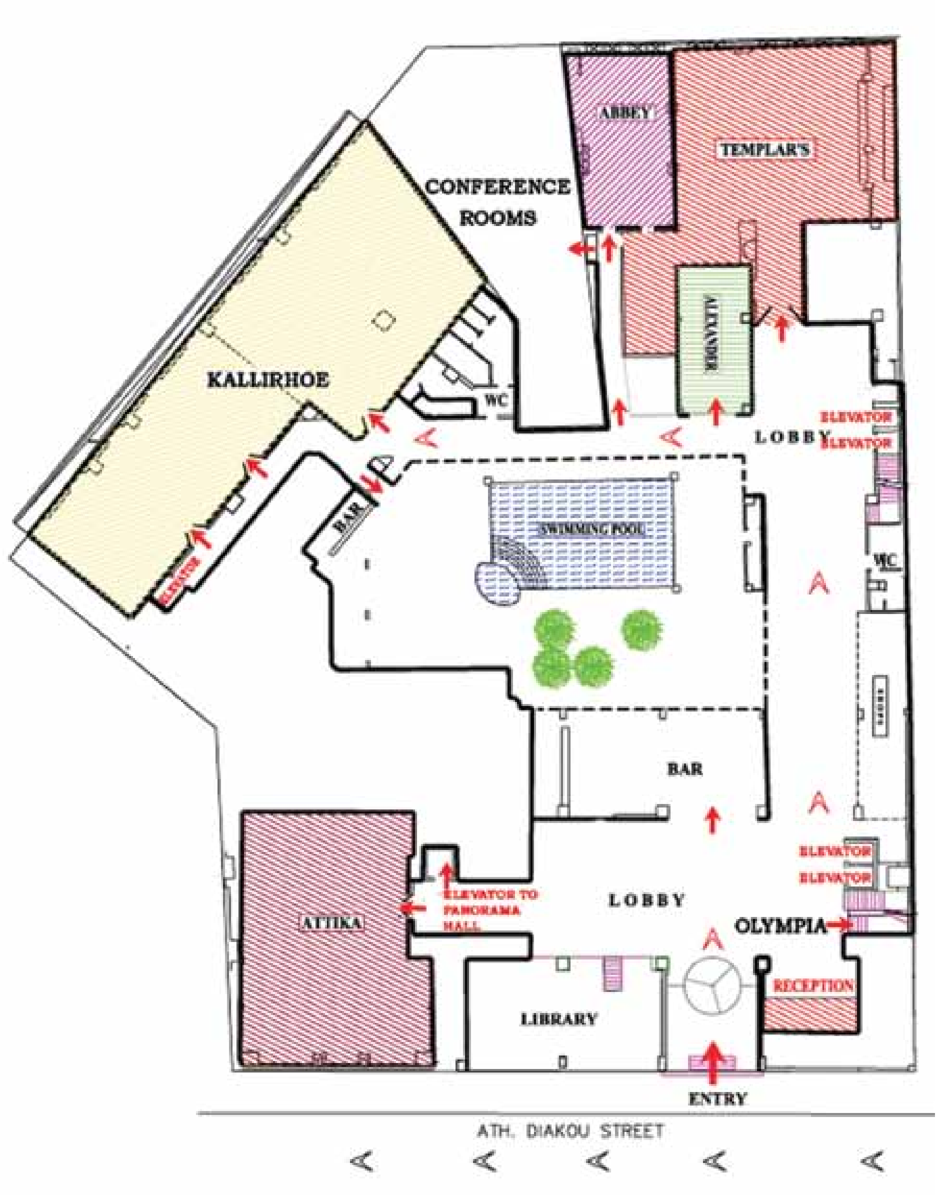
Other facilities

For Conference participants the following touristic services are available on request.

- Guided Tour to the New Acropolis Museum
- Half day Cape Sounion tour in the afternoon
- Full day to Delphi including lunch
- Full day to Argolis (Nafplion, Epidaurus, Mycenae) including lunch
- One day cruise to Saronic Islands of Hydra, Poros & Aegina including lunch on board

Venue Maps

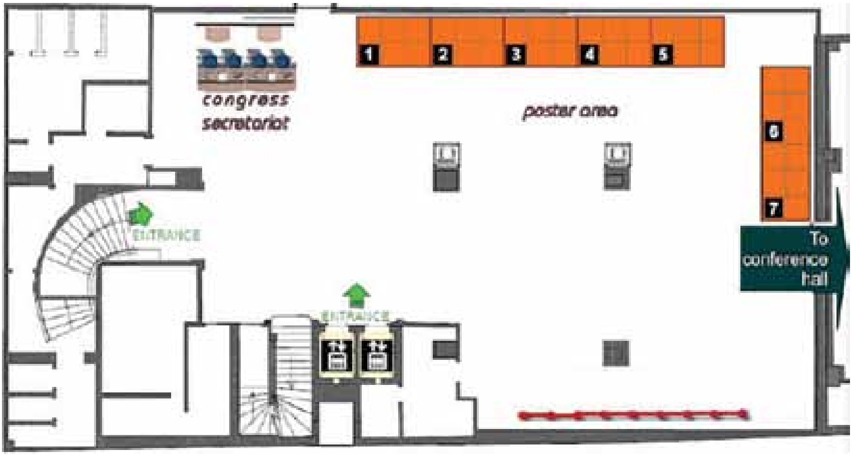
The Conference is held at Floors -1 (Olympia, Olympia Foyer), at Lobby area (Attica, Templar) and Floor 2 (Conference room 2) of the Royal Olympic Hotel. Some additional rooms at Floor 1 (Conference rooms 1A and 1B) will be used for workshops only on Friday 28. Coffee breaks and lunches will be served at the Olympia Foyer.



ROYAL OLYMPIC HOTEL

LOBBY - ΑΙΘΟΥΣΕΣ

Olympia Foyer



Room Summary

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

Room Name	Floor	Used for	Used on
Olympia (1&2)	-1	Keynotes, Tutorials ICDT Invited Lecture EDBT Sessions LWDM, GraphQ, ExploreDB	Mon, Tue, Wed, Thu, Fri
Attica	Lobby area	ICDT Sessions EDBT Sessions BeyondMR	Mon, Tue, Wed, Thu, Fri
Templar's	Lobby area	EDBT Sessions BX, EnDM	Tue, Wed, Thu, Fri
Olympia Foyer	-1	Demos/Posters	Tue, Wed
Conference room 2	2	EDBT Sessions MUD	Tue, Wed, Thu , Fri
Conference room 1A	1	PAIS	Fri
Conference room 1B	1	MSDM	Fri
Conference Grand	Lobby area	Organizers' Office	Mon, Tue, Wed, Thu, Fri

Social Events

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

Tuesday 25th March, 19:00: Welcome Party

Our evening will start with a guided tour at the new Acropolis Museum exhibits, followed by a cocktail at the Museum's restaurant on the second floor. The new Acropolis Museum is considered to be one of the most important Museums in the world, since it houses the unique collection of original sculptural masterpieces of Archaic and Classical Greek art from the sacred rock and citadel of ancient Athens. It has a total area of 25,000 square meters, with exhibition space of over 14,000 square meters, ten times more than that of the old museum on the Hill of the Acropolis.

These are mainly freestanding votive sculptures and important groups of architectural sculptures, which decorated the buildings erected on the Acropolis in the Archaic and Classical periods. The display also includes clay votive offerings. Other finds from the Acropolis, such as vases, bronze objects and relief sculptures, are displayed in the National Archaeological Museum, while the inscriptions are kept in the Epigraphical Museum. A noteworthy absence for the Acropolis Museum is the sculptures removed by Lord Elgin in the nineteenth century and currently displayed in the British Museum.

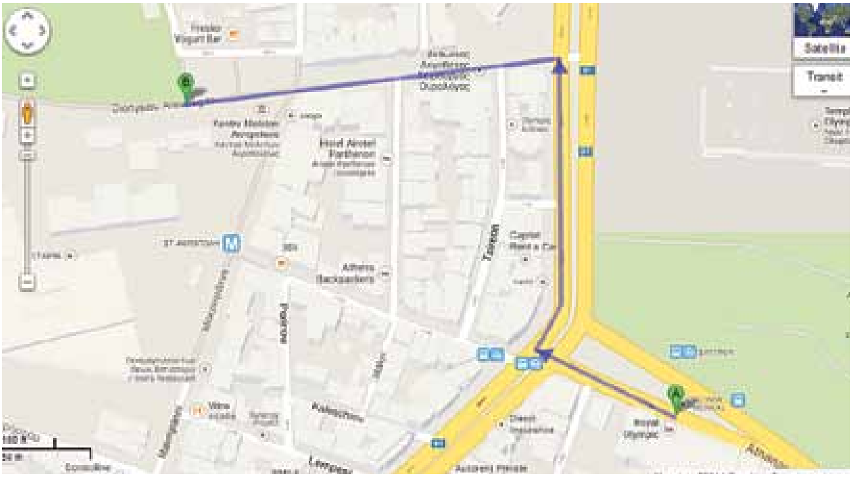
The museum is directly linked to the archaeological site of the Acropolis and to the extensive conservation work carried out on the sacred rock. It is under the supervision of the First Ephorate of Prehistoric and Classical Antiquities of the Ministry of Culture, which also oversees the construction of the New Acropolis Museum.

The new Acropolis Museum offers all the amenities expected in an international museum of the 21st century. The restaurant disposes a 700 square meter public terrace commanding a breathtaking view of the historic hills of Athens and proposes a variety of hot dishes based on traditional recipes, seasonally updated. According to a recent evaluation of the online site VirtualTourist.com the Acropolis restaurant appears in the Top 5 list of worldwide Museum Restaurants.

Meeting Point for on demand guided tour at the Acropolis Museum: 19:00 at the Hotel reception

- **Address: 15 Dionysiou Areopagitou Street, Athens 11742, Greece**

The Acropolis Museum is located in the historical area of Makriyianni, southeast of the Rock of the Acropolis, on Dionysiou Areopagitou Str, only a few meters from the Acropolis hill. The Museum entrance is located at the one end of the pedestrian walkway of Dionysiou Areopagitou Str., which constitutes the central route for the unified network of the city’s archaeological sites and by its own is considered one of the highlights of Athens. The ‘Acropolis’ Metro station is just on the east side of the Museum site.



From the Conference venue: Getting to the Acropolis Museum from the Royal Olympic Hotel is only 500 meters walk. Turn left onto *Athanassiou Diakou Str*., cross *Syngrou Ave*., continue straight onto *Athanassiou Diakou Str*., turn right onto *Stratigou Makriyianni Str*., and turn left onto *Dionysiou Areopagitou Str*.

Wednesday 26th March, from 19:30: Social Dinner at the Royal Olympic Hotel

The social dinner will be served at the “Ioannis Restaurant” in the roof garden of Royal Olympic Hotel in front of one of the most spectacular views of Athens.

The aperitif will be served starting from 19:30, and the dinner will be served from 20:30.

■ **Address: 28-34 Athanasiou Diakou Street, Athens, Greece**

Practical Information

Money exchange, currency

Euro is the official currency in Greece. Money exchange is available in most of banks and financial institutions. There are plenty of cash dispensers in Athens. Major international credit cards are widely but not always accepted in Greece (check in advance in small restaurants), and are not commonly used for small amounts. Banks are open from Monday to Friday until 2.00/2.30 pm, with flexible opening hours in the afternoon.

Electricity supply

Electricity in Greece, 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

Athens has a large, modern mass transit system to serve the needs of residents and visitors. It consists of city buses, electric trolley-buses, underground (“Metro”), tramway, and suburban railway (“Proastiakos”).

- **Metro:** From Athens International Airport “Eleftherios Venizelos” to the Conference venue. You will board the Metro Line 3 (Ag Marina – Douk. Plakentias – Airport), from the Airport’s Station and get off at Syntagma Station. Trains run every 30 minutes, 7 days a week from 6:30 a.m. to 11:30 p.m. The trip from/to the Airport to Syntagma station (Athens center) lasts 40 minutes. At Syntagma Station you switch lines in the direction of Elliniko (Line 2) and get off at the first Station, the Acropolis Station (see timetable at www.stasy.gr/index.php?id=70&L=1).
- **Tram** connects the city center with the southern seaside. There are 3 lines: Line 1 «Syntagma – SEF« linking downtown Athens to the Peace and Friendship Stadium, Line 2 «Syntagma – Voula» which runs between the city center and southern suburb of Voula and Line 3 «Voula – SEF» running along the coastal zone. Operating hours are 5:30 a.m. to 1:00 a.m. and on Fridays and Saturdays it operates

Omonoia Metro Station

approximately from 5:30 a.m. to 2:30 a.m.(see www.stasy.gr/index.php?id=33&no_cache=1&L=1).

- **Proastiakos:** The Suburban railway connects the Athens airport with the Athens Central Railway Station and Acharnai Railway Center, and through them to the National Railway network. Current timetables: Airport – Kiato and Kiato – Airport, Kiato – Patra bus connections, Piraeus – Athens – Halkida line. The Suburban railway departs every 15-25 minutes from the Athens Airport railway station to Plakentias station, where you can change trains and continue to the city center (Metro Line 3 to Peristeri), using the same ticket.
- **24-hour express buses:** All buses leave passengers at the Departures Level and depart from the Arrivals Level, between Exits 4 and 5. There are four routes:
  X93 : Kifissos KTEL (long-distance buses) Station - Airport
  X95 : Syntagma – Airport
  X96 : Port of Piraeus – Airport
  X97 : Eliniko Metro Station – Airport
  One-way travel time estimates: X93 (65 min), X95 (70 min), X96 (90 min), X97 (100 min). Allow sufficient time to travel as traffic conditions may cause delays.

Useful tips:

- While inside the trains, announcements are made before every stop.
- Tickets must be bought before boarding at the metro/tram stations cashiers or the bus terminals and by automatic tellers.
- Remember to validate your ticket

before you move on to the train platform or in the designated machines (orange colored) within the vehicles.

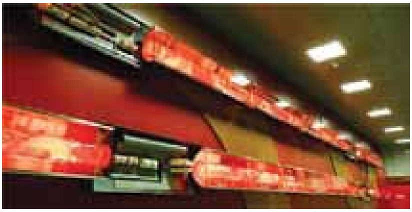
- Using any kind of transportation in Athens area cost 1,40€, for a duration of up to one and half-hours. Metro ticket for the airport cost 8€ (special prices for groups are also available).
- You may change transportation type, as many times you like within this time; you must validate your ticket at the beginning of the first ride.
- To stop a bus for embarkation you must make a hand signal to the driver. To disembark you have to notify the driver by pressing the stop button in time.

Taxis

A taxi from the airport to the city center costs a flat rate of €35 from 5:00 a.m. to midnight, and €50 from midnight to 5:00 a.m.

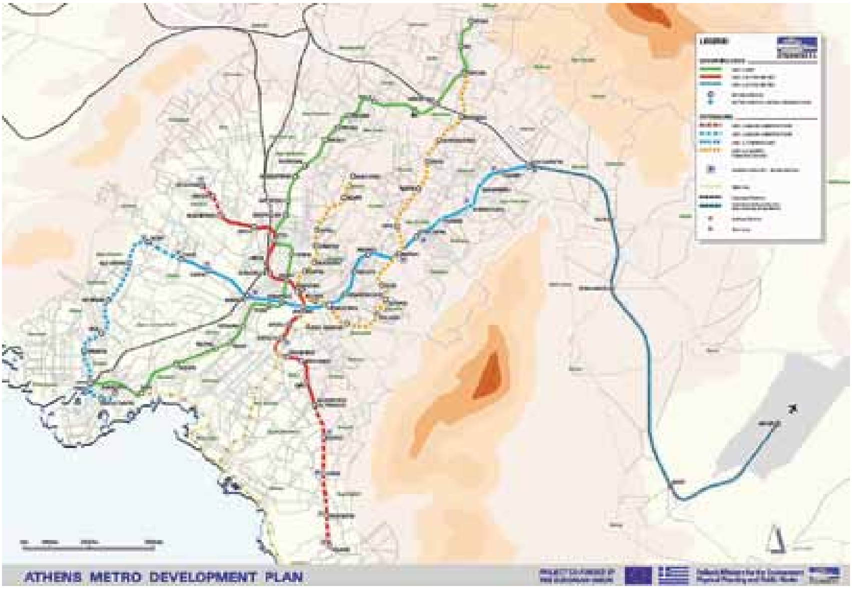
International calls

Dial 00 + country code + area code + phone number.
The international code number from abroad is +30 followed by the number code of the person you are calling, comprehensive of the "210" of Athens region.



Kerameikos Metro Station

Monastiraki Metro Station



Emergency calls

Medical Emergency Number	112
Ambulance Service	166
Marine Police Immediate intervention	108
Police - Immediate Response	100
Immediate social help	197
Counterterrorism agency	1014
Police Departments, Tel. Call center	1033, 10400
Emergencies Hospitals, Pharmacies	14944
Poisoning center	210 7793777
Fire Brigade	199
Tourist police	171
Athens traffic police	210 5284000
Forest fire service	191

Service of Lost or Stolen Credit Cards

American Express	210 339 7250
Diners	210 929 0200
Euro card	210 950 3673
Mastercard	00800118870303
Visa	00800116380304

Smoking

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

2. Program

Program at a Glance

Monday, March 24, 2014	
8:30-	Registration [Desk at Olympia Foyer]
Room	Olympia, Attica
8:45-9:00	ICDT Opening Session [Olympia]
9:00-10:30	ICDT Keynote 1: Frank Neven [Olympia]
10:30-11:00	*Coffee Break* *[Olympia Foyer]*
11:00-12:30	ICDT Research 1 (Attica)
12:30-14:00	*Lunch Break* *[Olympia Foyer]*
14:00-15:30	ICDT Research 2 (Attica)
15:30-16:00	*Coffee Break* *[Olympia Foyer]*
16:00-18:00	ICDT Research 3: Best Paper Awards (Attica)

Tuesday, March 25, 2014

8:30-	Registration [Desk at Olympia Foyer]				
Room	Olympia	Attica	Templar's	Conference room 2	Demo Stands
8:45-9:00	EDBT Opening Session [Olympia]				
9:00-10:30	EDBT Keynote 1: Serge Abiteboul [Olympia]				
10:30-11:00	*Coffee Break [Olympia Foyer]*				
11:00-12:30	EDBT Test of Time Award ICDT Test of Time Award				
12:30-14:00	*Lunch Break [Olympia Foyer]*				
14:00-15:30	Tutorial 1: N. Anciaux, B. Nguyen, I. Sandu-Popa	ICDT Research 4	EDBT Research 1	EDBT Research 2	Demo Session 1-a
15:30-16:00	*Coffee Break [Olympia Foyer]*				
16:00-17:30	ICDT Invited Lecture: Christian Bizer	EDBT Research 3	EDBT Research 4	EDBT Research 5	Demo Session 2-a
19:30-	Welcome Reception				

Wednesday, March 26, 2014

	Registration [Desk at Olympia Foyer]				
Rooms	Olympia	Attica	Templar’s	Conference room 2	Demo Stands
8:45-9:00	EDBT Opening Session [Olympia]				
9:00-10:30	ICDT Keynote 2: Christopher Re [Olympia]				
10:30-11:00	*Coffee Break* *[Olympia Foyer]*				
11:00-12:30	Tutorial 2: C. Mohan	ICDT Research 5	Industry & Applications 1	EDBT Research 6	
12:30-14:00	*Lunch Break* *[Olympia Foyer]*				
14:00-15:30	EDBT Vision Papers	ICDT Research 6	Industry & Applications 2	EDBT Research 7	Demo Session 2-b
15:30-16:00	*Coffee Break* *[Olympia Foyer]*				
16:00-18:00	EDBT Research 8	ICDT Research 7	Industry & Applications 3	EDBT Research 9	Demo Session 1-b
19:30-	Social Dinner				

Thursday, March 27, 2014

	Registration [Desk at Olympia Foyer]		
Rooms	Olympia	Attica	Templar’s
9:00-10:30	EDBT Keynote 2: Peter Boncz [Olympia]		
10:30-11:00	*Coffee Break* *[Olympia Foyer]*		
11:00-12:30	Tutorial 3: A. Artikis, G. Paliouras, A. Gal	EDBT Research 10	EDBT Research 11
12:30-14:00	*Lunch Break* *[Olympia Foyer]*		
14:00-15:30	Tutorial 3: A. Artikis, G. Paliouras, A. Gal	EDBT Research 12	EDBT Research 13
15:30-16:00	*Coffee Break* *[Olympia Foyer]*		
16:00-18:00	EDBT Research 14	EDBT Research 15	

Friday, March 28, 2014

Rooms	Olympia 1	Olympia 2	Attica	Templar’s	Conference room 2	Conference room 1A	Conference room 1B
Workshops	LWDM, GraphQ	ExploreDB	BeyondMR	BX, EnDM	MUD	PAIS	MSDM
8:45-9:00	Opening	Opening	Opening	BX Opening	Opening		Opening
9:00-10:30	Keynote Session	Keynote	Invited Talk & Session 1	BX Session 1	Session 1	Welcome Invited Talk Session 1	Keynote
10:30-11:00	*Coffee Break [Olympia Foyer]*						
11:00-12:30	Session 1	Session 1	Session 2	BX Session 2	Session 2	Session 2	Session 1
12:30-14:00	*Lunch Break [Olympia Foyer]*						
14:00-15:30	Session 2	Session 2	Invited Talk Session 3	EnDM Opening & EnDM Session1	Session 3	Invited Talk 2 Session 3	
15:30-16:00	*Coffee Break [Olympia Foyer]*						
16:00-17:30	Demo / Poster Session	Panel	Session 4	EnDM Session2	Session 4	Panel	

Detailed Program

Monday, March 24

8:45-9:00	ICDT Opening Session		

9:00-10:30	Keynote		**Frank Neven**
Room:	Olympia	Chair:	*Thomas Schwentick*

Remaining CALM in declarative networking

11:00-12:30	ICDT Research Session 1		***Distributed Computations***
Room:	Attica	Chair:	*Christian Bizer*

Foto Afrati, Anish Das Sarma, Anand Rajaraman, Pokey Rule, Semih Salihoglu and Jeffrey Ullman
Anchor-Points Algorithms for Hamming and Edit Distances Using MapReduce

Antoine Amarilli, Yael Amsterdamer and Tova Milo
On the Complexity of Mining Itemsets from the Crowd Using Taxonomies

Arvind Arasu and Raghav Kaushik
Oblivious Query Processing

14:00-15:30	ICDT Research Session 2		***Dynamic Complexity and XML***
Room:	Attica	Chair:	*Frank Neven*

Thomas Zeume and Thomas Schwentick
Dynamic Conjunctive Queries

Diego Figueira, Santiago Figueira and Carlos Areces
Basic Model Theory of XPath on Data Trees

Claire David, Piotr Hofman, Filip Murlak and Michal Pilipczuk
Synthesizing transformations from XML schema mappings

16:00-18:00	ICDT Research Session 3		***Best Paper Awards***
Room:	Attica	Chair:	*Nicole Schweikardt*

ICDT'14 Best Paper Award: Matthias Niewerth and Thomas Schwentick
Reasoning about XML Constraints based on XML-to-relational mappings

ICDT'14 Best Student Paper Award: Tom Ameloot
Deciding Correctness with Fairness for Simple Transducer Networks

ICDT'14 Best Newcomer Award: Todd Veldhuizen
Leapfrog Triejoin: A Simple, Worst-Case Optimal Join Algorithm

Tuesday, March 25

9:00 - 10:30	EDBT Opening Session		*Vassilis Christophides*
	Keynote		**Serge Abiteboul**
Room:	Olympia	Chair:	*Sihem Amer-Yahia*

Knowledge out there on the Web

11:00-12:30	ICDT EDBT		***Test of Time Awards***
Room:	Olympia	Chair:	Christine Collet Thomas Schwentick

ICDT Test of Time Award: Val Breazu-Tannen (now Tannen), Peter Buneman, and Limsoon Wong
Naturally Embedded Query Languages, ICDT 1992

EDBT Test of Time Award: Charu C. Aggarwal, and Philip S. Yu
A Condensation Approach to Privacy Preserving Data Mining, EDBT 2004

14:00-15:30	Tutorial Session 1	**N. Anciaux, B. Nguyen, I. Sandu-Popa**
Room:	Olympia	

Managing Personal Data with Strong Privacy Guarantees

14:00-15:30	ICDT Research Session 4		***Graph Databases***
Room:	Attica	Chair:	*Claire David*

Nadime Francis, Luc Segoufin and Cristina Sirangelo
Datalog Rewritings of Regular Path Queries using Views

Jelle Hellings
Conjunctive Context-Free Path Queries

Egor V. Kostylev, Juan L. Reutter and Domagoj Vrgoc
Containment of Data Graph Queries

14:00-15:30	EDBT Research Session 1		***Hadoop Optimization***
Room:	Templar’s	Chair:	Walif Aref

Mostafa Ead, Herodotos Herodotou, Ashraf Aboulnaga and Shivnath Babu
PStorM: Profile Storage and Matching for Feedback-Based Tuning of MapReduce Jobs

K. Ashwin Kumar, Jonathan Gluck, Amol Deshpande and Jimmy Lin
Optimization Techniques for “Scaling Down” Hadoop on Multi-Core, Shared-Memory Systems

Chuan Lei, Elke Rundensteiner and Mohamed Eltabakh
Redoop: Supporting Recurring Queries in Hadoop

14:00-15:30	EDBT Research Session 2		***Mapreduce Computation***
Room:	Conference room 2	Chair:	*Selcuk Candan*

Kasper Mullesgaard, Jens Laurits Pederseny, Hua Lu and Yongluan Zhou
Efficient Skyline Computation in MapReduce

Sergej Fries, Stephan Wels and Thomas Seidl
Projected Clustering for Huge Data Sets in MapReduce

Zhao Cao, Shimin Chen, Dongzhe Ma, Jianhua Feng and Min Wang
Efficient and Flexible Index Access in MapReduce

14:00-15:30	EDBT Demo Session 1-a	***Posters***
Room:	Olympia Foyer	

Jenny Rompa, Christos Tryfonopoulos, Costas Vassilakis and George Lepouras
Mindmap-Inspired Semantic Personal Information Management

Udayan Khurana, Srinivasan Parthasarathy and Deepak Turaga
READ: Rapid data Exploration, Analysis and Discovery

Max Heimel, Filip Haase, Martin Meinke, Sebastian Bre, Michael Saecker and Volker Markl
Demonstrating Self-Learning Algorithm Adaptivity in a Hardware-Oblivious Database Engine

Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos and Christos Tryfonopoulos
SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms

Robert Gottstein, Thorsten Peter, Ilia Petrov and Alejandro Buchmann
SIAS-V in Action: Snapshot Isolation Append Storage - Vectors on Flash

Silvana Castano, Alfio Ferrara and Stefano Montanelli
inWalk: Interactive and Thematic Walks inside the Web of Data

Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg and Ingeborg Torvik Solvberg

KIEV: a Tool for Extracting Semantic Relations from the World Wide Web

Gregory Smits, Olivier Pivert, Helene Jaudoin and Francois Paulus
AGGREGO SEARCH: Interactive Keyword Query Construction

16:00-17:30	ICDT Invited Lecture		**Christian Bizer**
Room:	Olympia	Chair:	*Nicole Schweikardt*

Search Joins with the Web

16:00-17:30	EDBT Research Session 3		***Stream and Parallel Processing***
Room:	Attica	Chair:	*Parth Nagarkar*

Ahmed M. Aly, Walid G. Aref, Mourad Ouzzani and Hosam M. Mahmoud
JISC: Adaptive Stream Processing Using Just-In-Time State Completion

Paolo Bellavista, Antonio Corradi, Spyros Kotoulas and Andrea Reale
Adaptive Fault-Tolerance for Dynamic Resource Provisioning in Distributed Stream Processing Systems

Steffen Zeuch, Johann-Christoph Freytag and Frank Huber
Adapting Tree Structures for Processing with SIMD Instructions

16:00-17:30	EDBT Research Session 4		***Multi-Queries and Concurrent Queries***
Room:	Templar's	Chair:	*Vasilis Vassalos*

Jennie Duggan, Olga Papaemmanouil, Ugur Cetintemel and Eli Upfal
Contender: A Resource Modeling Approach for Concurrent Query Performance Prediction

Venkatesh Raghavan and Elke Rundensteiner
CAQE: A Contract Driven Approach to Processing Concurrent Decision Support Queries

Shiwen Cheng, Anastasios Arvanitis, Marek Chrobak and Vagelis Hristidis
Multi-Query Diversification in Microblogging Posts

16:00-17:30	EDBT Research Session 5		***Provenance, Corroboration and Information Extraction***
Room:	Conference room 2	Chair:	*Grigoris Karvounarakis*

Nicole Bidoit, Melanie Herschel and Katerina Tzompanaki
Query-Based Why-Not Provenance with NedExplain

Minji Wu and Amélie Marian
Corroborating Facts from Affirmative Statements

Zhixu Li, Hongsong Li, Haixun Wang, Yi Yang, Xiangliang Zhang and Xiaofang Zhou
Overcoming Semantic Drift in Information Extraction

16:00-17:30	EDBT Demo Session 2-a	***Posters***
Room:	Olympia Foyer	

Roberto De Virgilio, Antonio Maccioni and Riccardo Torlone
R2G: a Tool for Migrating Relations to Graphs

Michael Gubanov and Michael Stonebraker
Large-scale Semantic Profile Extraction

Michael Pitts, Swapna Savvana, Senjuti Basu Roy and Vani Mandava
ALIAS: Author Disambiguation in Microsoft Academic Search Engine Dataset

Dong Wang, Lei Zou and Dongyan Zhao
gst-Store: An Engine for Large RDF Graph Integrating Spatiotemporal Information

Jishang Wei, Georgia Koutrika and Shanchan Wu
Learn2Learn: A Visual Educational System for Study Planning

Qi Li, Yuanyuan Xue, Jia Jia and Ling Feng
Helping Teenagers Relieve Psychological Pressures: A Micro-blog Based System

Clement Caron, Bernd Amann, Camelia Constantin and Patrick Giroux
WePIGE: The WebLab Provenance Information Generator and Explorer

Felix Beier, Nedal Alaqraa, Yuting Lai and Kai-Uwe Sattler
Learning Event Patterns for Gesture Detection

Wednesday, March 26

9:00-10:30	Keynote		**Christopher Ré**
Room:	Olympia	Chair:	*Nicole Schweikardt*

The Relational Join: New Theory and New Applications

11:00-12:30	Tutorial Session 2		**C. Mohan**
Room:	Olympia		

An In-Depth Look at Modern Database Systems

11:00-12:30	ICDT Research Session 5		***Uncertainty***
Room:	Attica	Chair:	*Claire David*

Serge Abiteboul, Daniel Deutch and Victor Vianu
Deduction with Contradictions in Datalog

Dany Maslowski and Jef Wijsen
Counting Database Repairs that Satisfy Conjunctive Queries with Self-Joins

Paraschos Koutris and Dan Suciu
A Dichotomy on the Complexity of Consistent Query Answering for Atoms with Simple Keys

11:00-12:30	EDBT Industrial Session 1		***Applications***
Room:	Templar’s	Chair:	Dimitris Kotzinos

Henrietta Dombrovskaya and Richard Lee
Talking to the Database in a Semantically Rich Way

Jagan Sankaranarayanan, Hakan Hacigumus, Haopeng Zhang and Mohamed Sarwat
SMILE: A Data Sharing Platform for Mobile Apps in the Cloud

Wei Tan, Sandeep Tata, Yuzhe Tang and Liang Fong
Diff-Index: Differentiated Index in Distributed Log-Structured Data Stores

11:00-12:30	EDBT Research Session 6		***Text and Sequence Mining***
Room:	Conference room 2	Chair:	*Maria Luisa Sapino*

Abhishek Mukherji, Elke Rundensteiner and Matthew Ward
COLARM: Cost-based Optimization for Localized Association Rule Mining

Deepak P, Atreyee Dey and Debapriyo Majumdar
Fast Mining of Interesting Phrases from Subsets of Text Corpora

Yongluan Zhou, Chunyang Ma, Qingsong Guo, Lidan Shou and Gang Chen
Sequence Pattern Matching over Time-Series Data with Temporal Uncertainty

14:00-15:30	ICDT Research Session 6		***Probabilistic Databases and Provenance***
Room:	Attica	Chair:	*Nicole Schweikardt*

Paul Beame, Jerry Li, Sudeepa Roy and Dan Suciu
Model Counting of Query Expressions: Limitations of Propositional Methods

Christopher Ré
The Theory of Zeta Graphs with an Application to Random Networks

Daniel Deutch, Tova Milo, Sudeepa Roy and Val Tannen
Circuits for Datalog Provenance

14:00-15:30	EDBT Vision Papers		***EDBT Vision Track***
Room:	Olympia	Chair:	Maurice Van Keulen

Thomas Heinis, Farhan Tauheed and Anastasia Ailamaki
Spatial Data Management Challenges in the Simulation Sciences

Michael Johnson, Jorge Pérez and James Terwilliger
What Can Programming Languages Say About Data Exchange?

David Broneske, Sebastian Breß, Max Heimel and Gunter Saake
Toward Hardware-Sensitive Database Operations

14:00-15:30	EDBT Research Session 7		***Distributed Query Processing***
Room:	Conference room 2	Chair:	*Norman Paton*

Siqiang Luo, Yifeng Luo, Shuigeng Zhou, Gao Cong and Jihong Guan
Distributed Spatial Keyword Querying on Road Networks

Johannes Niedermayer, Mario Nascimento, Matthias Renz, Peer Kröger and Hans-Peter Kriegel
Cost-Based Median Query Processing in Wireless Sensor Networks

George Tsatsanifos, Dimitris Sacharidis and Timos Sellis
RIPPLE: A Scalable Framework for Distributed Processing of Rank Queries

14:00-15:30	EDBT Industrial Session 2		***Applications***
Room:	Templar’s	Chair:	Yannis Stavrakas

Alexander Artikis, Matthias Weidlich, Francois Schnitzler, Ioannis Boutsis, Thomas Liebig, Nico Piatkowski, Christian Bockermann, Katharina Morik, Vana Kalogeraki, Jakub Marecek, Avigdor Gal, Shie

Mannor, Dimitrios Gunopulos and Dermot Kinane
Heterogeneous Stream Processing and Crowdsourcing for Urban Traffic Management

Maria Daltayanni, Luca de Alfaro, Panagiotis Papadimitriou and Panayiotis Tsaparas
On Assigning Implicit Reputation Scores in an Online Labor Marketplace

Khalid Belhajjame
Annotating the Behavior of Scientific Modules Using Data Examples: A Practical Approach

14:00-15:30	EDBT Demo Session 2-b	***Posters***
Room:	Olympia Foyer	

Roberto De Virgilio, Antonio Maccioni and Riccardo Torlone
R2G: a Tool for Migrating Relations to Graphs

Michael Pitts, Swapna Savvana, Senjuti Basu Roy and Vani Mandava
ALIAS: Author Disambiguation in Microsoft Academic Search Engine Dataset

Dong Wang, Lei Zou and Dongyan Zhao
gst-Store: An Engine for Large RDF Graph Integrating Spatiotemporal Information

Jishang Wei, Georgia Koutrika and Shanchan Wu
Learn2Learn: A Visual Educational System for Study Planning

Michael Gubanov and Michael Stonebraker
Large-scale Semantic Profile Extraction

Qi Li, Yuanyuan Xue, Jia Jia and Ling Feng
Helping Teenagers Relieve Psychological Pressures: A Micro-blog Based System

Clement Caron, Bernd Amann, Camelia Constantin and Patrick Giroux
WePIGE: The WebLab Provenance Information Generator and Explorer

Felix Beier, Nedal Alaqraa, Yuting Lai and Kai-Uwe Sattler
Learning Event Patterns for Gesture Detection

16:00-18:00	ICDT Research Session 7		***Algorithms and Queries***
Room:	Attica	Chair:	*Wim Martens*

Yeye He, Siddharth Barman and Jeffrey Naughton
On Load Shedding in Complex Event Processing

Andreas Kosmatopoulos, Kostas Tsichlas and Apostolos N. Papadopoulos
Dynamic Processing of Dominating Queries with Performance Guarantees

Simone Bova and Hubie Chen
The Complexity of Width Minimization for Existential Positive Queries

Francois Picalausa, George Fletcher, Jan Hidders and Stijn Vansummeren
Principles of Guarded Structural Indexing: On Guarded Simulations and Acyclic First Order Languages

16:00-18:00	EDBT Research Session 8		***Indexing and Cost Statistics***
Room:	Olympia	Chair:	*Stefan Manegold*

Parth Nagarkar and K. Selcuk Candan
HCS: Hierarchical Cut Selection for Efficiently Processing Queries on Data Columns using Hierarchical Bitmap Indices

Ingo Müller, Cornelius Ratsch and Franz Färber
Adaptive String Dictionary Compression in In-Memory Column-Store Database Systems

Cigdem Aslay, Nicola Barbieri, Francesco Bonchi and Ricardo Baeza-Yates
Online Topic-aware Influence Maximization Queries

Ramanujam S Halasipuram, Prasad M Deshpande and Sriram Padmanabhan
Determining Essential Statistics for Cost Based Optimization of an ETL Workflow

16:00-18:00	EDBT Research Session 9		***Matrix Factorization, Clustering and Probabilistic Data***
Room:	Conference room 2	Chair:	*Amélie Marian*

Chenghui Ren, Luyi Mo, Ben Kao, Reynold Cheng and David W. Cheung
CLUDE: An Efficient Algorithm for LU Decomposition Over a Sequence of Evolving Graphs

Mojgan Pourrajabi, Davoud Moulavi, Ricardo Campello, Arthur Zimek, Jörg Sander and Randy Goebel
Model Selection for Semi-Supervised Clustering

Tarique Anwar, Chengfei Liu, Hai L Vu and Christopher Leckie
Spatial Partitioning of Large Urban Road Networks

Sebastiaan van Schaik, Dan Olteanu and Robert Fink
ENFrame: A Platform for Processing Probabilistic Data

16:00-17:30	EDBT Industrial Session 3		***Applications***
Room:	Templar's	Chair:	Grigoris Karvounarakis

Martin Kaufmann, Peter Fischer, Norman May and Donald Kossmann
Benchmarking Bitemporal Database Systems: Ready for the Future or Stuck in the Past?

Jaroslaw Szlichta, Parke Godfrey, Jarek Gryz, Wenbin Ma, Weinan Qiu and Calisto Zuzarte
Business-Intelligence Queries with Order Dependencies in DB2

Souripriya Das, Jagannathan Srinivasan, Matthew Perry, Eugene Chong and Jayanta Banerjee
A Tale of Two Graphs: Property Graphs as RDF in Oracle

16:00-17:30	EDBT Demo Session 1-b	***Posters***
Room:	Olympia Foyer	

Jenny Rompa, Christos Tryfonopoulos, Costas Vassilakis and George Lepouras
Mindmap-Inspired Semantic Personal Information Management

Udayan Khurana, Srinivasan Parthasarathy and Deepak Turaga
READ: Rapid data Exploration, Analysis and Discovery

Max Heimel, Filip Haase, Martin Meinke, Sebastian Bre, Michael Saecker and Volker Markl
Demonstrating Self-Learning Algorithm Adaptivity in a Hardware-Oblivious Database Engine

Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos and Christos Tryfonopoulos
SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms

Robert Gottstein, Thorsten Peter, Ilia Petrov and Alejandro Buchmann
SIAS-V in Action: Snapshot Isolation Append Storage - Vectors on Flash

Silvana Castano, Alfio Ferrara and Stefano Montanelli
inWalk: Interactive and Thematic Walks inside the Web of Data

Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg and Ingeborg Torvik Solvberg
KIEV: a Tool for Extracting Semantic Relations from the World Wide Web

Gregory Smits, Olivier Pivert, Helene Jaudoin and Francois Paulus
AGGREGO SEARCH: Interactive Keyword Query Construction

Thursday, March 27

9:00-10:30	Keynote		**Peter Boncz**
Room:	Olympia	Chair:	*Georgia Koutrika*

Benchmarking Graph Data Management Systems

11:00-12:30	Tutorial Session 3	**A. Artikis, G. Paliouras**
Room:	Olympia	

Formal Methods for Event Processing

11:00-12:30	EDBT Research Session 10		***Keyword Search and Diversity***
Room:	Attica	Chair:	*Senjuti Basu Roy*

Chengyuan Zhang, Ying Zhang, Wenjie Zhang, Xuemin Lin, Muhammad Cheema and Xiaoyang Wang
Diversified Spatial Keyword Search On Road Networks

Bettina Fazzinga, Sergio Flesca, Filippo Furfaro and Francesco Parisi
Cleaning trajectory data of RFID-monitored objects through conditioning under integrity constraints

Ji Zhang, Wei-Shinn Ku, Min-Te Sun, Xiao Qin and Hua Lu
Multi-Criteria Optimal Location Query with Overlapping Voronoi Diagrams

11:00-12:30	EDBT Research Session 11		***Ranking***
Room:	Templar’s	Chair	*Bernd Amann*

Anastasios Arvanitis, Matthew Wiley and Vagelis Hristidis
Efficient Concept-based Document Ranking

Eleftherios Tiakas, George Valkanas, Apostolos N. Papadopoulos and Yannis Manolopoulos
Metric-Based Top-k Dominating Queries

Muhammad Cheema, Zhitao Shen, Xuemin Lin and Wenjie Zhang
A Unified Framework for Efficiently Processing Ranking Related Queries

14:00-15:30	Tutorial Session 3	**A. Artikis, G. Paliouras**
Room:	Olympia	

Formal Methods for Event Processing

14:00-15:30	EDBT Research Session 12		***Joins***
Room:	Attica	Chair:	*Torsten Grust*

Andrey Gubichev and Thomas Neumann
Exploiting the query structure for efficient join ordering in SPARQL queries

Angela Bonifati, Radu Ciucanu and Slawek Staworko
Interactive Inference of Join Queries

Bhupesh Chawda and Himanshu Gupta
Processing Interval Joins On Map-Reduce

14:00-15:30	EDBT Research Session 13		***Privacy-Aware Data Processing***
Room:	Templar's	Chair:	*Alfredo Cuzzocrea*

Haoran Li, Li Xiong and Xiaoqian Jiang
Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions

Quoc-Cuong To, Benjamin Nguyen and Philippe Pucheral
Privacy-Preserving Query Execution using a Decentralized Architecture and Tamper Resistant Hardware

Haohan Zhu, Xianrui Meng and George Kollios
Privacy Preserving Similarity Evaluation of Time Series Data

16:00-18:00	EDBT Research Session 14		***Graph Queries and Analytics***
Room:	Olympia	Chair:	*Venkatesh Raghavan*

Renê R. Veloso, Loïc Cerf, Wagner Meira Junior and Mohammed J. Zaki
Reachability Queries in Very Large Graphs: A Fast Refined Online Search Approach

Dritan Bleco and Yannis Kotidis
Graph Analytics on Massive Collections of Small Graphs

Arijit Khan, Francesco Bonchi, Aris Gionis and Francesco Gullo
Fast Reliability Search in Uncertain Graphs

Francesco Bonchi, Aristides Gionis, Francesco Gullo and Antti Ukkonen
Distance oracles in edge-labeled graphs

16:00-18:00	EDBT Research Session 15		***Privacy in Networks***
Room:	Attica	Chair:	*Irini Fundulaki*

Tamir Tassa and Francesco Bonchi
Privacy Preserving Estimation of Social Influence

Zach Jorgensen and Ting Yu
A Privacy-Preserving Framework for Personalized, Social Recommendations

Sadegh Nobari, Panagiotis Karras, Hweehwa Pang and Stéphane Bressan
L-opacity: Linkage-Aware Graph Anonymization

Aston Zhang, Xing Xie, Kevin Chen-Chuan Chang, Carl A. Gunter, Jiawei Han and Xiaofeng Wang
Privacy Risk in Anonymized Heterogeneous Information Networks

4. Keynotes

Date: Monday, March 24

Time: 9:00-10:30

Room:
Olympia

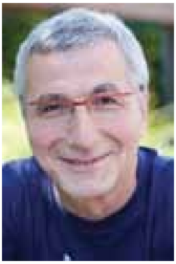
Speaker:
Frank Neven

Title: Remaining CALM in declarative networking

Abstract:

Declarative networking is an approach where distributed computations and networking protocols are modeled and programmed using formalisms based on Datalog. In his keynote speech at PODS 2010, Hellerstein made a number of intriguing conjectures concerning the expressiveness of declarative networking. One of those became popular under the name of the CALM conjecture (Consistency And Logical Monotonicity) and suggests a strong link between, on the one hand, eventually consistent and coordination-free distributed computations, and on the other hand, expressibility in monotonic Datalog. In this keynote, I will discuss recent results concerning the CALM conjecture and will point out future directions and challenges for the community.

Bio: Frank Neven is a full professor at Hasselt University. His main research area is that of database theory and systems. Recent research interests include datamining in bioinformatics, automatic schema inference, big data, cloud computing, and logic in databases. A significant part of his research has been devoted to applying automata theoretic techniques to the analysis of XML processing.



Date: Tuesday, March 25

Time: 9:00-10:30

Room:
Olympia

Speaker:
Serge Abiteboul

Title: Knowledge out there on the Web

Abstract:

The Web is turning from a collection of documents to a word of systems exchanging interlinked knowledge. This is radically changing the user's perspective who can now actively participate both in person or via systems (phones, cars, home applicances, etc.) We will survey different ways of acquiring knowledge on the Web. We will mention recent works on distributed knowledge management that addressed issues such as inconsistencies or privacy. We will also discuss novel technical challenges that arise.

Bio: Serge Abiteboul obtained his Ph.D. from the University of Southern California, and a Doctoral Thesis from the University of Paris-Sud. He became a researcher at the Institut National de Recherche en Informatique et Automatique in 1982. He was a Lecturer at the École Polytechnique and Visiting Professor at Stanford and Oxford University. He has been Chaire Professor at Collège de France in 2011-12 and Francqui Chair Professor at Namur University in 2012-2013. He co-founded the company Xyleme in 2000. Serge Abiteboul has received the ACM SIGMOD Innovation Award in 1998, the EADS Award from the French Academy of sciences in 2007 and the Milner Award from the Royal Society in 2013. He became a member of the French Academy of Sciences in 2008, and a member the Academy of Europe in 2011. He is currently PI of the Webdam ERC (2008-2013) and a member of the LSV at the École Normale Supérieure, Cachan. His research work focuses mainly on data, information and knowledge management, particularly on the Web. He is a member of the Conseil national du numérique and Chairman of the Scientific board of the Société d'Informatique de France.

Date: Wednesday, March 26 **Time:** 9:00-10:30

Room:
Olympia

Speaker:
Christopher Ré

Title: The Relational Join: New Theory and New Applications

Abstract:

A favorite of database theoreticians for decades, the relational join is back in the spotlight. Almost all major data processing systems (even many formerly "NoSQL" systems) implement some flavor of join processing. Theoretical understanding of join processing has taken a step forward. We have learned that the dogma of "one join at a time" leads to suboptimal running time; we understand how to perform joins in MapReduce environments, and we have learned that there are even more deep connections between join processing, discrete geometry, and constraint satisfaction. This talk will survey some of the recent results leading to these findings. Additionally, we will describe some unexpected applications of join processing in probabilistic inference that my research group has been using to build knowledge bases.

Bio: Christopher (Chris) Re is an assistant professor in the Department of Computer Science at Stanford University. The goal of his work is to enable users and developers to build applications that more deeply understand and exploit data. Chris received his PhD from the University of Washington in Seattle under the supervision of Dan Suciu. For his PhD work in probabilistic data management, Chris received the SIGMOD 2010 Jim Gray Dissertation Award. Chris's papers have received four best-paper or best-of-conference citations, including best paper in PODS 2012, best-of-conference in PODS 2010 twice, and one best-of-conference in ICDE 2009. Chris received an NSF CAREER Award in 2011 and an Alfred P. Sloan fellowship in 2013.

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

Room: Olympia

Speaker: Peter Boncz

Title: Benchmarking Graph Data Management Systems

Abstract:

Many so-called “Big Data” data management problems revolve around analysis of heterogeneous, and complexly structured data sets where the data is interrelated, thus forming a graph that connects billions of nodes. Here, the worth of the data and its analysis is not only in the attribute values of these nodes, but in the way these nodes are connected. Specific application areas that exhibit the growing need for management of such graph-shaped data include life science analytics, social network marketing and digital forensics. In this keynote, I will make the case that in order to evaluate pros and cons of using new emerging technology such as graph database systems or distributed graph programming frameworks, as well as more established technology such as MapReduce or even traditional (parallel) database systems, it is important to develop a new generation of benchmarks. In this context I will describe some of the work being conducted in the LDBC (Linked Data Benchmark Council - ldbc.eu), focusing on so-called “choke point” based benchmark development.

Bio: Peter Boncz holds appointments as tenured researcher at CWI and full professor at VU University Amsterdam. His academic background is in core database architecture, and his work focuses on the interaction between computer architecture and data management techniques. His specific contributions are in cache-conscious join methods, query processing in columnar database systems (specifically MonetDB), and vectorized query execution (Vectorwise). He received the Dutch ICT Regie Award 2006 for his role in the CWI spin-off company Data Distilleries. In 2008 he founded the CWI spin-off company around Vectorwise. He is also the co-recipient of the 2009 VLDB 10 Years Best Paper Award, and in 2013 received the Humboldt Research Award. Peter Boncz currently is the scientific director of the EU project LDBC, that aims to establish industry-strength graph and RDF benchmarks and benchmark practices.

5. ICDT Invited Lecture

Date: Tuesday, March 25 **Time:** 16:00-17:30

Room:
Olympia

Speaker:
Christian Bizer

Title:
Search Joins with the Web

Abstract:

The talk will discuss the concept of Search Joins. A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web in various formats. The challenge for Search Joins is to decide which Web tables to join with the local table in order to deliver high-quality results. Search joins are useful in various application scenarios. They allow for example a local table about cities to be extended with an attribute containing the average temperature of each city for manual inspection. They also allow tables to be extended with large sets of additional attributes as a basis for data mining, for instance to identify factors that might explain why the inhabitants of one city claim to be happier than the inhabitants of another.

In the talk, Christian Bizer will draw a theoretical framework for Search Joins and will highlight how recent developments in the context of Linked Data, RDFa and Microdata publishing, public data repositories as well as crowd-sourcing integration knowledge contribute to the feasibility of Search Joins in an increasing number of topical domains.

Bio: Christian Bizer explores questions concerning the development of global, decentralized information environments. His current research focus is the evolution of the World Wide Web from a medium for the publication of documents into a global data space. Christian Bizer has initialized the W3C Linking Open Data community effort which is interlinking large numbers of data sources on the Web. He also co-founded the DBpedia project which derives a comprehensive knowledge base from Wikipedia. Further results of his work include the Named Graphs data model which was adopted into the W3C SPARQL recommendation, the D2RQ mapping language which is widely used for publishing relational databases on the Web of Linked Data, the Silk - Linking Framework, and the Berlin SPARQL Benchmark. Christian Bizer holds an appointment as full professor at University of Mannheim. Before moving to Mannheim, he headed the Web-based Systems Group at Freie Universitat Berlin.

6. Tutorials

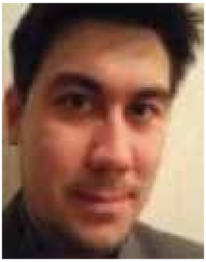
Managing Personal Data with Strong Privacy Guarantees

Presenters: Nicolas Anciaux, Benjamin Nguyen and Iulian Sandu-Popa

Managing personal data with strong privacy guarantees has become an important topic in an age where your glasses record and share everything you see, your wallet records and shares your financial transactions, and your set-top box records and shares your energy consumption. More and more alternatives are proposed based on user centric and decentralized solutions, capitalizing on the use of trusted personal devices controlling the data at the edges of the Internet.

In this tutorial, we review existing solutions and present a functional architecture encompassing such alternatives. We then expose the underlying techniques proposed recently and the open issues dealing with embedded data management in secure client devices and global query processing within a user centric architecture. We then conclude by presenting implementations of this approach, as the prefiguration of what can be perceived as the holy grail of personal data management.

Nicolas Anciaux is a researcher at INRIA Paris-Rocquencourt, France. He received his Ph.D. from University of Versailles in 2004 and was in 2005 and 2006 a researcher at University of Twente, Netherlands. His main areas of interest are core database systems, embedded databases, database security and privacy.



Benjamin Nguyen is Associate Professor at University of Versailles St-Quentin (UVSQ), member of the CNRS PRiSM Lab and INRIA Secured and Mobile Information Systems (SMIS) team. He received his Ph.D. from University of Paris-XI in 2003, joined UVSQ in 2004 and INRIA-SMIS in 2010. His current research topics revolve around privacy protection in data centric applications, personal data over-exposure and anonymization.

Iulian Sandu Popa is Assistant Professor in Computer Science at the University of Versailles Saint-Quentin (UVSQ) and member of INRIA-SMIS since 2012. He received his Ph.D. in Computer Science from UVSQ in 2009. His main research interests are embedded database management systems, spatiotemporal databases, and mobile data management, with a particular interest in topics revolving around privacy and personal data management.

An In-Depth Look at Modern Database Systems

Presenters: C. Mohan

This tutorial is targeted at a broad set of database systems and applications people. It is intended to let the attendees better appreciate what is really behind the covers of many of the modern database systems (e.g., NoSQL and NewSQL systems), going beyond the hype associated with these open source, commercial and research systems. The capabilities and limitations of such systems will be addressed. Modern extensions to decades old relational DBMSs will also be described. Some application case studies will also be presented.

Dr. C. Mohan has been an IBM researcher for 32 years in the information management area, impacting numerous IBM and non-IBM products, the research community and standards, especially with his invention of the ARIES family of locking and recovery algorithms, and the Presumed Abort commit protocol. This IBM, ACM and IEEE Fellow has also served as the IBM India Chief Scientist. In addition to receiving the ACM SIGMOD Innovation Award, the VLDB 10 Year Best Paper Award and numerous IBM awards, he has been elected to the US and Indian National Academies of Engineering, and has been named an IBM Master Inventor. This distinguished alumnus of IIT Madras received his PhD at the University of Texas at Austin. He is an inventor of 40 patents. He serves on the IBM Software Group Architecture Board's Council. More information can be found in his home page at http://bit.ly/CMohan

Formal Methods for Event Processing

Presenters: Alexander Artikis, Georgios Paliouras

Today's organisations require techniques for automated transformation of the large data volumes they collect into operational knowledge. This requirement may be addressed by employing event processing systems that detect activities/events of special significance within an organisation, given streams of `low-level' information that are very difficult to be utilised by humans. Numerous event processing approaches have been proposed in the literature. In this tutorial, we will review formal methods for event processing and discuss the open research issues of this field. We will present temporal reasoning systems, systems that explicitly deal with uncertainty, and machine learning techniques automating the construction and refinement of event structures. To illustrate the reviewed approaches we will use real-world case studies, including event processing for city transport and traffic management.

Alexander Artikis is a research associate in NCSR Demokritos (Athens, Greece). He holds a PhD from Imperial College London on norm-governed multi-agent systems, while his research interests lie in the areas of artificial intelligence and distributed systems. Alexander has been working on several international projects on event processing; currently, he is the technical director of the FP7 SPEEDD project that develops a Big Data system for proactive event-driven decision-making. Alexander has also developed a highly scalable, logic-based, open-source event processing system.

Georgios Paliouras is a senior researcher in NCSR Demokritos (Athens, Greece). He holds a PhD from Manchester University on machine learning for event recognition. He has performed basic and applied research in machine learning for the last 17 years. He is involved in many European and national research projects and has the role of scientific coordinator in some of them, including the FP7 SPEEDD project. He has given a number of invited talks and tutorials at various institutions and conferences, such as ECML/PKDD, DEBS and IJCAI.

ICDT Research Sessions

Monday, March 24			
11:00-12:30	ICDT Research Session 1		***Distributed Computations***
Room:	Attica	Chair:	*Christian Bizer*

Anchor-Points Algorithms for Hamming and Edit Distances Using MapReduce

Foto Afrati, Anish Das Sarma, Anand Rajaraman, Pokey Rule, Semih Salihoglu and Jeffrey Ullman

Algorithms for computing similarity joins in MapReduce were offered in [2]. Similarity joins ask to find input pairs that are within a certain distance d according to some distance measure. Here we explore the “anchor-points algorithm” of [2]. We continue looking at Hamming distance, and show that the method of that paper can be improved; in particular, if we want to find strings within Hamming dis- tance d, and anchor points are chosen so that every possible input is within Hamming distance k of some anchor point, then it is sufficient to send each input to all anchor points within distance (d/2)+k, rather than d+k as was suggested in the earlier paper. This improves on the communication cost of the MapReduce algorithm, i.e., reduces the amount of data transmitted among machines. Further, the same holds for edit distance, provided inputs all have the same length n and either the length of all anchor points is n − k or the length of all anchor points is n + k. We then explore the problem of finding small sets of anchor points for edit distance, which also provides an improvement on the communication cost. We give a close-to-optimal technique to extend anchor sets (called “covering codes”) from the k = 1 case to any k. We then give small covering codes that use either a single deletion or a single insertion, or – in one al- gorithm – two deletions. Discovering covering codes for edit distance is important in its own right, since very little work is known.

On the Complexity of Mining Itemsets from the Crowd Using Taxonomies

Antoine Amarilli, Yael Amsterdamer and Tova Milo

We study the problem of frequent itemset mining in domains where data is not recorded in a conventional database but only exists in human knowledge. We provide examples of such scenarios, and present a crowdsourcing model for them. The model uses the crowd as an oracle to find out whether an itemset is frequent or not, and relies on a known taxonomy of the item domain to guide the search for frequent itemsets. In the spirit of data mining with oracles, we analyze the complexity of this problem in terms of (i) crowd complexity, that measures the number of crowd questions required to identify the frequent itemsets; and (ii) computational complexity, that measures the computational effort required to choose the questions. We provide lower and upper complexity bounds in terms of the size and structure of the input taxonomy, as well as the size of a concise description of the output itemsets. We also provide constructive algorithms that achieve the upper bounds, and consider more efficient variants for practical situations.

Oblivious Query Processing

Arvind Arasu and Raghav Kaushik

Motivated by cloud security concerns, there is an increasing interest in database systems that can store and support queries over encrypted data. A common architecture for such systems is to use a trusted component such as a cryptographic co-processor for query processing that is used to securely decrypt data and perform computations in plaintext. The trusted component has limited memory, so most of the (input and intermediate) data is kept encrypted in an untrusted storage and moved to the trusted component on “demand.”

In this setting, even with strong encryption, the data access pat- tern from untrusted storage has the potential to reveal sensitive in- formation; indeed, all existing systems that use a trusted component for query processing over encrypted data have this vulnerability. In this paper, we undertake the first formal study of secure query processing, where an adversary having full knowledge of the query (text) and observing the query execution learns nothing about the underlying database other than the result size of the query on the database. We introduce a simpler notion, oblivious query processing, and show formally that a query admits secure query processing iff it admits oblivious query processing. We present oblivious query processing algorithms for a rich class of database queries involving selections, joins, grouping and aggregation. For queries not handled by our algorithms, we provide some initial evidence that designing oblivious (and therefore secure) algorithms would be hard via reductions from two simple, well-studied problems that are generally believed to be hard. Our study of oblivious query processing also reveals interesting connections to database join theory.

Monday, March 24			
14:00-15:30	ICDT Research Session 2		***Dynamic Complexity and XML***
Room:	Attica	Chair:	*Frank Neven*

Dynamic Conjunctive Queries

Thomas Zeume and Thomas Schwentick

The paper investigates classes of queries maintainable by conjunctive queries (CQs) and their extensions and restrictions in the dynamic complexity framework of Patnaik and Immerman. It studies the impact of union, atomic negation and quantification on the dynamic expressiveness of CQ, for the standard semantics as well as for Δ-semantics.

It turns out that, although there are many different combinations of these features, there exist basically five important fragments for the standard semantics, characterized by the addition of the following features to the possibility to build conjunctions over positive atoms: (1) arbitrary quantification and atomic negation, (2) existential quantification and atomic negation, (3) existential quantification, (4) atomic negation (but no quantification), and (5) conjunction only (and no quantification). Whether all these fragments are actually different remains mostly open, however, it is shown that (4) strictly subsumes (5). The fragments arising from Δ-semantics are also subsumed by the standard fragments (1), (2) and (4).

As a further result, all (statically) FO-definable queries are captured by fragment (2) and a complete characterization of these queries in terms of non-recursive dynamic CQ¬-programs is given.

Basic Model Theory of XPath on Data Trees

Diego Figueira, Santiago Figueira and Carlos Areces

We investigate model theoretic properties of XPath with data (in)equality tests over the class of data trees, i.e., the class of trees where each node contains a label from a finite alphabet and a data value from an infinite domain.

We provide notions of (bi)simulations for XPath logics containing the child, descendant, parent and ancestor axes to navigate the tree. We show that these notions precisely characterize the equivalence relation associated with each logic. We study formula complexity measures consisting of the number of nested axes and nested subformulas in a formula; these notions are akin to the notion of quantifier rank in first-order logic. We show characterization results for fine grained notions of equivalence and (bi)simulation that take into account these complexity measures. We also prove that positive fragments of these logics correspond to the formulas preserved under (non-symmetric) simulations. We show that the logic including the child axis is equivalent to the fragment of first-order logic invariant under the corresponding notion of bisimulation. If upward navigation is allowed the characterization fails but a weaker result can still be established. These results hold over the class of possibly infinite data trees and over the class of finite data trees.

Besides their intrinsic theoretical value, we argue that bi- simulations are useful tools to prove (non)expressivity results for the logics studied here, and we substantiate this claim with examples.

Synthesizing transformations from XML schema mappings

Claire David, Piotr Hofman, Filip Murlak and Michal Pilipczuk

XML schema mappings have been developed and studied in the context of XML data exchange, where a source document has to be restructured under the target schema according to certain rules. The rules are specified with a mapping, which consists of a set of source-to-target dependencies based on tree patterns. The problem of building a target document for a given source document and a mapping has polynomial data complexity, but is still intractable due to high combined complexity.

We consider a two layer architecture for building target in- stances, inspired by the Church synthesis problem. We view the mapping as a specification of a document transformation, for which an implementation must be found. The static layer inputs a map- ping and synthesizes a single XML-to-XML query implementing a valid transformation. The data layer amounts to evaluating this query on a given source document, which can be done by a specialized query engine, optimized to handle large documents.

We show that for a given mapping one can synthesize a query expressed in an XQuery-like language, which can be evaluated in time proportional to the evaluation time of the patterns used in the mapping. In general the involved constant is high, but it can be improved under additional assumptions. In terms of overall complexity, if the arity of patterns is considered constant, we obtain a fixed-parameter tractable procedure with respect to the mapping size, which improves previously known upper bounds.

Monday, March 24			
16:00-18:00	ICDT Research Session 3		***Best Paper Awards***
Room:	Attica	Chair:	Nicole Schweikardt

ICDT'14 Best Paper Award: Reasoning about XML Constraints based on XML-to-relational mappings

Matthias Niewerth and Thomas Schwentick

The paper introduces a simple framework for the specification of constraints for XML documents in which constraints are specified by (1) a mapping that extracts a relation from every XML document and (2) a relational constraint on the resulting relation. The mapping has to be generic with respect to the actual data values and the relational constraints can be of any kind. Besides giving a general undecidability result for first-order definable mappings and a general decidability result for MSO definable mappings for restricted functional dependencies, the paper studies the complexity of the implication problem for XML constraints that are specified by tree pattern queries and functional dependencies. Furthermore, it highlights how other specification languages for XML constraints can be formulated in the framework.

ICDT’14 Best Student Paper Award: Deciding Correctness with Fairness for Simple Transducer Networks

Tom Ameloot

Ensuring the correctness of a distributed system is an important challenge. Previously, two interpretations of correct- ness have been proposed: the first interpretation is about determinism, saying that all infinite fair computation traces produce the same output; and, the second interpretation is a confluence notion, saying that all finite computation traces can still be extended to produce the same output. A decidability result for the confluence notion was previously obtained for so-called simple transducer networks, a model from the field of declarative networking. In the cur- rent paper, we also present a decidability result for simple transducer networks, but this time for the first interpretation of correctness, with infinite fair computation traces. We also compare the expressivity of simple transducer networks under both interpretations.

ICDT’14 Best Newcomer Award: Leapfrog Triejoin: A Simple, Worst-Case Optimal Join Algorithm

Todd Veldhuizen

Recent years have seen exciting developments in join algorithms. In 2008, Atserias, Grohe and Marx (henceforth AGM) proved a tight bound on the maximum result size of a full conjunctive query, given constraints on the input relation sizes. In 2012, Ngo, Porat, R´e and Rudra (henceforth NPRR) devised a join algorithm with worst-case running time proportional to the AGM bound [8]. Our commercial database system LogicBlox employs a novel join algorithm, leapfrog triejoin, which compared conspicuously well to the NPRR algorithm in preliminary benchmarks. This spurred us to analyze the complexity of leapfrog triejoin. In this paper we establish that leapfrog triejoin is also worst-case optimal, up to a log factor, in the sense of NPRR. We improve on the results of NPRR by proving that leapfrog triejoin achieves worst-case optimality for finer-grained classes of database instances, such as those defined by constraints on projection cardinalities. We show that NPRR is not worst- case optimal for such classes, giving a counterexample where leapfrog triejoin runs in O(n log n) time and NPRR runs in Θ(n1.375) time. On a practical note, leapfrog triejoin can be implemented using conventional data structures such as B-trees, and extends naturally to $\exists 1$ queries. We believe our algorithm offers a useful addition to the existing toolbox of join algorithms, being easy to absorb, simple to implement, and having a concise optimality proof.

Tuesday, March 25			
14:00-15:30	ICDT Research Session 4		***Graph Databases***
Room:	Attica	Chair:	*Claire David*

Datalog Rewritings of Regular Path Queries using Views

Nadime Francis, Luc Segoufin and Cristina Sirangelo

We consider query answering views on graph databases, i.e. databases structured as edge-labeled graphs. We consider views and queries specified by Regular Path Queries. These are queries selecting pairs of nodes in a graph database that are connected via a path whose sequence of edge labels belongs to some regular language.

A view V determines a query Q if for all graph databases D, the view image VpDq always contains enough information to answer Q on D. In other words, there is a well defined function from VpDq to QpDq.

Our main result shows that when this function is mono- tone, there exists a rewriting of Q as a Datalog query over the view instance VpDq. In particular the query can be evaluated in time polynomial in the size of VpDq.

As a side result we also prove that it is decidable whether an RPQ query can be rewritten in Datalog using RPQ views.

Conjunctive Context-Free Path Queries

Jelle Hellings

In graph query languages, regular expressions are commonly used to specify the labeling of paths. A natural step in in- creasing the expressive power of these query languages is re- placing regular expressions by context-free grammars. With the

Conjunctive Context-Free Path Queries (CCFPQ) we introduce such a language based on the well-known Con- junctive Regular Path Queries (CRPQ).

First, we show that query evaluation of CCFPQ has polynomial time data complexity. Secondly, we look at the generalization of regular expressions, as used in CRPQ, to regular relations and show how similar generalizations can be applied to context-free grammars, as used in CCFPQ. Thirdly, we investigate the relations between the expressive power of CRPQ, CCFPQ, and their generalizations. In several cases we show that replacing regular expressions by context-free grammars does increase expressive power. Finally, we look at including context-free grammars in more powerful logics than conjunctive queries. We do so by adding negation and provide expressivity relations between the obtained languages.

Containment of Data Graph Queries

Egor V. Kostylev, Juan L. Reutter and Domagoj Vrgoc

The graph database model is currently one of the most popular paradigms for storing data, used in applications such as social networks, biological databases and the Semantic Web. Despite the popularity of this model, the development of graph database management systems is still in its infancy, and there are several fundamental issues regarding graph databases that are not fully understood. Indeed, while graph query languages that concentrate on topological properties are now well developed, not much is known about languages that can query both the topology of graphs and their underlying data.

Our goal is to conduct a detailed study of static analysis problems for such languages. In this paper we consider the containment problem for several recently proposed classes of queries that manipulate both topology and data: regular queries with memory, regular queries with data tests, and graph XPath. Our results show that the problem is in general undecidable for all of these classes. However, by allowing only positive data comparisons we find natural fragments that enjoy much better static analysis properties: the containment problem is decidable, and its computational complexity ranges from PSPACE-complete to EXPSPACE-complete. We also propose extensions of regular queries with an inverse operator, and study query evaluation and query containment for them.

Wednesday, March 26			
11:00-12:30	ICDT Research Session 5		***Uncertainty***
Room:	Attica	Chair:	*Claire David*

Deduction with Contradictions in Datalog

Serge Abiteboul, Daniel Deutch and Victor Vianu

We study deduction in the presence of inconsistencies. Following previous works, we capture deduction via datalog programs and inconsistencies through violations of functional dependencies (FDs). We study and compare two semantics for datalog with FDs: the first, of a logical nature, is based on inferring facts one at a time, while never violating the FDs; the second, of an operational nature, consists in a fixpoint computation in which maximal sets of facts consistent with the FDs are inferred at each stage.

Both semantics are nondeterministic, yielding sets of possible worlds. We introduce a PTIME (in the size of the extensional data) algorithm, that given a datalog program, a set of FDs and an input instance, produces a c-table representation of the set of possible worlds. Then, we propose to quantify nondeterminism with probabilities, by means of a probabilistic semantics. We consider the problem of capturing possible worlds along with their probabilities via probabilistic c-tables.

We then study classical computational problems in this novel context. We consider the problems of computing the probabilities of answers, of identifying most likely supports for answers, and of determining the extensional facts that are most influential for de- riving a particular fact. We show that the interplay of recursion and FDs leads to novel technical challenges in the context of these problems.

Counting Database Repairs that Satisfy Conjunctive Queries with Self-Joins

Dany Maslowski and Jef Wijsen

An uncertain database is defined as a relational database in which primary keys need not be satisfied. A block is a maximal subset of tuples of the same relation that agree on the primary key. A repair (or possible world) of an uncertain database is obtained by selecting exactly one tuple from each block. From a probabilistic database perspective, an uncertain database is a restricted kind of block-independent disjoint (BID) probabilistic database, where the restriction is that the probabilities of tuples in a block are equal and sum up to one.

For every fixed Boolean query q, the counting problem ♮CERTAINTY(q) takes as input an uncertain database db and asks to determine the number of repairs that satisfy q. A Boolean conjunctive query is self-join-free if no relation name occurs more than once in it. In previous work, it was proved that for every self-join-free Boolean conjunctive query q, the problem ♮CERTAINTY(q) is either in FP or ♮P-complete, and it is decidable which of the two cases applies. This complexity dichotomy has its analogue in BID probabilistic databases.

The current paper investigates the complexity of the problem ♮CERTAINTY(q) for Boolean conjunctive queries with self-joins. Our most appealing result is that for every Boolean conjunctive query q (possibly with self-joins) in which all primary keys consist of a single attribute, ♮CERTAINTY(q) is either in FP or ♮P-complete, and it is decidable which of the two cases applies. Significantly, no analogous dichotomy for conjunctive queries with self-joins is known for BID probabilistic databases.

A Dichotomy on the Complexity of Consistent Query Answering for Atoms with Simple Keys

Paraschos Koutris and Dan Suciu

We study the problem of consistent query answering under primary key violations. In this setting, the relations in a database violate the key constraints and we are interested in maximal subsets of the database that satisfy the constraints, which we call repairs. For a boolean query Q, the problem CERTAINTY(Q) asks whether every such repair satisfies the query or not; the problem is known to be always in coNP for conjunctive queries. However, there are queries for which it can be solved in polynomial time. It has been conjectured that there exists a dichotomy on the complexity of CERTAINTY(Q) for conjunctive queries: it is either in PTIME or coNP-complete. In this paper, we prove that the conjecture is indeed true for the case of conjunctive queries without self-joins, where each atom has as a key either a single attribute (simple key) or all attributes of the atom.

Wednesday, March 26			
14:00-15:30	ICDT Research Session 6		***Probabilistic Databases and Provenance***
Room:	Attica	Chair:	*Nicole Schweikardt*

Model Counting of Query Expressions: Limitations of Propositional Methods

Paul Beame, Jerry Li, Sudeepa Roy and Dan Suciu

Query evaluation in tuple-independent probabilistic databases is the problem of computing the probability of an answer to a query given independent probabilities of the individual tuples in a database instance. There are two main approaches to this problem: (1) in grounded inference one first obtains the lineage for the query and database instance as a Boolean formula, then performs weighted model counting on the lineage (i.e., computes the probability of the lineage given probabilities of its independent Boolean variables); (2) in methods known as lifted inference or extensional query evaluation, one exploits the high-level structure of the query as a first-order formula. Although it is widely believed that lifted inference is strictly more powerful than grounded inference on the lineage alone, no formal separation has previously been shown for query evaluation. In this paper we show such a formal separation for the first time.

We exhibit a class of queries for which model counting can be done in polynomial time using extensional query evaluation, whereas the algorithms used in state-of-the-art exact model counters on their lineages provably require exponential time. Our lower bounds on the running times of these exact model counters follow from new exponential size lower bounds on the kinds of d-DNNF representations of the lineages that these model counters (either explicitly or implicitly) produce. Though some of these queries have been studied before, no nontrivial lower bounds on the sizes of these representations for these queries were previously known.

The Theory of Zeta Graphs with an Application to Random Networks

Christopher Ré

Social, biological, and cyberphysical networks generate some of the most intriguing and valuable sources of data on the planet. For at least the last two decades, researchers have attempted to create formal (typically stochastic) models of these networks. We examine the database theory questions raised by these new models. We study a simple extension of Erdös–Rényi models that we call Zeta graphs. Zeta graphs are related to multiple-valued zeta functions, and we show that the expectation of a conjunctive query can be written as a linear combination of multiple-valued zeta functions. For queries on graphs, we use our results to devise a complete decision procedure for whether the probability that a query is true tends to 1 as the domain size tends to infinity. We apply our theory of Zeta graphs to describe the set of conjunctive graph queries that are true with probability 1 in another graph model in the literature that was described by Callaway, Hopcroft, Kleinberg, Newman, and Strogatz.

Circuits for Datalog Provenance

Daniel Deutch, Tova Milo, Sudeepa Roy and Val Tannen

The annotation of the results of database queries with provenance information has many applications. This paper studies provenance for datalog queries. We start by considering provenance representation by (positive) Boolean expressions, as pioneered in the theories of incomplete and probabilistic databases. We show that even for linear datalog programs the representation of provenance using Boolean expressions incurs a super-polynomial size blowup in data complexity. We address this with an approach that is novel in provenance studies, showing that we can construct in PTIME poly-size (data complexity) provenance representations as Boolean circuits. Then we present optimization techniques that embed the construction of circuits into semi- naive datalog evaluation, and further reduce the size of the circuits. We also illustrate the usefulness of our approach in multiple application domains such as query evaluation in probabilistic databases, and in deletion propagation. Next, we study the possibility of extending the circuit approach to the more general framework of semiring annotations introduced in earlier work. We show that for a large and useful class of provenance semirings, we can construct in PTIME poly-size circuits that capture the provenance.

Wednesday, March 26			
16:00-18:00	ICDT Research Session 7		***Algorithms and Queries***
Room:	Attica	Chair:	*Wim Martens*

On Load Shedding in Complex Event Processing

Yeye He, Siddharth Barman and Jeffrey Naughton

Complex Event Processing (CEP) is a stream processing model that focuses on detecting event patterns in continuous event streams. While the CEP model has gained popularity in the research com- munities and commercial technologies, the problem of gracefully degrading performance under heavy load in the presence of re- source constraints, or load shedding, has been largely overlooked. CEP is similar to "classical" stream data management, but addresses a substantially different class of queries. This unfortunately renders the load shedding algorithms developed for stream data processing inapplicable. In this paper we study CEP load shedding under various resource constraints. We formalize broad classes of CEP load-shedding scenarios as different optimization problems. We demonstrate an array of complexity results that reveal the hardness of these problems and construct shedding algorithms with performance guarantees. Our results shed some light on the difficulty of developing load-shedding algorithms that maximize utility.

Dynamic Processing of Dominating Queries with Performance Guarantees

Andreas Kosmatopoulos, Kostas Tsichlas and Apostolos N. Papadopoulos

The top-k dominating query returns the k database objects with the highest score with respect to their dominance score. The dominance score of an object p is simply the number of objects dominated by p, based on minimization or maximization preferences on the attribute values. Each object (tuple) is represented as a point in a multidimensional space, and therefore, the number of attributes equals the number of dimensions. The top-k dominating query combines the dominance concept of skyline queries with the ranking function of top-k queries and can be used as an important tool in multi-criteria decision making systems. In this work, we focus on the 2-dimensional space and present, for the first time, novel algorithms for top-k dominating query processing in main memory with non-trivial asymptotic guarantees. In particular, we focus on both the semi-dynamic case (only insertions are allowed) and the fully-dynamic case (insertions and deletions are supported). We perform a detailed cost analysis regarding the worst-case complexity of preprocessing, the worst-case complexity for the query cost and the worst-case and amortized complexity for updates (insertions and deletions) focusing on the RAM computation model. Our solutions require space linear with the number of points, which is very important especially for modern applications which manipulate massive datasets. In addition, we discuss the case of the word-RAM computation model, where slightly better results are obtained.

The Complexity of Width Minimization for Existential Positive Queries

Simone Bova and Hubie Chen

Existential positive queries are logical sentences built from conjunction, disjunction, and existential quantification, and are also known as select-project-join-union queries in data- base theory, where they are recognized as a basic and fundamental class of queries. It is known that the number of variables needed to express an existential positive query is the crucial parameter determining the complexity of evaluating it on a database, and is hence a natural measure from the perspective of query optimization and rewriting. In this article, we study the complexity of the natural decision problem associated to this measure, which we call the expressibility problem: Given an existential positive query and a number k, can the query be expressed using k (or fewer) variables? We precisely determine the complexity of the expressibility problem, showing that it is complete for the level Πp2 of the polynomial hierarchy. Moreover, we prove that the expressibility problem is undecidable in positive logic (that is, existential positive logic plus universal quantification), thus establishing existential positive logic as a maximal syntactic fragment where expressibility is decidable.

Principles of Guarded Structural Indexing

Francois Picalausa, George Fletcher, Jan Hidders and Stijn Vansummeren

We present a new structural characterization of the expressive power of the acyclic conjunctive queries in terms of guarded simulations, and give a finite preservation theorem for the guarded simulation invariant fragment of first order logic.

We discuss the relevance of these results as a formal basis for constructing so-called guarded structural indexes. Structural indexes were first proposed in the context of semi- structured query languages and later successfully applied as an XML indexation mechanism for XPath-like queries on trees and graphs. Guarded structural indexes provide a generalization of structural indexes from graph databases to relational databases.

EDBT Research Sessions

Tuesday, March 25			
14:00-15:30	EDBT Research Session 1		***Hadoop Optimization***
Room:	Templar’s	Chair:	*Walif Aref*

PStorM: Profile Storage and Matching for Feedback-Based Tuning of MapReduce Jobs

Mostafa Ead, Herodotos Herodotou, Ashraf Aboulnaga and Shivnath Babu

The MapReduce programming model has become widely adopted for large scale analytics on big data. MapReduce systems such as Hadoop have many tuning parameters, many of which have a significant impact on performance. The map and reduce functions that make up a MapReduce job are developed using arbitrary programming constructs, which make them black-box in nature and therefore renders it difficult for users and administrators to make good parameter tuning decisions for a submitted MapReduce job. An approach that is gaining popularity is to provide automatic tuning decisions for submitted MapReduce jobs based on feedback from previously executed jobs. This approach is adopted, for ex- ample, by the Starfish system. Starfish and similar systems base their tuning decisions on an execution profile of the MapReduce job being tuned. This execution profile contains summary information about the runtime behavior of the job being tuned, and it is assumed to come from a previous execution of the same job. Managing these execution profiles has not been previously studied. This paper presents PStorM, a profile store and matcher that accurately chooses the relevant profiling information for tuning a submitted MapReduce job from the previously collected profiling information. PStorM can identify accurate tuning profiles even for previously unseen MapReduce jobs. PStorM is currently integrated with the Starfish system, although it can be extended to work with any MapReduce tuning system. Experiments on a large number of MapReduce jobs demonstrate the accuracy and efficiency of pro- file matching. The results of these experiments show that the profiles returned by PStorM result in tuning decisions that are as good as decisions based on exact profiles collected during pervious ex- ecutions of the tuned jobs. This holds even for previously unseen jobs, which significantly reduces the overhead of feedback-driven profile-based MapReduce tuning.

Optimization Techniques for "Scaling Down" Hadoop on Multi-Core, Shared-Memory Systems

K. Ashwin Kumar, Jonathan Gluck, Amol Deshpande and Jimmy Lin

The underlying assumption behind Hadoop and, more generally, the need for distributed processing is that the data to be analyzed cannot be held in memory on a single machine. Today, this assumption needs to be re-evaluated. Although petabyte-scale datastores are increasingly common, it is unclear whether "typical" analytics tasks require more than a single high-end server. Additionally, we are seeing increased sophistication in analytics, e.g., machine learning, where we process smaller and more refined datasets. To address these trends, we propose "scaling down" Hadoop to run on multi-core, shared-memory machines. This paper presents a proto- type runtime called Hone ("Hadoop One") that is API compatible with Hadoop. With Hone, we can take an existing Hadoop application and run it efficiently on a single server. This allows us to take existing MapReduce algorithms and find the most suitable runtime environment for execution on datasets of varying sizes. For dataset sizes that fit into memory on a single machine, our experiments show that Hone is substantially faster than Hadoop running in pseudo-distributed mode. In some cases, Hone running on a single machine outperforms a 16-node Hadoop cluster.

Redoop: Supporting Recurring Queries in Hadoop

Chuan Lei, Elke Rundensteiner and Mohamed Eltabakh

The growing demand for large-scale data analytics ranging from online advertisement placement, log processing, to fraud detection, has led to the design of highly scalable data-intensive computing infrastructures such as the Hadoop platform. Recurring queries, repeatedly being executed for long periods of time on rapidly evolving high-volume data, have become a bedrock component in most of these analytic applications. Despite their importance, the plain Hadoop along with its state-of-art extensions lack built-in support for recurring queries. In particular, they lack efficient and scalable analytics over evolving datasets. In this work, we present the Redoop system, an extension of the Hadoop framework, designed to fill in this void. Redoop supports recurring queries as first- class citizen in Hadoop without sacrificing any of its core features. More importantly, Redoop deploys innovative window-aware opti- mization techniques for recurring query execution including adaptive window-aware data partitioning, window-aware task scheduling, and inter-window caching mechanisms. Redoop retains the fault-tolerance of MapReduce via automatic cache recovery and task re-execution support. Our extensive experimental study with real datasets demonstrates that Redoop achieves significant run- time performance gains of up to 9x speedup compared to the plain Hadoop.

Tuesday, March 25			
14:00-15:30	EDBT Research Session 2		***MapReduce Computation***
Room:	Conference room 2	Chair:	*Selçuk Candan*

Efficient Skyline Computation in MapReduce

Kasper Mullesgaard, Jens Laurits Pederseny, Hua Lu and Yongluan Zhou

Skyline queries are useful for finding interesting tuples from a large data set according to multiple criteria. The sizes of data sets are constantly increasing and the architecture of back-ends are switching from single-node environments to non-conventional paradigms like MapReduce. Despite the usefulness of skyline queries, existing works on skyline computation in MapReduce do not take full advantage of parallelism but still run significant parts serially. In this paper, we propose a novel approach to compute skylines efficiently in MapReduce. We design a grid partitioning scheme to di- vide the data space into partitions, and employ a bitstring to represent the partitions. The bitstring is efficiently obtained in MapReduce, and it clearly helps prune partitions (and tuples) that cannot have skyline tuples. Based on the grid partitioning, we propose two MapReduce algorithms to compute skylines. Both algorithms utilize the bitstring and dis- tribute the original tuples to multiple mappers and make use of them to compute local skylines in parallel. In particular, MapReduce Grid Partitioning based Single-Reducer Skyline

Computation (MR-GPSRS) employs a single reducer to assemble the local skylines appropriately to compute the glob- al skyline. In contrast, MapReduce Grid Partitioning based Multiple Reducer Skyline Computation (MR-GPMRS) further divides local skylines and distributes them to multiple reducers that compute the global skyline in an independent and parallel manner. The proposed algorithms are evaluated through extensive experiments, and the results show that MR-GPMRS significantly outperforms the alternatives in various settings.

Projected Clustering for Huge Data Sets in MapReduce

Sergej Fries, Stephan Wels and Thomas Seidl

Fast growing data sets with a very high number of attributes be- come a common situation in social, industry and scientific areas. A meaningful analysis of these data sets requires sophisticated data mining techniques as projected clustering that are able to deal with such complex data.

In this work, we investigate solutions for extending the state-of-the- art projected clustering algorithm P3C for large data sets in high- dimensional spaces. We show that the original model of the P3C algorithm is not suitable to deal with huge data sets. Therefore, we propose the necessary changes of the underlying clustering model and then present an efficient MapReduce-based implementation - our novel P3C+-MR algorithm. The effectiveness of the proposed changes on large data sets and the efficiency of the P3C+-MR algorithm are comprehensively evaluated on synthetic and real-world data sets. Additionally, we propose the P3C+-MR-Light algorithm, a simplified version of P3C+-MR that shows extraordinary good results in terms of runtime and result quality on large data sets. In the end, we compare our solutions to existing approaches.

Efficient and Flexible Index Access in MapReduce

Zhao Cao, Shimin Chen, Dongzhe Ma, Jianhua Feng and Min Wang

A popular programming paradigm in the cloud, MapReduce is extensively considered and used for “big data” analysis. Unfortunately, a great many “big data” applications require capabilities be- yond those originally intended by MapReduce, often burdening developers to write unnatural non-obvious MapReduce programs so as to twist the underlying system to meet the requirements. In this paper, we focus on a class of "big data" applications that in addition to MapReduce's main data source, require selective access to one or many data sources, e.g., various kinds of indices, knowledge bases, external cloud services.

We propose to extend MapReduce with EFind, an Efficient and Flexible index access solution, to better support this class of applications. EFind introduces a standard index access interface to MapReduce so that (i) developers can easily and flexibly express index access operations without unnatural code, and (ii) the EFind enhanced MapReduce system can automatically optimize the in- dex access operations. We propose and analyze a number of index access strategies that utilize caching, re-partitioning, and index locality to reduce redundant index accesses. EFind collects index statistics and performs cost-based adaptive optimization to improve index access efficiency. Our experimental results, using both real- world and synthetic data sets, show that EFind chooses execution plans that are optimal or close to optimal, and achieves a factor of 2x–8x improvements compared to an approach that accesses in- dices without optimization.

Tuesday, March 25			
16:00-17:30	EDBT Research Session 3		***Stream and Parallel Processing***
Room:	Attica	Chair:	*Parth Nagarkar*

JISC: Adaptive Stream Processing Using Just-In-Time State Completion

Ahmed M. Aly, Walid G. Aref, Mourad Ouzzani and Hosam M. Mahmoud

The continuous and dynamic nature of data streams may lead a query execution plan (QEP) of a long-running continuous query to become suboptimal during execution, and hence will need to be altered. The ability to perform an efficient and flawless transition to an equivalent, yet optimal QEP is essential for a data stream query processor. Such transition is challenging for plans with stateful binary operators, such as joins, where the states of the QEP have to be maintained during query transition without compromising the correctness of the query output. This paper presents Just-In-Time State Completion (JISC); a new technique for query plan migration. JISC does not cause any halt to the query execution, and thus allows the query to maintain steady output. JISC is applicable to pipelined as well as eddy-based query evaluation frameworks. Probabilistic analysis of the cost and experimental studies show that JISC in- creases the execution throughput during the plan migration stage by up to an order of magnitude compared to existing solutions.

Adaptive Fault-Tolerance for Dynamic Resource Provisioning in Distributed Stream Processing Systems

Paolo Bellavista, Antonio Corradi, Spyros Kotoulas and Andrea Reale

A growing number of applications require continuous processing of high-throughput data streams, e.g., financial analysis, network traffic monitoring, or Big Data analytics for smart cities. Stream processing applications typically re- quire specific quality-of-service levels to achieve their goals; yet, due to the high time-variability of stream characteristics, it is often inefficient to statically allocate the resources needed to guarantee application Service Level Agreements (SLAs). In this paper, we present LAAR, a novel method for adaptive replication that trades fault tolerance for in- creased capacity during load spikes. We have implemented and validated LAAR as a middleware layer on top of IBM InfoSphere Streams©. We have performed a wide set of experiments on an industrial-quality 60-core cluster deployment and we show that, under the assumption of only statistical knowledge of streams load distribution, LAAR can reduce resource consumption while guaranteeing an upper-bound on information loss in case of failures.

Adapting Tree Structures for Processing with SIMD Instructions

Steffen Zeuch, Johann-Christoph Freytag and Frank Huber

In this paper, we accelerate the processing of tree-based index structures by using SIMD instructions. We adapt the B+-Tree and prefix B-Tree (trie) by changing the search algorithm on inner nodes from binary search to k-ary search. The k-ary search enables the use of SIMD instructions, which are commonly available on most modern processors today. The main challenge for using SIMD instructions on CPUs is their inherent requirement for consecutive memory loads. The data for one SIMD load instruction must be located in consecutive memory locations and cannot be scattered over the entire memory. The original layout of tree-based index structures does not satisfy this constraint and must be adapted to enable SIMD usage. Thus, we introduce two tree adaptations that satisfy the specific constraints of SIMD instructions. We present two different algorithms for trans- forming the original tree layout into a SIMD-friendly layout. Additionally, we introduce two SIMD-friendly search algorithms designed for the new layout.

Our adapted B+-Tree speeds up search processes by a factor of up to eight for small data types compared to the original B+-Tree using binary search. Furthermore, our adapted prefix B-Tree enables a high search performance even for larger data types. We report a constant 14 fold speedup and an 8 fold reduction in memory consumption compared to the original B+-Tree.

Tuesday, March 25			
16:00-17:30	EDBT Research Session 4		***Multi-Queries and Concurrent Queries***
Room:	Templar’s	Chair:	*Vasilis Vassalos*

Contender: A Resource Modeling Approach for Concurrent Query Performance Prediction

Jennie Duggan, Olga Papaemmanouil, Ugur Cetintemel and Eli Upfal

Predicting query performance under concurrency is a difficult task that has many applications in capacity planning, cloud computing, and batch scheduling. We introduce Contender, a new resource- modeling approach for predicting the concurrent query performance of analytical workloads. Contender’s unique feature is that it can generate effective predictions for both static as well as ad- hoc or dynamic workloads with low training requirements. These characteristics make Contender a practical solution for real-world deployment.

Contender relies on models of hardware resource contention to predict concurrent query performance. It introduces two key metrics, Concurrent Query Intensity (CQI) and Query Sensitivity (QS), to characterize the impact of resource contention on query interactions. CQI models how aggressively concurrent queries will use the shared resources. QS defines how a query’s performance changes as a function of the scarcity of resources. Contender integrates these two metrics to effectively estimate a query’s concurrent execution latency using only linear time sampling of the query mixes.

Contender learns from sample query executions (based on known query templates) and uses query plan characteristics to generate latency estimates for previously unseen templates. Our experimental results, obtained from PostgreSQL/ TPC-DS, show that Contender's predictions have an error of 19% for known templates and 25% for new templates, which is competitive with the state-of- the-art while requiring considerably less training time.

CAQE: A Contract Driven Approach to Processing Concurrent Decision Support Queries

Venkatesh Raghavan and Elke Rundensteiner

Real-time analytical systems need to handle workloads comprised of expensive decision support queries with diverse quality of service requirements known as contracts. Contract driven multi-query processing, being an NP-hard problem, remains unaddressed to date. The traditional approach of blindly pipelining the entire input through a shared execution plan is not viable due to the diversity in query contracts. To tackle this challenge, we now develop a flexible model to express contracts and accompany it with an effective means to measure the run-time contract satisfaction. We propose our Contract-Aware Query Execution framework CAQE. In this work, we exploit the principle that "different portions of the in- put contribute to disparate subsets of queries with varying degrees of progressiveness." Therefore, CAQE's processing of the input chunks is driven by how the different query contracts are being met at run-time. To maximize the contract satisfaction of the work- load, CAQE leverages the dependencies of input chunks across the queries. This enables us to determine the impact of processing particular input chunks on improving the run-time contract satisfaction. Our experiments demonstrate the effectiveness of CAQE in increasing the overall contract satisfaction of the workload, specifically 2 fold better than existing multi-query processing techniques.

Multi-Query Diversification in Microblogging Posts

Shiwen Cheng, Anastasios Arvanitis, Marek Chrobak and Vagelis Hristidis

Effectively exploring data generated by microblogging services is challenging due to its high volume and production rate. To ad- dress this issue, we propose a solution that helps users effectively consume information from a microblogging stream, by filtering out redundant data. We formalize our approach as a novel optimization problem termed Multi-Query Diversification Problem (MQDP). In MQDP, the input consists of a list of microblogging posts and a set of user queries (e.g. news topics), where each query matches a subset of posts. The objective is to compute the smallest subset of posts that cover all other posts with respect to a "diversity dimension" that may represent time or, say, sentiment. Roughly, the solution (cover) has the property that each covered post has nearby posts in the cover that are collectively related to all queries relevant to this covered post.

This is distinct from previous single-query diversity problems, as we may have two nearby posts that are related to intersecting but not nested sets of queries, in which case none covers the other. Another key difference is that we do not define diversity in terms of post similarity, since posts are too short for this approach to be meaningful; instead, we focus on finding representative posts for ordered diversity dimensions like time and sentiment, which are critical in microblogging. For example, for time as the diversity dimension, the selected posts will show how certain news events unfolded over time.

We prove that MQDP is NP-hard and we propose an exact dynamic programming algorithm that is feasible for small problem instances. We also propose two approximate algorithms with provable approximation bounds, and show how they can be adapted for a streaming setting. Through comprehensive experiments on real data, we show that our algorithms efficiently and effectively generate diverse and representative posts.

Tuesday, March 25			
16:00-17:30	EDBT Research Session 5		***Provenance, Corroboration and Information Extraction***
Room:	Conference room 2	Chair:	*Grigoris Karvounarakis*

Query-Based Why-Not Provenance with NedExplain

Nicole Bidoit, Melanie Herschel and Katerina Tzompanaki

With the increasing amount of available data and transformations manipulating the data, it has become essential to analyze and debug data transformations. A sub-problem of data transformation analysis is to understand why some data are not part of the result of a relational query. One possibility to explain the lack of data in a query result is to identify where in the query we lost data pertinent to the expected outcome. A first approach to this so called why-not provenance has been recently proposed, but we show that this first approach has some shortcomings.

To overcome these shortcomings, we propose NedExplain, an algorithm to explain data missing from a query result. NedExplain computes the why-not provenance for monotone relational queries with aggregation. After providing necessary definitions, this paper contributes a detailed description of the algorithm. A comparative evaluation shows that it is both more efficient and effective than the state-of-the-art approach.

Corroborating Facts from Affirmative Statements

Minji Wu and Amélie Marian

Web sources often provide different and even conflicting in- formation. Simple voting-based strategies have already shown limitations at identifying the correct answer to a user query with the presence of unreliable sources. In order to identify the correct answer, corroboration techniques have been pro- posed and proved to be effective for such tasks. In this paper, we investigate the corroboration problem in which most or all facts have only affirmative statements from sources. A fact is either true or false, and an affirmative statement from a source indicates its support for a fact being true. Unfortunately, state-of-the-art corroboration techniques rely on conflicting information to differentiate the trustworthiness of the sources and we demonstrate their limitations in our scenario. Different from existing techniques that consider a single trust score for each source, we propose a novel algorithm that utilizes a multi-value trust score toward different subsets of facts. By considering the information entropy of the unknown facts, our algorithm incrementally evaluates facts and updates the estimates on the trust scores for the sources. We conduct experiments using both synthetic and real-world datasets and demonstrate that our algorithm significantly outperforms existing approaches in precision and accuracy.

Overcoming Semantic Drift in Information Extraction

Zhixu Li, Hongsong Li, Haixun Wang, Yi Yang, Xiangliang Zhang and Xiaofang Zhou

Semantic drift is a common problem in iterative information extraction. Previous approaches for minimizing semantic drift may incur substantial loss in recall. We observe that most semantic drifts are introduced by a small number of questionable extractions in the earlier rounds of iterations. These extractions subsequently intro- duce a large number of questionable results, which lead to the semantic drift phenomenon. We call these questionable extractions Drifting Points (DPs). If erroneous extractions are the "symptoms" of semantic drift, then DPs are the "causes" of semantic drift. In this paper, we propose a method to minimize semantic drift by identifying the DPs and removing the effect introduced by the DPs. We use isA (concept-instance) extraction as an example to demonstrate the effectiveness of our approach in cleaning information ex- traction errors caused by semantic drift. We perform experiments on a isA relation iterative extraction, where 90.5 million of isA pairs are automatically extracted from 1.6 billion web documents with a low precision. The experimental results show our DP cleaning method enables us to clean more than 90% incorrect instances with 95% precision, which outperforms the previous approaches we compare with. As a result, our method greatly improves the prevision of this large isA data set from less than 50% to over 90%.

Wednesday, March 26			
11:00-12:30	EDBT Research Session 6		***Text and Sequence Mining***
Room:	Conference room 2	Chair:	*Maria Luisa Sapino*

COLARM: Cost-based Optimization for Localized Association Rule Mining

Abhishek Mukherji, Elke Rundensteiner and Matthew Ward

Association rule mining typically focuses on discovering global rules valid across the entire dataset. Yet local rules valid for subsets of the dataset, while significantly different from global rules, are often also of tremendous importance to analysts. In this work, we tackle this overlooked problem of online mining of localized association rules. We provide support for analysts to interactively mine rules that are hidden in a global context yet are locally significant.

To tackle this problem we design a compact multidimensional itemset-based data partitioning (MIP-index). MIP-index offers efficient mining performance by utilizing precomputed results, while still allowing the user the flexibility of selecting any data subset of interest at run-time. We design a suite of alternative execution strategies for processing such localized mining requests. Optimization principles such as selection push-up, supported R-tree filter and differential treatment of contained and partially over- lapped MIPs are proposed. We analytically and experimentally demonstrate that different execution strategies are effective for different query scenarios. Given a localized mining query, our CO- LARM query optimizer takes a cost-based approach to identify the best strategy for execution. Through extensive experiments using benchmark data sets we demonstrate that the COLARM optimizer is highly accurate in online plan selection and discovering localized rules (otherwise hidden in the global context) in a diversity of localized mining requests.

Fast Mining of Interesting Phrases from Subsets of Text Corpora

Deepak P, Atreyee Dey and Debapriyo Majumdar

We address the problem of mining interesting phrases from subsets of a text corpus where the subset is specified using a set of features such as keywords that form a query. Previous algorithms for the problem have proposed solutions that involve sifting through a phrase dictionary based index or a document-based index where the solution is linear in either the phrase dictionary size or the size of the document subset. We propose the usage of an independence assumption between query keywords given the top correlated phrases, wherein the pre-processing could be reduced to discovering phrases from among the top phrases per each feature in the query. We then outline an indexing mechanism where per-keyword phrase lists are stored either in disk or memory, so that popular aggregation algorithms such as No Random Access and Sort-merge Join may be adapted to do the scoring at real-time to identify the top interesting phrases. Though such an approach is expected to be approximate, we empirically illustrate that very high accuracies (of over 90%) are achieved against the results of exact algorithms. Due to the simplified list-aggregation, we are also able to provide response times that are

orders of magnitude better than state-of-the-art algorithms. Interestingly, our disk-based approach outperforms the in-memory baselines by up to hundred times and sometimes more, confirming the superiority of the proposed method.

Sequence Pattern Matching over Time-Series Data with Temporal Uncertainty

Yongluan Zhou, Chunyang Ma, Qingsong Guo, Lidan Shou and Gang Chen

In this paper, we consider complex pattern matching over event data generated from error-prone sources such as low-cost wireless motes, RFID. Such data are often imprecise in both their values and their timestamps. While there are existing works addressing the problem of spatial uncertainty (i.e. the uncertainty of the data values), relatively little attention has been paid to the problem of temporal uncertainty (i.e. the uncertainty of the event timestamps). As a step to fill this gap, we formulate the problem of matching complex sequence patterns over time-series data with temporal un- certainty and propose a new indexing structure to organize the in- formation of the uncertain sequences and a set of efficient pattern query processing algorithms. We conduct an extensive experimental study on both synthetic and real datasets. The results indicate that the query processing algorithms based on our index structure can dramatically improve the query performance.

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 AAA 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 26			
14:00-15:30	EDBT Vision Papers		***EDBT Vision Track***
Room:	Olympia	Chair:	*Maurice Van Keulen*

Spatial Data Management Challenges in the Simulation Sciences

Thomas Heinis, Farhan Tauheed and Anastasia Ailamaki

Scientists in many disciplines have progressively been using simulations to better understand the natural systems they study. Faster hardware, as well as increasingly precise instruments, allow the construction and simulation of progressively advanced models of various systems.

Governed by algorithms and equations, the spatial models at the core of simulations are changed and updated at every simulation step through spatial queries, implementing massive updates. Therefore, the efficient execution of these numerous spatial queries is essential.

Two reasons render current spatial indexes inadequate for simulation applications. First, to ensure quick access to data, most of the spatial models in simulations are stored in memory. Most spatial access methods, however, have been optimized for use on disk and are not efficient in memory. Second, in every time step of a simulation, almost all spatial elements change their position, challenging update mechanisms for spatial indexes.

In this paper we discuss how these challenges create opportunities for exciting data management research.

What Can Programming Languages Say About Data Exchange?

Michael Johnson, Jorge Pérez and James Terwilliger

Data Exchange, defined generally, is the process of taking data structured under one schema and transforming it into data structured under another independent schema. This process is present in enough scenarios both theoretical and practical that it has been addressed in many different ways. Most prominent amongst the solutions to the problem is that proposed by database literature, in which one constructs schema mappings, using (a subset of) first-order predicate calculus, to establish the high-level relationship among the database schemas participating in the exchange. From a schema mapping an executable process is derived to per- form the exchange. This line of research has made significant progress and come to impressive findings, but has some theoretical and practical shortcomings as well. For instance, there are theoretical limitations as to how to compose or invert such mappings in a complete and unique way, which is a barrier to making such mappings bidirectional. It is possible to address some of these shortcomings by looking to solutions from a different discipline—a construct from the programming language literature called a lens—that addresses similar problems from a different perspective. By combining solutions from these two disciplines, one ends up with a new direction of research as well as a result that might be greater than the sum of its parts.

Toward Hardware-Sensitive Database Operations

David Broneske, Sebastian Breß, Max Heimel and Gunter Saake

Satisfying the performance needs of tomorrow typically implies using modern processor capabilities (such as single instruction, multiple data) and co-processors (such as graphics processing units) to accelerate database operations. Algorithms are typically hand-tuned to the underlying (co-)processors. This solution is error-prone, introduces high implementation and maintenance cost and one implementations is not portable to other (co-)processors. To this end, we argue for a combination of database research with modern software-engineering approaches. We emphasize our vision of generating optimized database algorithms tailored to used (co-)processors from a common code base. With this, we maximize performance while minimizing implementation and maintenance effort of hardware-tailored database operations.

Wednesday, March 26			
14:00-15:30	EDBT Research Session 7		***Distributed Query Processing***
Room:	Conference room 2	Chair:	*Norman Paton*

Distributed Spatial Keyword Querying on Road Networks

Siqiang Luo, Yifeng Luo, Shuigeng Zhou, Gao Cong and Jihong Guan

Spatial-keyword queries on road networks are receiving in- creasing attention with the prominence of location-based services. There is a growing need to handle queries on road networks in distributed environments because a large net- work is typically distributed over multiple machines and it will improve query throughput. However, all the existing work on spatial keyword queries is based on a centralized setting. In this paper, we develop a distributed solution to answering spatial keyword queries on road networks. Example queries include “find locations near a supermarket and a hospital,” and “find Chinese restaurants within 500 meters from my current location.” We define an operation for answering such queries and reduce the problem of answering a query into computing a function of such operations. We pro- pose a new distributed index that enables each machine to independently evaluate the operation on its network fragment in a distributed setting. We theoretically prove the space optimality of the proposed index technique. We con- duct experiments with a distributed setting. Experimental results demonstrate the promising performance of our method.

Cost-Based Median Query Processing in Wireless Sensor Networks

Johannes Niedermayer, Mario Nascimento, Matthias Renz, Peer Kröger and Hans-Peter

A major concern when processing queries within a wireless sensor network is to minimize the energy consumption of the network nodes, thus extending the networks lifetime. One way to achieve this is by minimizing the amount of communication required to answer queries. In this paper we investigate exact continuous quantile queries, focusing on the particular case of the median query. Many recently proposed algorithms determine a quantile by performing a series of refining histogram queries. For that class of queries, we recently proposed a cost-model to estimate the optimal number of histogram buckets within an algorithm for minimizing the energy consumption of a query. In this paper, we extend that algorithm for continuous queries. Furthermore we also offer a new refinement-based algorithm that employs a heuristic to minimize the number of message transmissions. Our experiments, using synthetic and real datasets, show that despite its theoretical runtime complexity our heuristic solution is able to perform significantly better than histogram-based approaches.

RIPPLE: A Scalable Framework for Distributed Processing of Rank Queries

George Tsatsanifos, Dimitris Sacharidis and Timos Sellis

We introduce a generic framework, termed RIPPLE, for processing rank queries in decentralized systems. Rank queries are particularly challenging, since the search area (i.e., which tuples qualify) can- not be determined by any peer individually. While our proposed framework is generic enough to apply to all decentralized structured systems, we show that when coupled with a particular distributed hash table (DHT) topology, it offers guaranteed worst-case performance. Specifically, rank query processing in our framework exhibits tunable polylogarithmic latency, in terms of the network size. Additionally we provide a means to trade-off latency for communication and processing cost. As a proof of concept, we apply RIPPLE for top-k query processing. Then, we consider skyline queries, and demonstrate that our framework results in a method that has better latency and lower overall communication cost than existing approaches over DHTs. Finally, we provide a RIPPLE- based approach for constructing a k-diversified set, which, to the best of our knowledge, is the first distributed solution for this problem. Extensive experiments with real and synthetic datasets validate the effectiveness of our framework.

Wednesday, March 26			
16:00-18:00	EDBT Research Session 8		***Indexing and Cost Statistics***
Room:	Olympia	Chair:	*Stefan Manegold*

HCS: Hierarchical Cut Selection for Efficiently Processing Queries on Data Columns using Hierarchical Bitmap Indices

Parth Nagarkar and K. Selçuk Candan

When data are large and query processing workloads consist of data selection and aggregation operations (as in online analytical processing), column-oriented data stores are generally the preferred choice of data organization, because they enable effective data compression, leading to significantly reduced IO. Most column- store architectures leverage bitmap indices, which themselves can be compressed, for answering queries over data columns. Column- domains (e.g., geographical data, categorical data, biological taxonomies, organizational data) are hierarchical in nature, and it may be more advantageous to create hierarchical bitmap indices, that can help answer queries over different sub-ranges of the domain. However, given a query workload, it is critical to choose the appropriate subset of bitmap indices from the given hierarchy. Thus, in this paper, we introduce the cut-selection problem, which aims to help identify a subset (cut) of the nodes of the domain hierarchy, with the appropriate bitmap indices. We discuss inclusive, exclusive, and hybrid strategies for cut-selection and show that the hybrid strategy can be efficiently computed and returns optimal (in terms of IO) results in cases where there are no memory constraints. We also show that when there is a memory availability constraint, the cut-selection problem becomes difficult and, thus, present efficient cut-selection strategies that return close to optimal results, especially in situations where the memory limitations are very strict (i.e., the data and the hierarchy are much larger than the available memory). Experiment results confirm the efficiency and effective- ness of the proposed cut-selection algorithms.

Adaptive String Dictionary Compression in In-Memory Column-Store Database Systems

Ingo Müller, Cornelius Ratsch and Franz Färber

Domain encoding is a common technique to compress the columns of a column store and to accelerate many types of queries at the same time. It is based on the assumption that most columns contain a relatively small set of distinct values, in particular string columns. In this paper, we argue that domain encoding is not the end of the story. In real world systems, we observe that a substantial amount of the columns are of string types. Moreover, most of the memory space is consumed by only a small fraction of these columns.

To address this issue, we make three main contributions: First we survey several approaches and variants for dictionary compression, i. e., data structures that store the dictionary of domain encoding in a compressed way. As expected, there is a trade-off between size of the data structure and its access performance. This observation can be used to compress rarely accessed data more than frequently accessed data. Furthermore the question which approach has the best compression ratio for a certain column heavily depends on specific characteristics of its content. Consequently, as a second contribution, we present non-trivial sampling schemes for all our dictionary formats, enabling us to estimate their size for a given column. This way it is possible to identify compression schemes specialized for the content of a specific column.

Third, we draft how to fully automate the decision of the dictionary format. We sketch a compression manager that selects the most appropriate dictionary format based on column access and up- date patterns, characteristics of the underlying data, and costs for set-up and access of the different data structures. We evaluate an off-line prototype of a compression manager using a variation of the TPC-H benchmark [15]. The compression manager can con- figure the database system to be anywhere in a large range of the space / time trade-off with a fine granularity, providing significantly better trade-offs than any fixed dictionary format.

Online Topic-aware Influence Maximization Queries

Cigdem Aslay, Nicola Barbieri, Francesco Bonchi and Ricardo Baeza-Yates

Influence maximization is the key algorithmic problem be- hind viral marketing: it requires to identify a set of influential users in a social network, who, when convinced to adopt a product, shall influence other users in the network, leading to a large number of adoptions. Although real world users evidently have different degrees of interest and authoritativeness on different topics, the bulk of the literature on influence maximization is topic-blind, in the sense that it treats all items as they were the same.

In this paper we study Topic-aware Influence Maximization (TIM) queries: given a directed social graph, where the arcs are associated with a topic-dependent user-to-user social influence strength, and given a budget k, the problem requires to find a set of k users (named seed set) that we shall target in a viral marketing campaign for a given new item (described as a distribution over topics) in order to maximize its adoption. Our goal is to answer such queries in milliseconds, thus enabling online social influence analytics, what-if simulation, and marketing decision making.

The main challenge here is the enormous number of potential queries: any possible distribution over the topic space (i.e., any possible item) induces a different probabilistic graph, and thus a different instance of the standard influence maximization problem, for which efficiency and scalability are still unsolved problems.

Given these computational challenges, we propose to build an index over pre-computed solutions for a limited number of possible queries. Our proposal, INFLEX, employs a tree- based index for similarity search with Bregman divergences, to eefficiently retrieve a good-enough set of neighbor points for the query item. Then it performs rank aggregation on their seed sets to produce the final answer to the query. Experimental results on real data show that INFLEX can provide in few milliseconds a solution very similar (Kendall-τ distance < 0.1) to the one produced by the best known offline computation (which usually takes several days).

Determining Essential Statistics for Cost Based Optimization of an ETL Workflow

Ramanujam S Halasipuram, Prasad M Deshpande and Sriram Padmanabhan

Many of the ETL products in the market today provide tools for design of ETL workflows, with very little or no support for optimization of such workflows. Optimization of ETL workflows pose several new challenges compared to traditional query optimization in database systems. There have been many attempts both in the industry and the research community to support cost-based optimization techniques for ETL Workflows, but with limited success. Non-availability of source statistics in ETL is one of the major challenges that precludes the use of a cost based optimization strategy. However, the basic philosophy of ETL workflows of design once and execute repeatedly allows interesting possibilities for determining the statistics of the input. In this paper, we propose a frame- work to determine various sets of statistics to collect for a given workflow, using which the optimizer can estimate the cost of any alternative plan for the workflow. The initial few runs of the workflow are used to collect the statistics and future runs are optimized based on the learned statistics. Since there can be several alternative sets of statistics that are sufficient, we propose an optimization framework to choose a set of statistics that can be measured with the least overhead. We experimentally demonstrate the effective- ness and efficiency of the proposed algorithms.

Wednesday, March 26			
16:00-18:00	EDBT Research Session 9		***Matrix Factorization, Clustering and Probabilistic Data***
Room:	Conference room 2	Chair:	*Amélie Marian*

CLUDE: An Efficient Algorithm for LU Decomposition Over a Sequence of Evolving Graphs

Chenghui Ren, Luyi Mo, Ben Kao, Reynold Cheng and David W. Cheung

In many applications, entities and their relationships are represented by graphs. Examples include the WWW (web pages and hyperlinks) and bibliographic networks (authors and co-authorship). A graph can be conveniently modeled by a matrix from which various quantitative measures are derived. Some example measures include PageRank and SALSA (which measure nodes' importance), and Personalized PageRank and Random Walk with Restart (which measure proximities between nodes). To compute these measures, linear systems of the form Ax = b, where A is a matrix that captures a graph's structure, need to be solved. To facilitate solving the linear system, the matrix A is often decomposed into two triangular matrices (L and U). In a dynamic world, the graph that models it changes with time and thus is the matrix A that represents the graph. We consider a sequence of evolving graphs and its associated sequence of evolving matrices. We study how LU-decomposition should be done over the sequence so that (1) the decomposition is efficient and (2) the resulting LU matrices best preserve the sparsity of the matrices A's (i.e., the number of extra non-zero entries introduced in L and U are minimized.) We propose a cluster-based algorithm CLUDE for solving the problem. Through an experimental study, we show that CLUDE is about an order of magnitude faster than the traditional incremental update algorithm. The number of extra non-zero entries introduced by CLUDE is also about an order of magnitude fewer than that of the traditional algorithm. CLUDE is thus an efficient algorithm for LU decomposition that produces high-quality LU matrices over an evolving matrix sequence.

Model Selection for Semi-Supervised Clustering

Mojgan Pourrajabi, Davoud Moulavi, Ricardo Campello, Arthur Zimek, Jörg Sander and Randy Goebel

Although there is a large and growing literature that tackles the semi-supervised clustering problem (i.e., using some labeled objects or cluster-guiding constraints like "must-link" or "cannot-link"), the evaluation of semi-supervised clustering approaches has rarely been discussed. The application of cross-validation techniques, for example, is far from straight- forward in the semi-supervised setting, yet the problems associated with evaluation have yet to be addressed. Here we summarize these problems and provide a solution.

Furthermore, in order to demonstrate practical applicability of semi-supervised clustering methods, we provide a method for model selection in semi-supervised clustering based on this sound evaluation procedure. Our method allows the user to select, based on the available information (labels or constraints), the most appropriate clustering model (e.g., number of clusters, density-parameters) for a given problem.

Spatial Partitioning of Large Urban Road Networks

Tarique Anwar, Chengfei Liu, Hai L Vu and Christopher Leckie

The rapid global migration of people towards urban areas is multiplying the traffic volume on urban road networks. As a result these networks are rapidly growing in size, in which different sub-networks exhibit distinctive traffic flow patterns.

In this paper, we propose a scalable framework for traffic congestion-based spatial partitioning of large urban road networks. It aims to identify different sub-networks or partitions that exhibit homogeneous traffic congestion patterns internally, but heterogenous to others externally. To this end, we develop a two-stage procedure within our framework that first transforms the large road graph into a well-structured and condensed supergraph via clustering and link aggregation based on traffic density and adjacency connectivity, respectively. We then devise a spectral theory based novel graph cut (referred as α-Cut) to partition the supergraph and compare its performance with that of an existing method for partitioning urban networks. Our results show that the proposed method outperforms the normalized cut based existing method in all the performance evaluation metrics for small road networks and provides good results for much larger networks where other methods may face serious problems of time and space complexities.

ENFrame: A Platform for Processing Probabilistic Data

Sebastiaan van Schaik, Dan Olteanu and Robert Fink

This paper introduces ENFrame, a unified data processing platform for querying and mining probabilistic data. Us- ing ENFrame, users can write programs in a fragment of Python with constructs such as bounded-range loops, list comprehension, aggregate operations on lists, and calls to external database engines. The program is then interpreted probabilistically by ENFrame.

The realisation of ENFrame required novel contributions along several directions. We propose an event language that is expressive enough to succinctly encode arbitrary correlations, trace the computation of user programs, and allow for computation of discrete probability distributions of program variables. We exemplify ENFrame on three clustering algorithms: k-means, k-medoids, and Markov clustering. We introduce sequential and distributed algorithms for computing the probability of interconnected events exactly or approximately with error guarantees.

Experiments with k-medoids clustering of sensor readings from energy networks show orders-of-magnitude improvements of exact clustering using ENFrame over naïve clustering in each possible world, of approximate over exact, and of distributed over sequential algorithms.

Thursday, March 27			
11:00-12:30	EDBT Research Session 10		***Keyword Search and Diversity***
Room:	Attica	Chair:	*Senjuti Basu Roy*

Diversified Spatial Keyword Search On Road Networks

Chengyuan Zhang, Ying Zhang, Wenjie Zhang, Xuemin Lin, Muhammad Cheema and Xiaoyang Wang

With the increasing pervasiveness of the geo-positioning technologies, there is an enormous amount of spatio-textual objects available in many applications such as location based services and social networks. Consequently, various types of spatial keyword searches which explore both locations and textual descriptions of the objects have been intensively studied by the research communities and commercial organizations. In many important applications (e.g., location based services), the closeness of two spatial objects is measured by the road network distance. Moreover, the result diversification is becoming a common practice to enhance the quality of the search results. Motived by the above facts, in this paper we study the problem of diversified spatial keyword search on road networks which considers both the relevance and the spatial diversity of the results. An efficient signature-based inverted indexing technique is pro- posed to facilitate the spatial keyword query processing on road networks. Then we develop an efficient diversified spatial keyword search algorithm by taking advantage of spatial keyword pruning and diversity pruning techniques. Comprehensive experiments on real and synthetic data clearly demonstrate the efficiency of our methods.

Cleaning trajectory data of RFID-monitored objects through conditioning under integrity constraints

Bettina Fazzinga, Sergio Flesca, Filippo Furfaro and Francesco Parisi

A probabilistic framework is introduced for reducing the inherent uncertainty of trajectory data collected for RFID-monitored objects. The framework represents the position of an object at each instant as a random variable over the set of possible locations. The probability density function of this random variable is initialized according to an a-priori probability distribution, and then revised by conditioning it w.r.t. the event that integrity constraints are satisfied. In particular, integrity constraints implied by the structure of the map of locations and the motility characteristics (such as the maximum speed) of the monitored objects are exploited (namely, direct unreachability, latency and minimum traveling time constraints). The efficiency and effectiveness of the proposed approach are assessed experimentally on synthetic data.

Multi-Criteria Optimal Location Query with Overlapping Voronoi Diagrams

Ji Zhang, Wei-Shinn Ku, Min-Te Sun, Xiao Qin and Hua Lu

This paper presents a novel optimal location selection problem, which can be applied to a wide range of applications. After providing a formal definition of the novel query type, we explore an intuitive approach that sequentially scans all possible object combinations in the search space. Then, we propose an Overlapping Voronoi Diagram (OVD) model that defines OVDs and Minimum OVDs, and construct an algebraic structure under an OVD overlap operation. Based on the OVD model, we design an advanced approach to answer the query. Due to the high complexity of Voronoi diagram overlap computation, we improve the overlap operation by replacing the real boundaries of Voronoi diagrams with their Minimum Bounding Rectangles (MBR). We also propose a cost-bound iterative approach that efficiently processes a large number of Fermat-Weber problems. Our experimental results show that the proposed algorithms can evaluate the novel query type effectively and efficiently.

Thursday, March 27			
11:00-12:30	EDBT Research Session 11		***Ranking***
Room:	Templar’s	Chair:	*Bernd Amann*

Efficient Concept-based Document Ranking

Anastasios Arvanitis, Matthew Wiley and Vagelis Hristidis

Recently, there is increased interest in searching and computing the similarity between Electronic Medical Records (EMRs). A unique characteristic of EMRs is that they consist of ontological concepts derived from biomedical ontologies such as UMLS or SNOMED- CT. Medical researchers have found that it is effective to search and find similar EMRs using their concepts, and have proposed sophisticated similarity measures. However, they have not addressed the performance and scalability challenges to support searching and computing similar EMRs using ontological concepts. In this paper, we formally define these important problems and show that they pose unique algorithmic challenges due to the nature of the search and similarity semantics and the multi-level relationships between the concepts. In particular, the similarity between two EMRs is a function of the minimum semantic distance from each concept of one document to a concept of the other and vice versa. We present an efficient algorithm to compute the similarity between two EMRs. Then, we propose an early-termination algorithm to search for the top-k most relevant EMRs to a set of concepts, and to find the top-k most similar EMRs to a given EMR. We experimentally evaluate the performance and scalability of our methods on a large real EMR data set.

Metric-Based Top-k Dominating Queries

Eleftherios Tiakas, George Valkanas, Apostolos N. Papadopoulos and Yannis Manolopoulos

Top-k dominating queries combine the natural idea of selecting the k best items with a comprehensive “goodness” criterion based on dominance. A point p1 dominates p2 if p1 is as good as p2 in all attributes and is strictly better in at least one. Existing works address the problem in settings where data objects are multidimensional points. However, there are domains where we only have access to the distance between two objects. In cases like these, attributes reflect distances from a set of input objects and are dynamically generated as the input objects change. Consequently, prior works from the literature can not be applied, despite the fact that the dominance relation is still meaningful and valid. For this reason, in this work, we present the first study for processing top-k dominating queries over distance-based dynamic attribute vectors, defined over a metric space. We propose four progressive algorithms that utilize the proper- ties of the underlying metric space to efficiently solve the problem, and present an extensive, comparative evaluation on both synthetic and real world data sets.

A Unified Framework for Efficiently Processing Ranking Related Queries

Muhammad Cheema, Zhitao Shen, Xuemin Lin and Wenjie Zhang

The computation of k-lower envelope is a classical problem and has been very well studied for main memory non-indexed data. In this paper, we study the problem from the database perspective and present the first algorithm which utilizes the presence of the index and achieves access optimality, i.e., it accesses a node of the index only if the correctness of the results cannot be guaranteed with- out accessing this node. We also demonstrate the applications of k-lower envelope in ranking systems. Let an object be called valuable if it is one of the top-k objects according to at least one linear scoring function. In this paper, we answer the following important questions that may be asked by different users: 1) I am not sure what scoring function I should use, therefore, return me the set of valuable objects so that I can select an object I like the most; 2) How can I modify the attributes (e.g., price) of my product such that it becomes a valuable object; 3) What are the preference functions for which a given object is among the top-k objects. These three questions are formalized and called k-snippet, k-depth con- tour and reverse top-k query, respectively. We propose a unified framework to solve these queries by utilizing k-lower envelope as a common foundation. Our main algorithm is access optimal for k-snippet and k-lower envelope computation. We also demonstrate its access optimality for the k-depth contour problem when k is smaller than the minimum number of objects in any leaf node of the index structure. Our algorithms outperform state-of-the-art algorithms by more than an order of magnitude in terms of both CPU and I/O cost.

Thursday, March 27			
14:00-15:30	EDBT Research Session 12		***Joins***
Room:	Attica	Chair:	*Torsten Grust*

Exploiting the query structure for efficient join ordering in SPARQL queries

Andrey Gubichev and Thomas Neumann

The join ordering problem is a fundamental challenge that has to be solved by any query optimizer. Since the high-performance RDF systems are often implemented as triple stores (i.e., they represent RDF data as a single table with three attributes, at least conceptually), the query optimization strategies employed by such systems are often adopted from relational query optimization. In this paper we show that the techniques borrowed from traditional SQL query optimization (such as Dynamic Programming algorithm or greedy heuristics) are not immediately capable of handling large SPARQL queries. We introduce a new join ordering algorithm that performs a SPARQL-tailored query simplification. Furthermore, we present a novel RDF statistical synopsis that accurately estimates cardinalities in large SPARQL queries. Our experiments show that this algorithm is highly superior to the state-of-the-art SPARQL optimization approaches, including the RDF-3X's original Dynamic Programming strategy.

Interactive Inference of Join Queries

Angela Bonifati, Radu Ciucanu and Slawek Staworko

We investigate the problem of inferring join queries from user interactions. The user is presented with a set of candidate tuples and is asked to label them as positive or negative depending on whether or not she would like the tuples as part of the join result. The goal is to quickly infer an arbitrary n-ary join predicate across two relations by keeping the number of user interactions as minimal as possible. We assume no prior knowledge of the integrity constraints be- tween the involved relations. This kind of scenario occurs in several application settings, such as data integration, reverse engineering of database queries, and constraint inference. In such scenarios, the database instances may be too big to be skimmed. We explore the search space by using a set of strategies that let us prune what we call "uninformative" tuples, and directly present to the user the informative ones i.e., those that allow to quickly find the goal query that the user has in mind. In this paper, we focus on the inference of joins with equality predicates and we show that for such joins deciding whether a tuple is uninformative can be done in polynomial time. Next, we propose several strategies for presenting tuples to the user in a given order that lets minimize the number of interactions. We show the efficiency and scalability of our approach through an experimental study on both benchmark and synthetic datasets. Finally, we prove that adding projection to our queries makes the problem intractable.

Processing Interval Joins On Map-Reduce

Bhupesh Chawda and Himanshu Gupta

In this paper we investigate the problem of processing multi- way interval joins on map-reduce platform. We look at join queries formed by interval predicates as defined by Allen's interval algebra. These predicates can be classified in two groups: colocation based predicates and sequence based predicates. A colocation predicate requires two intervals to share at least one common point while a sequence predicate requires two intervals to be disjoint. An interval join query can therefore be thought of as belonging to one of the three classes: (a) queries containing only colocation based predicates, (b) queries containing only sequence based predicates and (c) queries containing both classes of predicates. We address these three classes of join queries, discuss the challenges and present novel approaches for processing these queries on map-reduce platform. We also discuss why the current approaches developed for handling join queries on real-valued data can not be directly used to handle interval joins. We finally extend the approaches developed to handle join queries containing multiple interval attributes as well as join queries containing both interval as well as non-interval attributes. Through experimental evaluations both on synthetic and real life datasets, we demonstrate that the proposed approaches comfortably outperform naive approaches.

Wednesday, March 27			
14:00-15:30	EDBT Research Session 13		***Privacy-Aware Data Processing***
Room:	Templar's	Chair:	*Alfredo Cuzzocrea*

Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions

Haoran Li, Li Xiong and Xiaoqian Jiang

Differential privacy has recently emerged in private statistical data release as one of the strongest privacy guarantees. Most of the existing techniques that generate differentially private histograms or synthetic data only work well for single dimensional or low-dimensional histograms. They become problematic for high dimensional and large domain data due to increased perturbation error and computation complexity. In this paper, we propose DPCopula, a differentially private data synthesization technique using Copula functions for multi-dimensional data. The core of our method is to compute a differentially private copula function from which we can sample synthetic data. Copula functions are used to describe the dependence between multivariate random vectors and allow us to build the multivariate joint distribution using one-dimensional marginal distributions. We present two methods for estimating the parameters of the copula functions with differential privacy: maximum likelihood estimation and Kendall's τ estimation. We present formal proofs for the privacy guarantee as well as the convergence property of our methods. Extensive experiments using both real datasets and synthetic datasets demonstrate that DPCopula generates highly accurate synthetic multi- dimensional data with significantly better utility than state- of-the-art techniques.

Privacy-Preserving Query Execution using a Decentralized Architecture and Tamper Resistant Hardware

Quoc-Cuong To, Benjamin Nguyen and Philippe Pucheral

Current applications, from complex sensor systems (e.g. quantified self) to online e-markets acquire vast quantities of personal information which usually ends-up on central servers. Decentralized architectures, devised to help individuals keep full control of their data, hinder global treatments and queries, impeding the development of services of great interest. This paper promotes the idea of pushing the security to the edges of applications, through the use of secure hardware devices controlling the data at the place of their acquisition. To solve this problem, we propose secure distributed querying protocols based on the use of a tangible physical element of trust, reestablishing the capacity to perform global computations without revealing any sensitive information to central servers. There are two main problems when trying to support SQL in this context: perform joins and perform aggregations. In this paper, we study the subset of SQL queries without joins and show how to secure their execution in the presence of honest-but-curious attackers. Cost models and experiments demonstrate that this approach can scale to nationwide infrastructures.

Privacy Preserving Similarity Evaluation of Time Series Data

Haohan Zhu, Xianrui Meng and George Kollios

Privacy preserving issues of time series databases in financial, medical and transportation applications have become more and more important recently. A key problem in time series databases is to compute the similarity between two different time series. Despite some recent work on time series security and privacy, there is very limited progress on securely computing the similarity between two time series. In this paper, we consider exactly this problem in a two- party setting (client and server). In particular, we want to compute the similarity between two time series, one from the client and the other from the server, without revealing the actual time series to the other party. Only the value of the similarity should be revealed to both parties at the end. At the same time, we want to do the computation as efficiently as possible. Therefore, we propose practical protocols for computing the similarity (or distance) for time series using two popular and well known functions: Dynamic Time Warping and Discrete Fŕechet Distance. Since both of these functions require dynamic programming to be computed, our protocols not only encrypt the original time series data, but also try to hide intermediate results, including the matrix of the dynamic programming algorithm and the path of the optimal solution. The protocols combine partial ho- momorphic encryption and random offsets to protect intermediate information and at the same time provide efficient computation. Unlike previous approaches that are mostly theoretical, our protocols are scalable and easy to implement. We also provide an experimental evaluation where we assess the scalability and practicality of our schemes using both synthetic and real datasets.

Wednesday, March 27			
16:00-18:00	EDBT Research Session 14		***Graph Queries and Analytics***
Room:	Olympia	Chair:	*Venkatesh Raghavan*

Reachability Queries in Very Large Graphs: A Fast Refined Online Search Approach

Renê R. Veloso, Loïc Cerf, Wagner Meira Junior and Mohammed J. Zaki

A key problem in many graph-based applications is the need to know, given a directed graph G and two vertices u, v ∈ G, whether there is a path between u and v, i. e., if u reaches v. This problem is particularly challenging in the case of very large real-world graphs. A common approach is the pre- processing of the graphs, in order to produce an efficient index structure, which allows fast access to the reachability information of the vertices. However, the majority of existing methods can not handle very large graphs. We propose, in this paper, a novel indexing method called FELINE (Fast rEfined onLINE search), which is inspired by Dominance Graph Drawing. FELINE creates an index from the graph representation in a two-dimensional plane, which provides reachability information in constant time for a significant portion of queries. Experiments demonstrate the efficiency of FELINE compared to state-of-the-art approaches.

Graph Analytics on Massive Collections of Small Graphs

Dritan Bleco and Yannis Kotidis

Emerging applications face the need to store and query data that are naturally depicted as graphs. Building a Business Intelligence (BI) solution for graph data is a formidable task. Relational databases are frequently criticized for being unsuitable for managing graph data. Graph databases are gaining popularity but they have not yet reached the same maturity level with relational systems. In this pa- per we identify a large spectrum of applications that generate graph data with specific characteristics that make them candidate for being stored in a relational system. We describe a novel framework where data and queries are both treated as abstract graph structures that can be decomposed into simpler structural elements. We complement this abstract framework with a description of a system that utilizes three different means of expediting user queries: (i) a flat description of the graph records using a column-oriented storage model, (ii) use of bitmap columns for enabling fast access to parts of these graph records and (iii) a novel framework for selecting and materializing graph views that significantly expedite retrieval of records in response to a graph query. To the best of our knowledge we are the first to report results using datasets consisting of hundreds of millions of graphs with billions nodes, edges and mea- sure values using a single database server running of a commodity node. Our results demonstrate that our platform is orders of magnitude faster than alternative systems that natively handle graph data and a straightforward relational implementation. Moreover, our materialization techniques (that account for about 10% of extra disk space) are able to reduce the query execution times further, by up to 94% compared to an evaluation plan that is oblivious to the existing materialized graphs views in the database.

Fast Reliability Search in Uncertain Graphs

Arijit Khan, Francesco Bonchi, Aris Gionis and Francesco Gullo

Uncertain, or probabilistic, graphs have been increasingly used to represent noisy linked data in many emerging application scenarios, and have recently attracted the attention of the database re- search community. A fundamental problem on uncertain graphs is reliability, which deals with the probability of nodes being reachable one from another. Existing literature has exclusively focused on reliability detection, which asks to compute the probability that two given nodes are connected.

In this paper we study reliability search on uncertain graphs, which we define as the problem of computing all nodes reach- able from a set of query nodes with probability no less than a given threshold. Existing reliability-detection approaches are not well-suited to efficiently handle the reliability-search problem. We propose RQ-tree, a novel index which is based on a hierarchical clustering of the nodes in the graph, and further optimized using a balanced-minimum-cut criterion. Based on RQ-tree, we define a fast filtering-and-verification online query-evaluation strategy that relies on a maximum-flow-based candidate-generation phase, fol- lowed by a verification phase consisting of either a lower-bounding method or a sampling technique. The first verification method re- turns no incorrect nodes, thus guaranteeing perfect precision, completely avoids sampling, and is more efficient. The second verification method ensures instead better recall.

Extensive experiments on real-world uncertain graphs show that our methods are very efficient—over state-of-the-art reliability- detection methods, we obtain a speed-up up to five orders of magnitude; as well as accurate—our techniques achieve precision > 0.95 and recall usually higher than 0.75.

Distance oracles in edge-labeled graphs

Francesco Bonchi, Aristides Gionis, Francesco Gullo and Antti Ukkonen

A fundamental operation over edge-labeled graphs is the computation of shortest-path distances subject to a constraint on the set of permissible edge labels.

Applying exact algorithms for such an operation is not a viable option, especially for massive graphs, or in scenarios where the distance computation is used as a primitive for more complex computations.

In this paper we study the problem of efficient approximation of shortest-path queries with edge-label constraints, for which we de- vise two indexes based on the idea of landmarks: distances from all vertices of the graph to a selected subset of landmark vertices are pre-computed and then used at query time to efficiently approximate distance queries. The major challenge to face is that, in principle, an exponential number of constraint label sets needs to be stored for each vertex-landmark pair, which makes the index pre-computation and storage far from trivial. We tackle this challenge from two different perspectives, which lead to indexes with different characteristics: one index is faster and more accurate, but it requires more space than the other.

We extensively evaluate our techniques on real and synthetic datasets, showing that our indexes can efficiently and accurately estimate label-constrained distance queries.

Wednesday, March 27			
16:00-18:00	EDBT Research Session 15		***Privacy in Networks***
Room:	Attica	Chair:	*Irini Fundulaki*

Privacy Preserving Estimation of Social Influence

Tamir Tassa and Francesco Bonchi

Exploiting word-of-mouth effect to create viral cascades in social networks is a very appealing possibility from the marketing standpoint. However, in order to set up an effective viral marketing campaign, one has first to accurately estimate social influence. This is usually done by analyzing user activity data. As we point out in this paper, the data analysis and sharing that is needed to estimate social influence raises important privacy issues that may jeopardize the le- gal, ethical and societal acceptability of such practice, and in turn, the concrete applicability of viral marketing in the real world.

In this paper we devise secure multiparty protocols that allow a group of service providers and a social networking platform to jointly compute social influence, in a privacy preserving manner.

A Privacy-Preserving Framework for Personalized, Social Recommendations

Zach Jorgensen and Ting Yu

We consider the problem of producing item recommendations that are personalized based on a user's social network, while simultaneously preventing the disclosure of sensitive user-item preferences (e.g., product purchases, ad clicks, web browsing history, etc.). Our main contribution is a privacy- preserving framework for a class of social recommendation algorithms that provides strong, formal privacy guarantees under the model of differential privacy. Existing mechanisms for achieving differential privacy lead to an unacceptable loss of utility when applied to the social recommendation problem. To address this, the proposed framework incorporates a clustering procedure that groups users according to the natural community structure of the social network and significantly reduces the amount of noise required to satisfy differential privacy. Although this reduction in noise comes at the cost of some approximation error, we show that the benefits of the former significantly outweigh the latter. We explore the privacy-utility trade-off for several different instantiations of the proposed framework on two real-world data sets and show that useful social recommendations can be produced without sacrificing privacy. We also experimentally compare the proposed framework with several existing differential privacy mechanisms and show that the proposed framework significantly outperforms all of them in this set- ting.

L-opacity: Linkage-Aware Graph Anonymization

Sadegh Nobari, Panagiotis Karras, Hweehwa Pang and Stéphane Bressan

The wealth of information contained in online social networks has created a demand for the publication of such data as graphs. Yet, publication, even after identities have been removed, poses a privacy threat. Past research has suggested ways to publish graph data in a way that prevents the re-identification of nodes. However, even when identities are effectively hidden, an adversary may still be able to infer linkage between individuals with sufficiently high confidence. In this paper, we focus on the privacy threat arising from such link disclosure. We suggest L-opacity, a sufficiently strong privacy model that aims to control an adversary's confidence on short multi- edge linkages among nodes. We propose an algorithm with two variant heuristics, featuring a sophisticated look-ahead mechanism, which achieves the desired privacy guarantee after a few graph modifications. We empirically evaluate the performance of our algorithm, measuring the alteration inflicted on graphs and various utility metrics quantifying spectral and structural graph properties, while we also compare them to a recently proposed, albeit limited in generality of scope, alternative. Thereby, we demonstrate that our algorithms are more general, effective, and efficient than the competing technique, while our heuristic that preserves the number of edges in the graph constant fares better overall than one that reduces it.

Privacy Risk in Anonymized Heterogeneous Information Networks

Aston Zhang, Xing Xie, Kevin Chen-Chuan Chang, Carl A. Gunter, Jiawei Han and Xiaofeng Wang

Anonymized user datasets are often released for research or industry applications. As an example, t.qq.com released its anonymized users' profile, social interaction, and recommendation log data in KDD Cup 2012 to call for recommendation algorithms. Since the entities (users and so on) and edges (links among entities) are of multiple types, the released social network is a heterogeneous in- formation network. Prior work has shown how privacy can be com- promised in homogeneous information networks by the use of specific types of graph patterns. We show how the extra information derived from heterogeneity can be used to relax these assumptions. To characterize and demonstrate this added threat, we formally de- fine privacy risk in an anonymized heterogeneous information net- work to identify the vulnerability in the possible way such data are released, and further present a new de-anonymization attack that exploits the vulnerability. Our attack successfully de-anonymized most individuals involved in the data—for an anonymized 1,000- user t.qq.com network of density 0.01, the attack precision is over 90% with a 2.3-million-user auxiliary network.

EDBT Industry & Application Sessions

Wednesday, March 26			
11:00-12:30	EDBT Industry & Applications Session 1		***Applications***
Room:	Templar’s	Chair:	Dimitris Kotzinos

Talking to the Database in a Semantically Rich Way

Henrietta Dombrovskaya and Richard Lee

Conventional recommendations for Object Oriented application design include the concept of Object-Relational Mapping and suggest clear separation of business logic from interaction with the database. While these requirements seem natural to application developers, it prevents them from using the full power of the database engine, and thereby become the most essential source of application performance degradation. Acknowledging the widespread usage of the above concepts, our approach provides an algorithm for “splitting” logic between different layers of classes. We identify the parts of logic that are essential for data retrieval and thereby belong to the database, and the parts of logic that drive the computation or other data transformation and can reside in the application model. Although the splitting logic algorithm, as yet, is not implemented in any tool, we consider it an important part of the application design process. In our paper we provide examples of redesigned methods as well as before-and-after performance data from the production system.

SMILE: A Data Sharing Platform for Mobile Apps in the Cloud

Jagan Sankaranarayanan, Hakan Hacigumus, Haopeng Zhang and Mohamed Sarwat

We identify an opportunity to share data among mobile apps hosted in the cloud, thus helping users improve their mobile experience, while resulting in cost savings for the cloud provider. In this work, we propose a platform for sharing data among mobile apps hosted in the cloud. A “sharing” is specified by a triple consisting of: (a) a set of data sources to be shared, (b) a set of specified transformations on the shared data, and (c) a staleness (freshness) requirement on the shared data. The platform addresses the following two main challenges: What sharings to admit into the system under a set of specified constraints, how to implement a sharing at a low cost while maintaining the desired level of staleness. We show that reductions in costs are achievable by exploiting the commonalities between the different sharings in the platform. Experimental evaluation is performed with a cloud platform containing 25 sharings among mobile apps with realistic datasets containing user, social, location and checkin data. Our platform is able to maintain the sharings with very few violations, even under a very high update rate. Our results show that our method results in a cost savings of over 35% for the cloud provider, while enabling an improved mobile experience for users.

Diff-Index: Differentiated Index in Distributed Log-Structured Data Stores

Wei Tan, Sandeep Tata, Yuzhe Tang and Liang Fong

Log-Structured-Merge (LSM) Tree gains much attention recently because of its superior performance in write-intensive workloads. LSM Tree uses an append-only structure in memory to achieve low write latency; at memory capacity, in-memory data are flushed to other storage medium (e.g. disk). Consequently, read access is usually slower comparing to write. These specific features of LSM, including no in-place update and asymmetric read/write performance raise unique challenges in index maintenance for LSM. The structural difference between LSM and B-Tree also prevents mature B-Tree based approaches from being directly applied. To address the issues of index maintenance for LSM, we propose Diff-Index to support a spectrum of index maintenance schemes to suit different objectives in index consistency and performance. The schemes consist of sync-full, sync-insert, async-

simple and async-session. Experiments on our HBase implementation quantitatively demonstrate that Diff-Index offers various performance/consistency balance and satisfactory scalability while avoiding global coordination. Sync-insert and async-simple can reduce 60%-80% of the overall index update latency when compared to the baseline sync-full; async-simple can achieve superior index update performance with a mild inconsistency. Diff-Index exploit LSM features such as versioning and the flush-compact process, to achieve goals of concurrency control and failure recovery with low complexity and overhead.

Wednesday, March 26			
14:00-15:30	EDBT Industry & Applications Session 2		***Applications***
Room:	Templar’s	Chair:	Yannis Stavrakas

Heterogeneous Stream Processing and Crowdsourcing for Urban Traffic Management

Alexander Artikis, Matthias Weidlich, Francois Schnitzler, Ioannis Boutsis, Thomas Liebig, Nico Piatkowski, Christian Bockermann, Katharina Morik, Vana Kalogeraki, Jakub Marecek, Avigdor Gal, Shie Mannor, Dimitrios Gunopulos and Dermot Kinane

Intelligent traffic and transport management involves the use of large, uncertain data streams to identify and effectively manage issues of congestion and quality of service. In particular, urban traffic has been in the eye of the storm for many years now and gathers increasing interest as cities become bigger, crowded, and ``smart”. We present a system for heterogeneous stream processing and crowdsourcing supporting intelligent urban traffic management. Complex events related to traffic congestion (trends) are detected from heterogeneous sources involving fixed sensors mounted on intersections and mobile sensors mounted on public transport vehicles. To deal with the inherent data veracity, a crowdsourcing component handles and resolves sensor disagreement. Furthermore, to deal with data sparsity, a traffic modeling component offers information in areas with low sensor coverage. We demonstrate the proposed system with a real-world use-case from the city of Dublin, Ireland.

On Assigning Implicit Reputation Scores in an Online Labor Marketplace

Maria Daltayanni, Luca de Alfaro, Panagiotis Papadimitriou and Panayiotis Tsaparas

In online labor marketplaces, such as oDesk and Mechanical Turk, two parties are involved; employers and workers. Employers post jobs on an online platform and candidate workers apply for openings, based on their qualifications, skills and interests. Beyond skill matching and past employer ratings (often biased), true quality of applicants is crucial for selecting the best hire for each job. Computing reputation scores for workers based on their true performance, not only saves time spent on profile review and interviewing, but also yields objective hire recommendations. We build a reputation system that uses link analysis on real application data to rank workers by true quality. Reputation scores not only contribute significantly as recommenders in hiring, but also are used towards building strategy decision rules for applying to jobs. Our algorithms apply machine learning, efficiently tackling large scale crowd-sourced data from the history of job applications and hiring.

Annotating the Behavior of Scientific Modules Using Data Examples: A Practical Approach

Khalid Belhajjame

A major issue that arises when designing scientific experiments (i.e., workflows) is that of identifying the modules (which are often “black boxes”), that are suitable for performing the steps of the experiment. To assist scientists in the task of identifying suitable modules, semantic annotations have been proposed and used to describe scientific modules. Different facets of the module can be described using semantic annotations. Our experience with scientists from modern sciences such as bioinformatics, biodiversity and astronomy, however, suggests that most of semantic annotations that are available are confined to the description of the domain of input and output parameters of modules. Annotations specifying the behavior of the modules, as to the tasks they play, are rarely specified. To address this issue, we argue in this paper that data examples are an intuitive and effective means for understanding the behavior of scientific modules. We present a heuristic for automatically generating data examples that annotate scientific modules without relying on the existence of the module specifications, and show through an empirical evaluation that uses real-world scientific modules the effectiveness of the heuristic proposed.

Wednesday, March 26			
16:00-17:30	EDBT Industry & Applications Session 3		***Applications***
Room:	Templar’s	Chair	Grigoris Karvounarakis

Benchmarking Bitemporal Database Systems: Ready for the Future or Stuck in the Past?

Martin Kaufmann, Peter Fischer, Norman May and Donald Kossmann

After more than a decade of a virtual standstill, the adoption of temporal data management features has recently picked up speed, driven by customer demand and the inclusion of temporal expressions into SQL:2011. Most of the big commercial DBMS now include support for bitemporal data and operators. In this paper, we perform a thorough analysis of these commercial temporal DBMS: We investigate their architecture, determine their performance and study the impact of performance tuning. This analysis utilizes our recent (TPCTC 2013) benchmark proposal, which includes a comprehensive temporal workload definition. The results of our analysis show that the support for temporal data is still in its infancy: All systems store their data in regular, statically partitioned tables and rely on standard indexes as well as query rewrites for their operations. As shown by our measurements, this causes considerable performance variations on slight workload variations and a significant effort for performance tuning. In some cases, there is considerable overhead for temporal operations even after extensive tuning.

Business-Intelligence Queries with Order Dependencies in DB2

Jaroslaw Szlichta, Parke Godfrey, Jarek Gryz, Wenbin Ma, Weinan Qiu and Calisto Zuzarte

Business-intelligence queries often involve SQL functions and algebraic expressions. There can be clear semantic relationships between a column’s values and the values of a function over that column. A common property is monotonicity: as the column’s values ascend, so do the function’s values. This we call an order dependency (OD). Queries can be evaluated more efficiently when the query optimizer uses order dependencies. They can be run even faster when the optimizer can also reason over known ODs to infer new ones. Order dependencies can be declared as integrity constraints, and they can be detected automatically for many types of SQL functions and algebraic expressions. We present optimization techniques using ODs for queries that involve join, order by, group by, partition by, and distinct. Essentially, ODs can further exploit interesting orders to eliminate or simplify potentially expensive sorts in the query plan. We evaluate these techniques over our implementation in IBM DB2 V10 using the TPC-DS benchmark schema and some customer inspired queries. Our experimental results demonstrate a significant performance gain. We additionally devise an algorithm for testing logical implication for ODs which is polynomial over the size of the set of given ODs. We show that the inference algorithm which we have implemented in DB2 is sound and complete over sets of ODs over natural domains. This enables the optimizer to infer useful ODs from known ODs.

A Tale of Two Graphs: Property Graphs as RDF in Oracle

Souripriya Das, Jagannathan Srinivasan, Matthew Perry, Eugene Chong and Jayanta Banerjee

Graph Databases are gaining popularity, owing to pervasiveness of graph data in social networks, physical sciences, networking, and web applications. A majority of these databases are based on the property graph model, which is characterized as key/value-based, directed, and multi-relational. In this paper, we consider the problem of supporting property graphs as RDF in Oracle Database. We introduce a property graph to RDF transformation scheme. The main challenge lies in representing the key/value properties of property graph edges in RDF. We propose three models: 1) named graph based, 2) subproperty based, and 3) (extended) reification based, all of which can be supported with RDF capabilities in Oracle Database. These models are evaluated with respect toease of SPARQL query formulation, jo in complexities, skewness in generated RDF data, query performance, and storage overhead. An experimental study with a real-life Twitter Social network dataset on Oracle Database 12c demonstrates the feasibility of representing property graphs as RDF and presents a quantitative performance comparison of the proposed models.

EDBT Demo Sessions

Tuesday, March 25 and Wednesday, March 26		
14:00-15:30 (25th) 16:00-17:30 (26th)	EDBT Demo Session 1	
Room:	Olympia Foyer	

Mindmap-Inspired Semantic Personal Information Management

Jenny Rompa, Christos Tryfonopoulos, Costas Vassilakis and George Lepoura

Users nowadays need to manage large amounts of information, including documents, e-mails, contacts, and multimedia content. To facilitate the tasks of organisation, maintenance, and retrieval of personal information, a number of semantics-based methods have emerged; these methods employ (personal) ontologies as an underlying infrastructure for organising and querying the personal information space. In this paper we present OntoFM, a novel personal information management tool that offers a mindmap-inspired interface to facilitate user interactions with the information base. Besides serving as an information retrieval aid, OntoFM allows the user to specify and update the semantic links between information items, constituting thus a complete personal information management tool.

READ: Rapid data Exploration, Analysis and Discovery

Udayan Khurana, Srinivasan Parthasarathy and Deepak Turaga

Exploratory data analysis (EDA) is the process of discovering important characteristics of a dataset or finding data-driven insights in the corresponding domain.

EDA is a human intensive process involving data management, analytic flow deployment and model creation, and data visualization and interpretation. It involves extensive use of analyst time, effort, and skill in data processing as well as domain expertise. In this paper, we introduce READ, a mixed initiative system for accelerating exploratory data analysis. The key idea behind READ is to decompose the exploration process into components that can be independently specified and automated. These components can be defined, reused or extended using simple choice points that are expressed using inference rules, planning logic, and reactive user interfaces and visualization. READ uses a formal specification of the analytic process for automated model space enumeration, workflow composition, deployment, and model validation and clustering. READ aims to reduce the time required for exploration and understanding of a dataset from days to minutes.

Demonstrating Self-Learning Algorithm Adaptivity in a Hardware-Oblivious Database Engine

Max Heimel, Filip Haase, Martin Meinke, Sebastian Bre, Michael Saecker and Volker Markl

The increasingly heterogeneous modern hardware landscape is forcing database vendors to rethink basic design decisions: With more and more architectures to support, the traditional approach of building on hand-tuned operators might simply become too cost- and labor-intensive.

With this problem in mind, we introduced the notion of a hardware-oblivious database engine, which avoids device-specific optimizations and targets multiple different hardware architectures from a single code-base. We demonstrated the feasibility of this concept through Ocelot, a prototypical hardware-oblivious database that uses OpenCL to provide operators that can run on multiple architectures.

In this demonstration, we show how we modified Ocelot to support self-learning algorithm adaptivity: The ability to automatically learn which algorithms are optimal for a given operation on a given hardware architecture. We present how to specify operators that can be executed by multiple algorithms, provide details about the underlying learning and decision routines, and demonstrate how our system picks the optimal algorithm when running on systems with multiple devices, such as CPUs and graphics cards.

SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms

Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos and Christos Tryfonopoulos

Publishing data about individuals, in a privacy-preserving way, has led to a large body of research. Meanwhile, algorithms for anonymizing datasets, with relational or transaction attributes, that preserve data truthfulness, have attracted significant interest from organizations. However, selecting the most appropriate algorithm is still far from trivial, and tools that assist data publishers in this task are needed. In response, we develop SECRETA, a system for analyzing the effectiveness and efficiency of anonymization algorithms. Our system allows data publishers to evaluate a specific algorithm, compare multiple algorithms, and combine algorithms for anonymizing datasets with both relational and transaction attributes. The analysis of the algorithm(s) is performed, in an interactive and progressive way, and results, including attribute statistics and various data utility indicators, are summarized and presented graphically.

SIAS-V in Action: Snapshot Isolation Append Storage - Vectors on Flash

Robert Gottstein, Thorsten Peter, Ilia Petrov and Alejandro Buchmann

Multi-Version Database Management Systems (MV-DBMS) are wide-spread and can effectively address the characteristics of new storage technologies such as Flash, yet they are mainly optimized for traditional storage. A modification of a tuple in a MV-DBMS results in a new version of that item and the invalidation of the old version. Under Snapshot Isolation (SI) the invalidation is performed as an in-place update, which is suboptimal for Flash. We introduce Snapshot Isolation Append Storage – Vectors (SIAS-V), which avoids the invalidation related updates by organising tuple versions as a simple linked list and by utilizing bitmap vectors representing different states of a single version. SIAS-V sequentializes writes and reduces the write-overhead by appending in tuple-version granularity, writing out only completely filled pages, and eliminating in-place invalidation.

In this demonstration we showcase the SIAS-V implementation in PostgreSQL side-to-side with SI. Firstly, we demonstrate that the I/O distribution of PostgreSQL under a TPC-C style workload, exhibits a dominant small-sequential write pattern for SIAS-V, as opposed to a random write dominated pattern under SI. Secondly, we demonstrate how the dense packing of tuple-versions on pages under SIAS-V reduces significantly the amount of data written. Thirdly, we show that SIAS-V yields to stable write performance and low transaction response times under mixed loads. Last but not least, we demonstrate that SIAS-V also provides performance improvements for traditional HDDs.

inWalk: Interactive and Thematic Walks inside the Web of Data

Silvana Castano, Alfio Ferrara and Stefano Montanelli

The goal of this paper is to demonstrate inWalk, an interactive web-based system for linked data exploration featured by the notion of inCloud and thematic walk. The demonstration focuses on the key functionalities of the system for smart data aggregation and navigation.

KIEV: a Tool for Extracting Semantic Relations from the World Wide Web

Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg and Ingeborg Torvik Solvberg

Deriving knowledge from information stored in unstructured documents is a major challenge. The proliferation of knowledge sharing communities such as Wikipedia urge for automatic methods to construct a knowledge base consisting of entities and their relationships for advanced querying.

More specifically, binary relationships representing a fact between two entities can be extracted to populate semantic triple stores or large knowledge bases. In this paper, we present our novel tool KIEV to fulfil this task. It combines a discovery process and a verification process for the entities and the type of relationship. We finally demonstrate three use cases for which KIEV is useful.

AGGREGO SEARCH: Interactive Keyword Query Construction

Gregory Smits, Olivier Pivert, Helene Jaudoin and Francois Paulus

AGGREGO SEARCH offers a novel keyword-based query solution for end users in order to retrieve precise answers from semantic data sources. Contrary to existing approaches, AGGREGO SEARCH suggests grammatical connectors from natural languages during the query formulation step in order to specify the meaning of each keyword, thus leading to a complete and explicit definition of the intent of the search. An example of such a query is it name of person at the head of company and author of article about “business intelligence”. In order to help users formulate such connected keywords queries, a specific autocompletion strategy has been developed. A translation of the user keyword query into SPARQL is performed on-the-fly during the interactive query construction process. For this demonstration, we show how AGGREGO SEARCH has been integrated on top of a mediation system to let users intuitively define explicit and precise keyword queries in order to extract knowledge distributed in heterogeneous large semantic data sources.

Tuesday, March 25 and Wednesday, March 26		
16:00-17:30 (25th) 14:00-15:30 (26th)	EDBT Demo Session 2	
Room:	Olympia Foyer	

R2G: a Tool for Migrating Relations to Graphs

Roberto De Virgilio, Antonio Maccioni and Riccardo Torlone

We present R2G, a tool for the automatic migration of databases from a relational to a Graph Database Management System (GDBMS). GDBMSs provide a flexible and efficient solution to the management of graph-based data (e.g., social and semantic Web data) and, in this context, the conversion of the persistent layer from a relational to a graph format can be very beneficial for an application. R2G provides a thorough solution to this problem with a minimal impact to the application layer: it transforms a relational database r into a graph database g and any conjunctive query over r into a graph query over g. Constraints defined over r are suitably used in the translation to minimize the number of data access required by graph queries. The approach refers to an abstract notion of graph database and this allows R2G to map relational database into different GDBMSs. The demonstration of R2G allows the direct comparison of the relational and the graph approaches to data management.

ALIAS: Author Disambiguation in Microsoft Academic Search Engine Dataset

Michael Pitts, Swapna Savvana, Senjuti Basu Roy and Vani Mandava

We present a system called ALIAS, that is designed to search for duplicate authors from Microsoft Academic Search Engine dataset. Author-ambiguity is a prevalent problem in this dataset, as many authors publish under several variations of their own name, or different authors share similar or same name. ALIAS takes an author name as an input (who may or may not exist in the corpus), and outputs a set of author names from the database, that are determined as duplicates of the input author. It also provides a confidence score with each output. Additionally, ALIAS has the feature of finding a Top-k list of similar authors, given an input author name. The underlying techniques heavily rely on a mix of learning, mining, and efficient search techniques, including partitioning, clustering, supervised learning using ensemble algorithms, and indexing to perform efficient search to enable fast response for near real time user interaction. While the system is designed using Academic Search Engine data, the proposed solution is generic and could be extended to other problems in the category of entity disambiguation. In this demonstration paper, we describe different components of ALIAS and the intelligent algorithms associated with each of these components to perform author name disambiguation or similar authors finding.

gst-Store: An Engine for Large RDF Graph Integrating Spatiotemporal Information

Dong Wang, Lei Zou and Dongyan Zhao

In this paper, we present a spatiotemporal information integrated RDF data management system, called gst-Store. In gst-Store, some entities have spatiotemporal features, and some statements have valid time intervals and occurring locations. We introduce some spatiotemporal assertions into the SPARQL query language to answer the spatiotemporal range queries and join queries. Some examples are listed to demonstrate our demo.

Learn2Learn: A Visual Educational System for Study Planning

Jishang Wei, Georgia Koutrika and Shanchan Wu

The large collection of educational data provides the opportunity to study how students learn and can be a source of valuable knowledge both for students when planning their studies and for educators and administrators for improving their curricula and services. In our work, we mine course relationships and student consumption patterns found in the data. We present a visual analysis system, Learn-2Learn, that mines, visualizes, and allows interaction with such relationships for user-guided study planning and analysis.

Large-scale Semantic Profile Extraction

Michael Gubanov and Michael Stonebraker

Enriching existing data is a well understood and appreciated problem in many contexts. Users of the Web search-engines would love sometimes the engine to do the job of automatically consolidating all pieces of needed information without the need to hop among many Web pages. Large enterprises having millions of data sources with overlapping and complimentary information is another famous example.

Content providers on the Web (or another information source) usually exhibit a specific focus/domain of their postings. For example, information at www.nasdaq.com is usually in financial domain (stock market), Britney Spears is mostly tweeting about music, and the same is usually true for most information sources.

In this paper, we evaluate our distributed system extracting more than a million of semantic user profiles from the Web and justify its applicability to distributed expert mining and ranking of search results.

Helping Teenagers Relieve Psychological Pressures: A Micro-blog Based System

Qi Li, Yuanyuan Xue, Jia Jia and Ling Feng

The rapid development of economy and society brings unprecedentedly intensive competition and adolescent psychological pressures to current teenagers. If these psychological pressures could not be resolved properly, they will turn to mental problems, which will finally lead to serious consequences, such as suicide or aggressive behaviors. Traditional face-to-face psychological diagnosis and treatment cannot meet the demand of relieving teenagers' stress completely due to its lack of timeliness and diversity. With micro-blog becoming a popular media channel for teenagers' information acquisition, interaction, self-expression and emotion release, we present a system called tHelper for sensing and easing teenagers' psychological pressures in study, communication, affection, or self-recognition through micro-blog. The system adopts Gaussian Process to classify a teenager's pressure (i.e., pressure category, as well as pressure level) based on a number of features extracted from his/her tweets. Then the system provides various methods to help pressurized teenagers relieve their stress according to the sensing results, by sending positive stories, proverbs, pictures, or cheerful jokes, suggesting simple breathing and muscle relaxation exercises, guiding the teenagers to write down something for self expression, and at the worst case notifying their guardians (who have registered the system beforehand) via mobile phone messages. tHelper demonstrates how microblog can be turned into a new kind of adolescent mental education mode to complement and enhance the traditional face-to-face treatment by psychological doctors.

WePIGE: The WebLab Provenance Information Generator and Explorer

Clement Caron, Bernd Amann, Camelia Constantin and Patrick Giroux

In this demonstration we will present and illustrate a new provenance model for generating fine-grained data and service dependencies within XML data processing workflows. These mappings are defined as XPath queries and allow to describe the dependencies of a service result on its inputs. In this demonstration we present WePIGE, a set of tools and interfaces for generating and exploring fine-grained data provenance information. WePIGE is part of the WebLab platform, an open environment for integrating such services into complex media mining workflows. We will demonstrate the usage of WePIGE for assisting experts in the exploration of data and process dependencies generated by workflow executions and in the definition of provenance mapping rules.

Learning Event Patterns for Gesture Detection

Felix Beier, Nedal Alaqraa, Yuting Lai and Kai-Uwe Sattler

Usability often plays a key role when software is brought to market, including clearly structured workflows, the way of presenting information to the user, and, last but not least, how he interacts with the application. In this context, input devices as 3D cameras or (multi-)touch displays became omnipresent in order to define new intuitive ways of user interaction.

State-of-the-art systems tightly couple application logic with separate gesture detection components for supported devices. Hard-coded rules or static models obtained by applying machine learning algorithms on many training samples are used in order to robustly detect a pre-defined set of gesture patterns. If possible at all, it becomes difficult to extend these sets with new patterns or to modify existing ones – difficult for both, application developers and end users. Further, adding gesture support for legacy software or for additional devices becomes difficult with this hard-wired approach. In previous research we demonstrated how the database community can contribute to this challenge by leveraging complex event processing on data streams to express gesture patterns. While this declarative approach decouples application logic from gesture detection components, its major drawback was the non-intuitive definition of gesture queries.

In this paper, we present an approach that is related to density-based clustering in order to find declarative gesture descriptions using only a few samples. We demonstrate the algorithms on mining definitions for multi-dimensional gestures from the sensor data stream that is delivered by a Microsoft Kinect 3D camera, and provide a way for non-expert users to intuitively customize gesture-controlled user interfaces – even during runtime.

7. Workshops

Fourth International Workshop on Linked Web Data Management (LWDM) & Third International Workshop on Querying Graph Structured Data (GraphQ)

Room: Olympia 1

8:45-9:00	**Opening**
9:00-10:30	LWDM & GraphQ joint Keynote
10:30	**Coffee Break**
11:00-12:30	Session 1: Graph and Linked Data Querying
11:00	An Event-Driven Approach for Querying Graph-Structured Data Using Natural Language **Richard A. Frost, Wale Agboola, Eric Matthews, and Jon Donais**
11:15	Quantifying the Connectivity of a Semantic Warehouse **Yannis Tzitzikas, Nikos Minadakis, Yannis Marketakis, Pavlos Fafalios, Carlo Allocca, and Michalis Mountantonakis**
11:30	GraphMCS: Discover the Unknown in Large Data Graphs **Elena Vasilyeva, Maik Thiele, Christof Bornhövd, and Wolfgang Lehner**

11:45	Scalable Numerical SPARQL Queries over Relational Databases **Minpeng Zhu, Silvia Stefanova, Thanh Truong, and Tore Risch**
12:00	Graph-driven Exploration of Relational Databases for Efficient Keyword Search **Roberto De Virgilio, Antonio Maccioni, and Riccardo Torlone**
12:15	Similarity Recognition in the Web of Data **Alfio Ferrara, Lorenzo Genta, and Stefano Montanelli**
12:30	**Lunch**
14:00-15:30	Session 2: Applications, Tools and Mining for Graph and Linked Data
14:00	Implementing Iterative Algorithms with SPARQL **Robert Techentin, Barry Gilbert, Adam Lugowski, Kevin Deweese, John Gilbert, Eric Dull, Mike Hinchey, and Steve Reinhardt**
14:15	Mining Diverse Friends from Social Networks **Alfredo Cuzzocrea and Carson K. Leung**
14:30	A Map-Reduce Algorithm for Querying Linked Data based on Query Decomposition into Stars **Christos Nomikos, Manolis Gergatsoulis, Eleftherios Kalogeros, and Matthew Damigos**
14:45	TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples **Kostas Patroumpas, Michalis Alexakis, Giorgos Giannopoulos, and Spiros Athanasiou**
15:00	Frequent Pattern Mining from Dense Graph Streams **Juan J. Cameron, Alfredo Cuzzocrea, Fan Jiang, and Carson K. Leung**
15:15	Performance Optimization for Querying Social Network Data **Florian Holzschuher and René Peinl**
15:30	**Coffee Break**
16:00-17:30	LWDM & GraphQ wrap-up Session
17:30	**Closing**

International Workshop on Exploratory Search in Databases and the Web (ExploreDB)

Room: Olympia 2

8:45-9:00	**Welcome and Opening by the Workshop Chairs**
9:00-10:30	Keynote Talk
9:00	Exploring Big Data using Visual Analytics **Daniel A. Keim**
10:30	**Coffee Break**
11:00-12:30	Research Session 1
11:00	On the Suitability of Skyline Queries for Data Exploration **Sean Chester, Michael Lind Mortensen, and Ira Assent**
11:30	Hippalus: Preference-enriched Faceted Exploration **Panagiotis Papadakos and Yannis Tzitzikas**
12:00	Multi-Engine Search and Language Translation **Steven Simske, Igor Boyko, and Georgia Koutrika**
12:30	**Lunch**
14:00-15:30	Research Session 2
14:00	Exploring RDF/S Evolution using Provenance Queries **Haridimos Kondylakis and Dimitris Plexousakis**
14:30	Skyline Ranking a la IR **George Valkanas, Apostolos N. Papadopoulos, and Dimitrios Gunopulos**

15:00	The DisC Diversity Model **Marina Drosou and Evaggelia Pitoura**
15:30	**Coffee Break**
16:00-17:30	Panel
17:30	**Closing**

International Workshop on Algorithms for MapReduce and Beyond (BeyondMR)

Room: Attica

8:45-9:00	**Opening**
9:30-10:30	Session 1: Algorithm Design for MapReduce
9:00	Principles and Pitfalls in Algorithm Design for MapReduce **Sergei Vassilvitskii**
10:00	Scheduling MapReduce Jobs on Unrelated Processors **Dimitris Fotakis, Ioannis Milis, Emmanouil Zampetakis, and Georgios Zois**
10:30	**Coffee Break**
11:00-12:30	Session 2: Applications of MapReduce
11:00	Binary Theta-Joins using MapReduce: Efficiency Analysis and Improvements **Ioannis Koumarelas, Athanasios Naskos, and Anastasios Gounaris**
11:30	On the Design Space of MapReduce ROLLUP Aggregates **Duy-Hung Phan, Matteo Dell'Amico, and Pietro Michiardi**
12:00	Determining the k in k-means with MapReduce **Thibault Debatty, Pietro Michiardi, and Wim Mees**
12:30	**Lunch**
14:00-15:30	Session 3: Recursion in Data Flow Systems
14:00	Building Improved Data Processing Languages on Naiad **Frank McSherry**
15:00	Tagged Dataflow: a Formal Model for Iterative Map-Reduce **Angelos Charalambidis, Nikolaos Papaspyrou, and Panos Rondogiannis**

15:30	**Coffee Break**
16:00-17:30	Session 4: Systems Similar to MapReduce
16:00	Processing Regular Path Queries on Giraph **Maurizio Nolé and Carlo Sartiani**
16:30	Graph-Parallel Entity Resolution using LSH & IMM **Pankaj Malhotra, Puneet Agarwal, and Gautam Shroff**
17:00	Modular Data Clustering - Algorithm Design beyond MapReduce **Martin Hahmann, Dirk Habich, and Wolfgang Lehner**
17:30	**Closing**

Third International Workshop on Bidirectional Transformations (BX)

Room: Templar’s

8:45	**Opening**
9:00-10:40	Session 1: Types, Transformations, and Benchmark
9:00	Implementing a Bidirectional Model Transformation Language as an Internal DSL in Scala **Arif Wider**
9:15	Towards a framework for multi-directional model transformations **Nuno Macedo, Alcino Cunha, Hugo Pacheco**
9:30	Formalizing Semantic Bidirectionalization with Dependent Types **Helmut Grohne, Andres Löh, and Janis Voigtländer**
9:45	Group discussion
10:00	BenchmarX **Anthony Anjorin, Manuel Alcino Cunha, Holger Giese, Arend Rensink, and Andy Schürr**
10:15	Towards a Repository of Bx Examples **James Cheney, Jeremy Gibbons, James McKinna, and Perdita Stevens**
10:30	Group discussion
10:40	**Coffee Break**
11:00-12:30	Session 2: Databases, Monads, and Lenses
11:00	Intersection Schemas as a Dataspace Integration Technique **Richard Brownlow and Alex Poulovassilis**

11:15	Bidirectional Transformations in Database Evolution: A Case Study "At Scale" **Mathieu Beine, Nicolas Hames, Jens Weber, and Anthony Cleve**
11:30	Group discussion
11:40	Entangled State Monads **Faris Abou-Saleh, James Cheney, Jeremy Gibbons, James McKinna, and Perdita Stevens**
11:55	Spans of Lenses **Michael Johnson and Robert Rosebrugh**
12:10	Group discussion and closing

Third Workshop on Energy Data Management (EnDM)

Room: Templar's

14:00-15:30	Session 1: Modeling Energy Data
14:00	Pipeline Production Data Model **Jitao Yang, Yu Fan, Yinliang Liu, Hui Deng, and Yang Lin**
14:30	Renewable Energy Data Sources in the Semantic Web with OpenWatt **Davide Lamanna and Antonio Maccioni**
15:00	A Generic Ontology for Prosum er-Oriented Smart Grid **Syed Gillani, Frederique Laforest, and Gauthier Picard**
15:30	**Coffee Break**
16:00-17:00	Session 2: Energy Analytics
16:00	Computing Electricity Consumption Profiles from Household Smart Meter Data **Omid Ardakanian, Negar Koochakzadeh, Rayman Preet Singh, Lukasz Golab, and S Keshav**
16:30	ECAST: A Benchmark Framework for Renewable Energy Forecasting Systems **Robert Ulbricht, Ulrike Fischer, Lars Kegel, Wolfgang Lehner, and Hilko Donker**
17:00-17:30	**Panel:** Energy Data Management: Where Are We Headed**?**
17:30	**Closing**

International Workshop on Mining Urban Data (MUD)

Room: Conference Room 2

8:45	**Opening - Introduction by the Organizers – INSIGHT Overview**
9:30-10:25	Session A
9:00	Invited Talk #1
9:45	Mining Trajectory Data for Discovering Communities of Moving Objects **Corrado Loglisci, Donato Malerba, and Apostolos N. Papadopoulos**
10:05	A Case Study in Preprocessing Mobile Sensing Data for Urban Mobility Analysis **Indre Zliobaite and Jaakko Hollmen**
10:25	**Coffee Break**
11:00-12:40	Session B
11:00	Crowd Density Estimation for Public Transport Vehicles **Marcus Handte, Muhammad Umer Iqbal, Stephan Wagner, Wolfgang Apolinarski, Pedro Jose Marron, Eva Maria Munoz Navarro, Santiago Martinez, Sara Izquierdo Barthelemy, and Mario Gonzalez Fernandez**
11:20	Traffic Incident Detection Using Probabilistic Topic Model **Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi**
11:40	Predictive Trip Planning – Smart Routing in Smart Cities **Thomas Liebig, Nico Piatkowski, Christian Bockermann, and Katharina Morik**
12:00	Addressing the Sparsity of Location Information on Twitter **Dimitris Kotzias, Ted Lappas, and Dimitrios Gunopulos**

12:20	Efficient Dissemination of Emergency Information using a Social Network **Iouliana Litou, Ioannis Boutsis, and Vana Kalogeraki**
12:40	**Lunch**
14:00-15:30	Session C
14:00	Invited Talk #2
14:45	Crowdsourcing Turning restrictions for OpenStreetMap **Alexandros Efentakis, Sotiris Brakatsoulas, Nikos Grivas, and Dieter Pfoser**
15:05	Short Papers >> Fast Forward >> • Big Data Analytics for Smart Mobility - A Case Study **Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Mirco Nanni, and Dino Pedreschi** • Smart Applications for Smart City: a Contribution to Innovation **Simona Citrigno, Sabrina Graziano, Francesco Lupia, and Domenico Saccà** • Analysis of Relationships between Road Traffic Volumes and Weather: Exploring Spatial Variation **Jaakko Rantala and James Culley** • SiCi Explorer: Situation Monitoring of Cities in Social Media Streaming Data **Andreas Weiler** • A Cascading Wavelet-Feed Forward Neural Network Approach for Forecasting Traffic Flow **Md. Mostafizur Rahman, Atsuhiro Takasu, and Hafiz Md. Hasan Babu** • Combining a Gauss-Markov model And Gaussian Process for Traffic Prediction in Dublin City Center **François Schnitzler, Thomas Liebig, Shie Mannor, and Katharina Morik**
15:30	**Coffee Break (Poster Set-Up, Poster Session)**
16:00-16:45	Session D
16:00	Poster Session (Continued)
16:25	Invited Talk #3

17:10	Sensing Urban Sounscapes **Tae Hong Park, Jonathan Turner, Michael Musick, Jun Hee Lee, Christopher Jacoby, Charlie Mydlarz, and Justin Salamon**
17:30	**Closing**

Seventh International Workshop on Privacy and Anonymity in the Information Society (PAIS)

Room: Conference Room 1A

8:45	**Opening**
9:00-10:15	Invited Talk #1
9:00	A Hybrid Approach for Privacy-preserving Record Linkage **Murat Kantarcioglu**
10:15-10:30	Session1
10:15	Clustering-based Multidimensional Sequence Data Anonymization **Morvarid Sehatkar and Stan Matwin**
10:30	**Coffee Break**
11:00-12:30	Session 2
11:00	Efficient Multi-User Indexing for Secure Keyword Search **Eirini C. Micheli, Giorgos Margaritis, and Stergios V. Anastasiadis**
11:30	Community Detection in Anonymized Social Networks **Alina Campan, Yasmeen Alufaisan, and Traian Marius Truta**
12:00	Secure Multi-Party Linear Regression **Fida Dankar, Renaud Brien, Carlisle Adams, and Stan Matwin**
12:30	**Lunch**
14:00-15:15	Invited Talk #2

14:00	Data Anonymization: The Challenge from Theory to Practice **Ting Yu**
15:15-15:30	Session 3
15:15	A Privacy Preserving Model for Ownership Indexing in Distributed Storage Systems **Tiejian Luo, Zhu Wang, and Xiang Wang**
15:30	**Coffee Break**
16:00-17:30	Panel
17:30	**Closing**

International Workshop on Multimodal Social Data Management (MSDM)

Room: Conference Room 1B

8:45	**Opening**
9:00-10:30	Keynote
9:00	Social Data and Multimedia Analytics for News and Events Applications **Yiannis Kompatsiaris**
9:50	Discussion
10:30	**Coffee Break**
11:00-12:30	Session 2: Social Data Management
11:00	Event Identification and Tracking in Social Media Streaming Data **Andreas Weiler, Michael Grossniklaus and Marc H. Scholl**
11:30	Recommendation of Multimedia Objects for Social Network Applications **Flora Amato, Francesco Gargiulo, Vincenzo Moscato, Fabio Persia and Antonio Picariello**
12:00	Estimating Completeness in Streaming Graphs **Malay Bhattacharyya, Supratim Bhattacharya, and Sanghamitra Bandyopadhyay**
12:30	**Closing**

Conference Organization

General Chair

Vassilis Christophides, University of Crete, Greece, and
Technicolor R&I Center Paris, France

EDBT Program Chair

Sihem Amer-Yahia, CNRS - LIG, France

ICDT Program Chair

Nicole Schweikardt, University of Frankfurt, Germany

Workshop Chair

Selcuk Candan, Arizona State University, USA

Tutorial Chair

Minos Garofalakis, Technical University of Crete, Greece

Demo Chair

Stratos Idreos, Harvard University, USA

Industrial Chair

Anastasios Kementsietsidis, IBM Research, USA

Proceedings Chair

Vincent Leroy, University of Grenoble and
CNRS - LIG, France

Sponsorship Chair

Dimitris Kotzinos, University of Cergy-Pontoise, France and
FORTH-ICS, Greece

Publicity Chair

Grigoris Karvounarakis, LogicBlox, USA

Website Chair

Yannis Stavrakas, IMIS - Athena RC, Greece

Local Executive Chair

Efi Papastavropoulou, Triaena Tours & Congress, Greece

Organization Support

Graphics

FORTH-ICS: Theodossia Bitzou

Conference brochure layout: Myriki & Co
Conference brochure print: Psimythi Ltd

All photos of Athens included in this booklet have been copied by the FB group “Every Saturday in Athens”.

Conference Helpers

Triaena Tours & Congress: Nikos Aneziris, Simos Fasouliotis, Dimitris Tsertos

EKT: Panagiotis Stathopoulos, Konstantinos Christidis, Marios Alexandrakis, Giannis Desypris