Data Lab @ UOI

Members:


Panos Vassiliadis


Evaggelia Pitoura


Panayiotis Tsaparas


Nikos Mamoulis

...+ 14 students

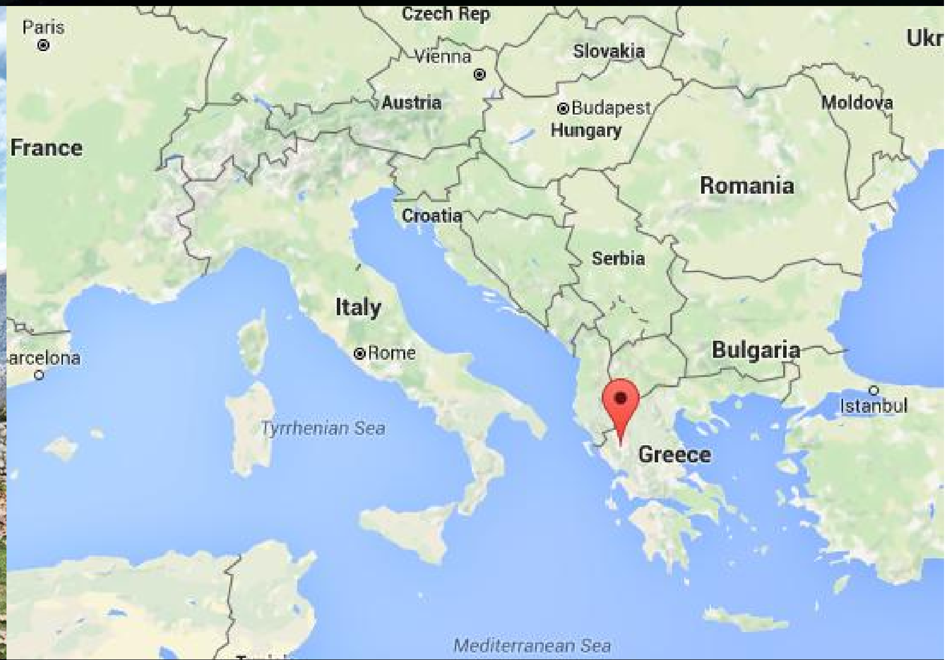
Research Topics:

- Data Warehousing (ETL and OLAP)
- Data Visualization
- Graph Data Analytics, Evolving Graphs
- Query Results Diversification
- Social Media Data Analysis
- Spatial Data Management and Analysis

D. A. T. A.

Data Algorithms Technologies Architectures

University of Ioannina



Outline

- CineCubes: Aiding data workers gain insights from OLAP queries (Panos Vassiliadis, pvassil@cs.uoi.gr)
- Analysis of large social networks that evolve over time (Evaggelia Pitoura, pitoura@cs.uoi.gr)
- Extracting Knowledge from Social Networks and Their Content (Panayiotis Tsaparas, tsap@cs.uoi.gr)
- Search and Analysis of spatially-enriched data (Nikos Mamoulis, nikos@cs.uoi.gr)

Caught somewhere in time

- **Query result = (<u>just</u>) a set of tuples**

- No difference from the 70's when this assumption was established and tailored for
  - what people had available then
    - … a green/orange monochrome screen
    - … a dot-matrix(?) printer
    - … nothing else
  - users being programmers

Photos copied from http://en.wikipedia.org/

NO MORE JUST SETS OF TUPLES!

REPLACE QUERY ANSWERING WITH INSIGHT GAINING!

... but how?

Cinecubes produce small stories presented as data movies ...

to orthogonally combine the following tasks:

STORY

- expand a query result with the results of complementary queries which allow the user to contextualize and analyze the information content of the original query.

with INSIGHTS

- extract meaningful, important patterns, or “highlights” from the query results

Presented as a MOVIE

- present the results (a) properly visualized; (b) enriched with an automatically extracted text that comments on the result; (c) vocally enriched, i.e., enriched with audio that allows the user not only to see, but also hear

http://www.cs.uoi.gr/~pvassil/projects/cinecubes/

1	Assoc	Post-grad	Some-college	University
Gov	40.73	43.58	38.38	42.14
Private	41.06	45.19	38.73	43.06
Self-emp	46.68	47.24	45.70	46.61

Original query

Here, you can see the answer of the original query. You have specified education to be equal to 'Post-Secondary', and work to be equal to 'With-Pay'. We report on Avg of Hrs grouped by education at level 2, and work at level 1. We highlight the largest values with red and the lowest values with blue.

Column Some-college has 2 of the 3 lowest values.

Row Self-emp has 3 of the 3 highest values.

Row Gov has 2 of the 3 lowest values.

2	**Post-Secondary**	Without Post-Secondary
Gov	**41.12**	38.97
Private	**41.06**	39.40
Self-emp	**46.39**	44.84

Summary for education

In this graphic, we put the original request in context by comparing the value 'Post-Secondary' for education at level 3 with its sibling values. We calculate the Avg of Hrs while fixing education at level 4 to be equal to "ALL", and work at level 2 to be equal to "With-Pay'. We highlight the reference cells with bold, the highest value with red and the lowest value with blue.

Compared to its sibling we observe that in 3 out of 3 cases Post-Secondary has higher value than Without-Post-Secondary.

Act I (sl. 2,3)

In this slide, we drill-down one level for all values of dimension work at level 0. For each cell we show both the Avg of Hrs and the number of tuples that correspond to it in parentheses. ...

Column Post-grad has 4 of the 6 highest values.

Column Some-college has 4 of the 6 lowest values.

3	Assoc	Post-grad	Some-college	University
With-Pay	**41.62**	**44.91**	**39.41**	**43.44**
Without-pay	50.00	-	35.33	-

Summary for work

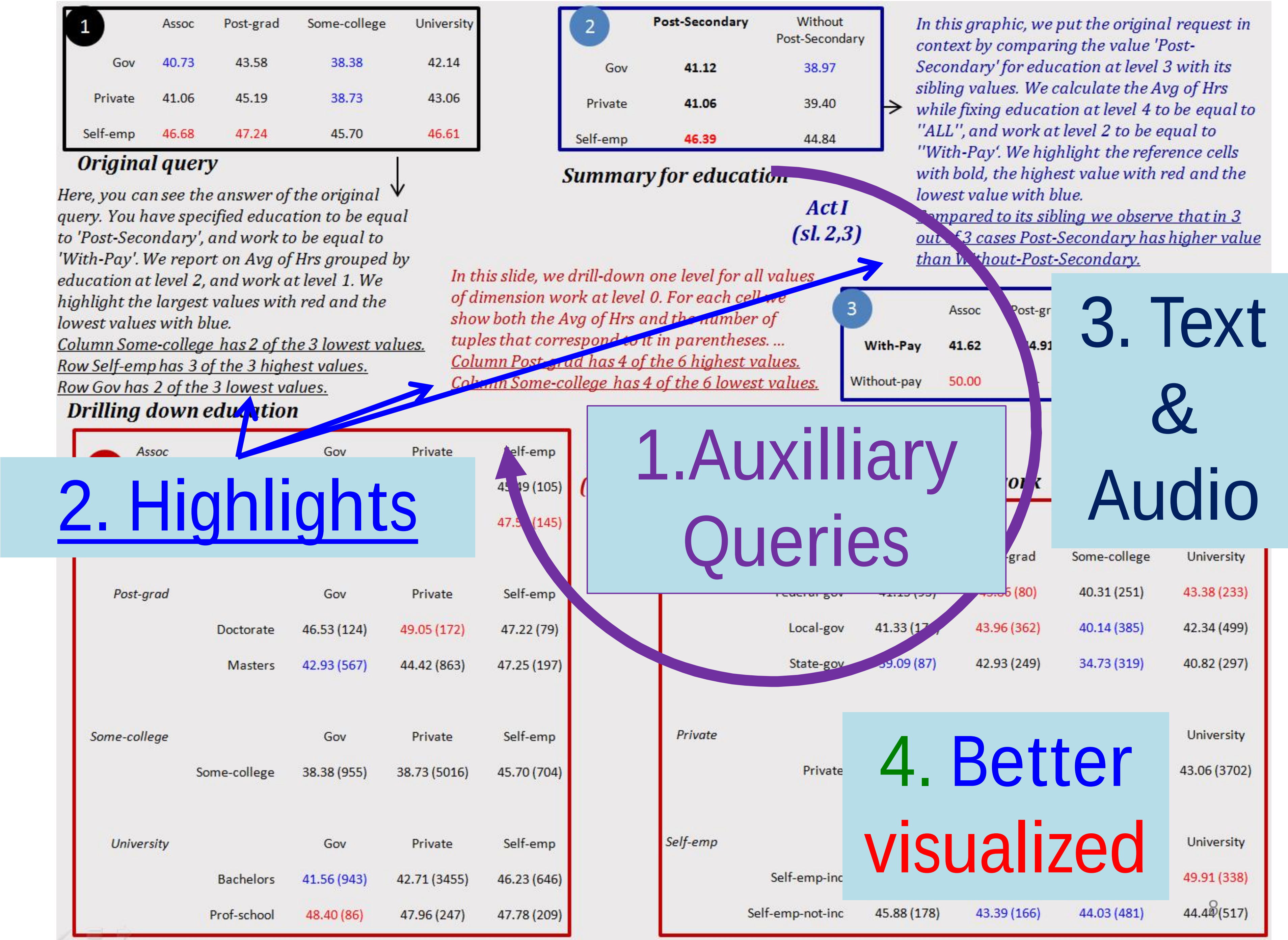
Drilling down education

5		Gov	Private	Self-emp
Assoc		Gov	Private	Self-emp
	Assoc-acdm	39.91 (182)	40.87 (720)	45.49 (105)
	Assoc-voc	41.61 (169)	41.20 (993)	47.55 (145)
Post-grad		Gov	Private	Self-emp
	Doctorate	46.53 (124)	49.05 (172)	47.22 (79)
	Masters	42.93 (567)	44.42 (863)	47.25 (197)
Some-college		Gov	Private	Self-emp
	Some-college	38.38 (955)	38.73 (5016)	45.70 (704)
University		Gov	Private	Self-emp
	Bachelors	41.56 (943)	42.71 (3455)	46.23 (646)
	Prof-school	48.40 (86)	47.96 (247)	47.78 (209)

Act II (sl. 3,4)

Drilling down work

4		Assoc	Post-grad	Some-college	University
Gov		Assoc	Post-grad	Some-college	University
	Federal-gov	41.15 (93)	43.86 (80)	40.31 (251)	43.38 (233)
	Local-gov	41.33 (171)	43.96 (362)	40.14 (385)	42.34 (499)
	State-gov	39.09 (87)	42.93 (249)	34.73 (319)	40.82 (297)
Private		Assoc	Post-grad	Some-college	University
	Private	41.06 (1713)	45.19 (1035)	38.73 (5016)	43.06 (3702)
Self-emp		Assoc	Post-grad	Some-college	University
	Self-emp-inc	48.68 (72)	53.05 (110)	49.31 (223)	49.91 (338)
	Self-emp-not-inc	45.88 (178)	43.39 (166)	44.03 (481)	44.44 (517)


1
Assoc Post-grad Some-college University
Gov 40.73 43.58 38.38 42.14
Private 41.06 45.19 38.73 43.06
Self-emp 46.68 47.24 45.70 46.61
Original query
Here, you can see the answer of the original query. You have specified education to be equal to 'Post-Secondary', and work to be equal to 'With-Pay'. We report on Avg of Hrs grouped by education at level 2, and work at level 1. We highlight the largest values with red and the lowest values with blue.
Column Some-college has 2 of the 3 lowest values.
Row Self-emp has 3 of the 3 highest values.
Row Gov has 2 of the 3 lowest values.
2
Post-Secondary Without Post-Secondary
Gov 41.12 38.97
Private 41.06 39.40
Self-emp 46.39 44.84
Summary for education
In this graphic, we put the original request in context by comparing the value 'Post-Secondary' for education at level 3 with its sibling values. We calculate the Avg of Hrs while fixing education at level 4 to be equal to "ALL", and work at level 2 to be equal to "With-Pay'. We highlight the reference cells with bold, the highest value with red and the lowest value with blue.
Act I
(sl. 2,3)
In this slide, we drill-down one level for all values of dimension work at level 0. For each cell we show both the Avg of Hrs and the number of tuples that correspond to it in parentheses. ...
Column Post-grad has 4 of the 6 highest values.
Column Some-college has 4 of the 6 lowest values.
3
Assoc
With-Pay 41.62
Without-pay 50.00
Drilling down education
Assoc Gov Private Self-emp
Post-grad Gov Private Self-emp
Doctorate 46.53 (124) 49.05 (172) 47.22 (79)
Masters 42.93 (567) 44.42 (863) 47.25 (197)
Some-college Gov Private Self-emp
Some-college 38.38 (955) 38.73 (5016) 45.70 (704)
University Gov Private Self-emp
Bachelors 41.56 (943) 42.71 (3455) 46.23 (646)
Prof-school 48.40 (86) 47.96 (247) 47.78 (209)
Some-college University
40.31 (251) 43.38 (233)
Local-gov 43.96 (362) 40.14 (385) 42.34 (499)
State-gov 42.93 (249) 34.73 (319) 40.82 (297)
Private University
Private 43.06 (3702)
Self-emp University
Self-emp-inc 49.91 (338)
Self-emp-not-inc 45.88 (178) 43.39 (166) 44.03 (481) 44.44 (517)
1.Auxilliary Queries
2. Highlights
3. Text & Audio
4. Better visualized

Play

An example of a slide…

	Assoc	Post-grad	Some-college	University
Gov	40.73	43.58	38.38	42.14
Private	41.06	45.19	38.73	43.06
Self-emp	46.68	47.24	45.70	46.61

Here, you can see the answer of the original query. You have specified education to be equal to 'Post-Secondary‘, and work to be equal to 'With-Pay'. We report on Avg of work hours per week grouped by education at level 2. and work at level 1 .

You can observe the results in this table. We highlight the largest values with red and the lowest values with blue color.

Column Some-college has 2 of the 3 lowest values.

Row Self-emp has 3 of the 3 highest values.

Row Gov has 2 of the 3 lowest values.

Cinecubes resources

Panos Vassiliadis

Dept. of Computer Science & Engineering
Univ. Ioannina, Hellas

pvassil@cs.uoi.gr

Readings, Presentations and Demo

http://www.cs.uoi.gr/~pvassil/projects/cinecubes/

Code

https://github.com/DAINTINESS-Group/CinecubesPublic.git

Dimitrios Gkesoulis, Panos Vassiliadis, Petros Manousis. CineCubes: Aiding data workers gain insights from OLAP queries. Information Systems, Volume 53, October-November 2015, Pages 60 - 86.

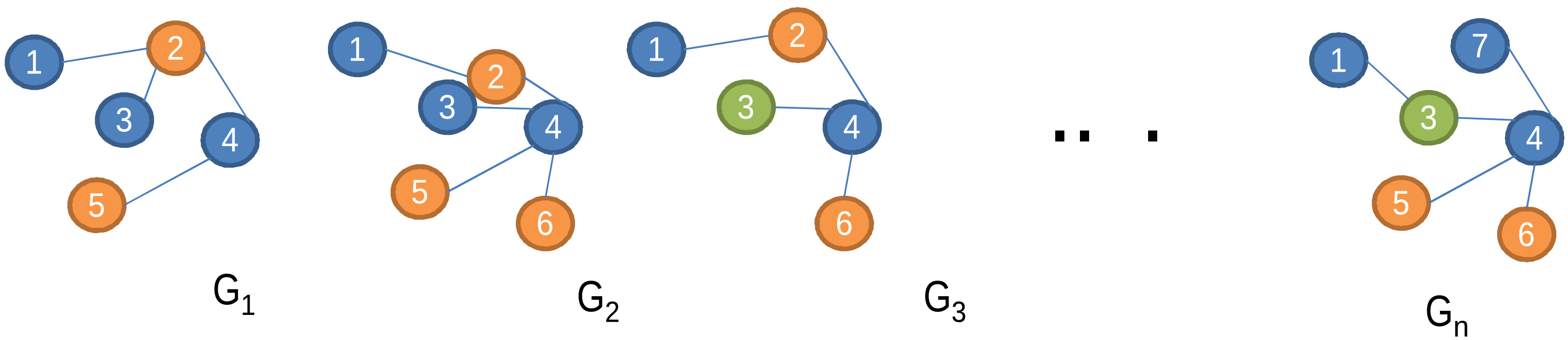
Dimitrios Gkesoulis, Panos Vassiliadis. CineCubes: Cubes as Movie Stars with Little Effort. DOLAP 2013, pp. 3 - 10, 28 October 2013, San Fransisco, USA

Outline

- CineCubes: Aiding data workers gain insights from OLAP queries (Panos Vassiliadis, pvassil@cs.uoi.gr)
- Analysis of large social networks that evolve over time (Evaggelia Pitoura, pitoura@cs.uoi.gr)
- Extracting Knowledge from Social Networks and Their Content (Panayiotis Tsaparas, tsap@cs.uoi.gr)
- Search and Analysis of spatially-enriched data (Nikos Mamoulis, nikos@cs.uoi.gr)

Time-Evolving Graphs

Evolving graph: A sequence of graph *snapshots* G_t at time instance t



- Storage issues
- Indexing issues
- Many novel historical graph queries that:
  - (a) have a time-range dimension *(when)*,
  - (b) consider the durability of results *(how long/how often)*, and
  - (c) capture time evolution (e.g., monitor).

Pattern Matching

Given: graph $G(V, E, L)$, L: V Σ*

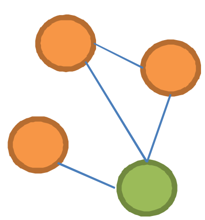
pattern $P(V_P, E_P, L_P)$

Find all subgraphs $m = (V_m, E_m, L_m)$ of G, such that, there exists a *bijective function f :* V_p V_m :

- for all u *in* V_P, $L_p(u)$ $L_m(f(u))$ and
- for each edge (u, v) E_p, $(f(u), f(v))$ E_m

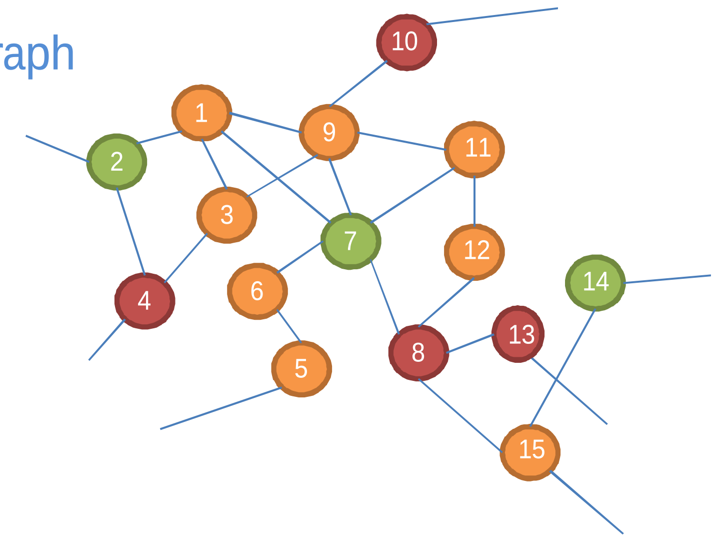
Graph m is called a *match* of P in G

Pattern



color - label

Graph



Identify the most durable matches: the matches that exist for the largest time interval, either collectively (i.e., in the largest number of graph snapshots), or continuously (i.e., in consecutive graph snapshots)

Definition

Duration of a set of time intervals *I*

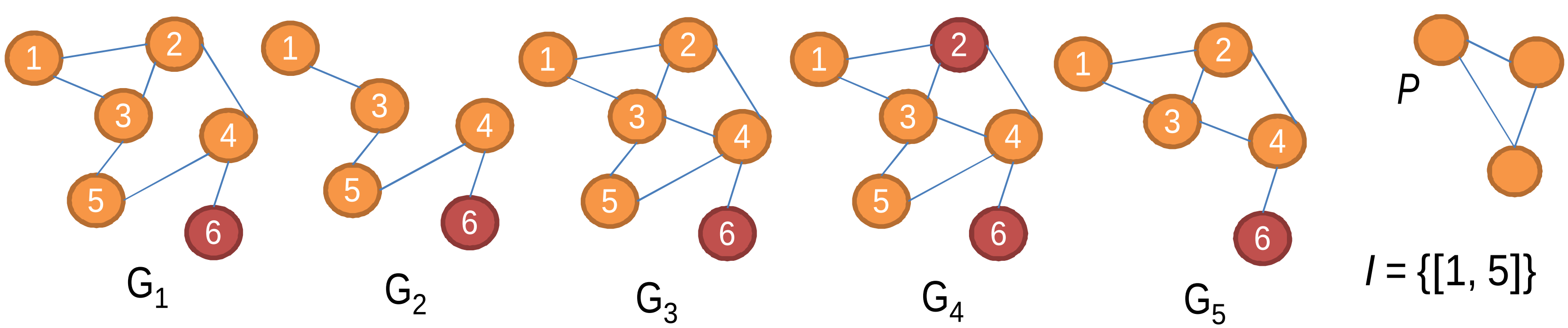
- *collective duration*: the number of time instants in *I*
- *continuous duration*: the duration of the longest time interval in *I*

Example *I* = {[1, 3],[5, 10], [12, 13]} – Collective: 11, Continuous: 6

(Durable Graph Pattern Matching): Given an evolving graph *G[i, j],* a graph pattern query *P* and a set *I* of time intervals:

- A *collective-time durable graph pattern query* finds the matches *m* such that *lifespan(m) I* has the largest collective duration.
- A *continuous-time durable graph pattern query* finds the matches *m* such that *lifespan(m) I* has the largest continuous duration.

time interval intersection/ lifespan: set of time intervals that an element (a node/edge/match, etc) exists



match [1 2 3] lifespan: {[1,1], [3,3], [5,5]} top-collective with collective duration 3
match [3 4 5] lifespan {[3, 4]} top-continuous with continuous duration 2

Durable Graph Pattern algorithm

A Filter-and-Verify algorithm based on three basic concepts :

1. *Version Graph* representation of the snapshot sequence (life-span annotated union graph of the sequence with an efficient in-memory representation using bit-arrays)
2. *Time Graph Indexes* used to filter candidate for matching nodes and refining them
3. *ϑ-duration threshold* that dynamically estimates the minimum duration of a candidate match using the indexes

Historical Reachability Queries

Problem definition

Given nodes *u* and *v* and a set of time intervals *I* are *u* and *v* reachable in *I*

- *Disjunctive*, in at least one time instant in *I*
- *Conjunctive*, in all time instants in *I*

Key concepts

- Find *Strongly-Connected-Components* (SCC) in each graph snapshot
- Use bi-partite matching to *map* SCC at different snapshots effectively
- Store reachability for SCC + extended *2HOP* index

For more information

- K. Semertzidis, and E. Pitoura: *Durable Graph Pattern Queries on Historical Graphs*, *ICDE 2016*

- K. Semertzidis, E. Pitoura, K. Lillis: *TimeReach: Historical Reachability Queries on Evolving Graphs*, *EDBT 2015*

Older papers

- K. Semertzidis, and E. Pitoura: *Time Traveling in Graphs using a Graph Database*, in Proc. of the 5th International Workshop on Querying Graph Structured Data (GraphQ 2016), in conjunction with the EDBT/ICDT 2016
- G. Koloniari, E. Pitoura: Partial view selection for evolving social graphs. *GRADES 2013*
- G. Koloniari, D. Souravlias, E. Pitoura: On Graph Deltas for Historical Queries. Workshop on Online Social Systems (*WOSS 2012*), in conjunction with the VLDB 2012

Outline

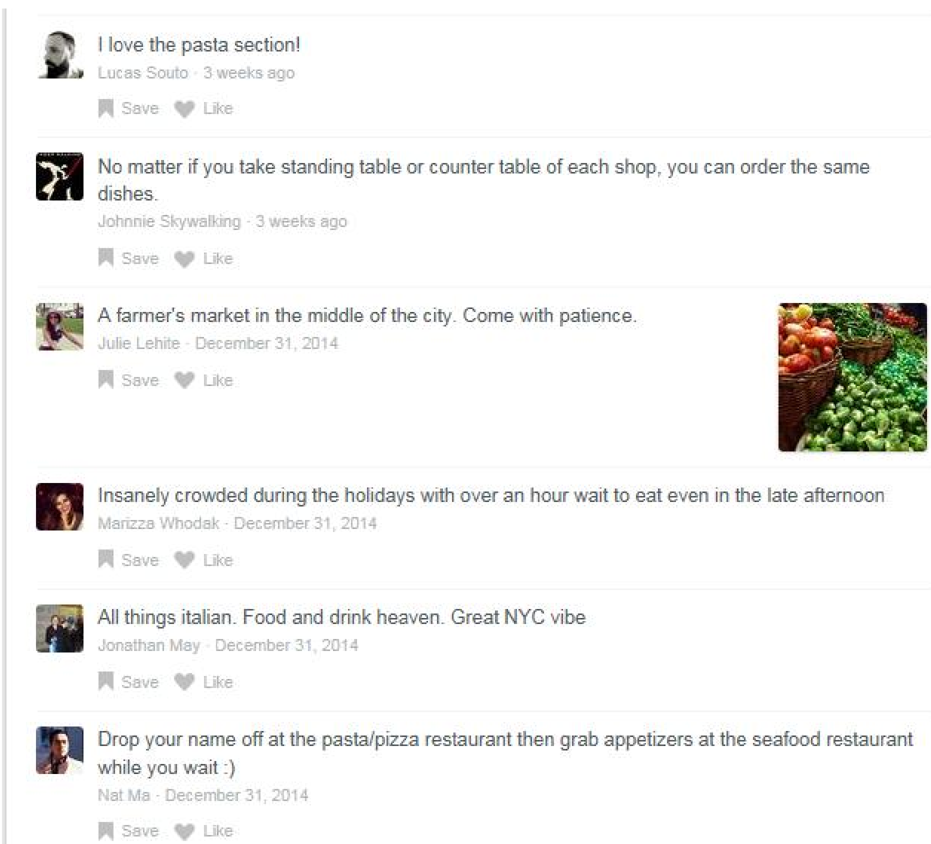
- CineCubes: Aiding data workers gain insights from OLAP queries (Panos Vassiliadis, pvassil@cs.uoi.gr)
- Analysis of large social networks that evolve over time (Evaggelia Pitoura, pitoura@cs.uoi.gr)
- Extracting Knowledge from Social Networks and Their Content (Panayiotis Tsaparas, tsap@cs.uoi.gr)
- Search and Analysis of spatially-enriched data (Nikos Mamoulis, nikos@cs.uoi.gr)

Extracting Knowledge from Social Networks and Their Content

- Social Networks and Media produce huge volume of data with great commercial and scientific value
  - We need new algorithms for modeling and analyzing social networked data.
- Two areas of research
  - Mining user-generated reviews and micro-reviews
  - Understanding relationships and dynamic processes in networks.

Mining Reviews and Micro-Reviews

- Micro-reviews (tips): A new type of User Generated Content!
  - Bite-size reviews (usually under 200 characters) commonly posted on social media or check-in services.



foursquare

facebook

Facebook Places
Who. What. When. And now where.

real-time movie reviews from across the twitterverse
flicktweets

Micro-Review Summarization

- Micro-reviews are highly informative but:
  - Short, repetitive, and in large volumes– tedious to go through them
  - Mostly consumed on mobile devices with limited screen space – hard to go through them
- We need a summary that:
  - Covers the salient points in the micro-reviews
  - Reasonable length
  - Flowing text, easy to read

Reviews vs Micro-Reviews

Reviews

off-site

elaborate and comprehensive

well-written, narrative/descriptive flow

Micro-Reviews

on-site check-ins

concise and distilled

brusque and curt

Reviews and Micro-Reviews are complementary to each other

Goal: Use Reviews to summarize micro-review content

Using reviews to summarize micro-reviews

- Two approaches:
  - Use micro-reviews to select a small set of reviews that capture the review content
    - Model the problem as a coverage problem where we want to cover the micro-reviews while not adding redundant content

    [T-S. Nguyen, H. W. Lauw, P. Tsaparas, *Review Selection Using Micro-reviews*. IEEE Transactions on Knowledge and Data Engineering, TKDE,27(4), 1098-1111, 2015.]
  - Synthesize a new review from review snippets that capture the information in the micro-reviews in a compact form.
    - Use MDL to find a compact and representative subset of snippets

    [T-S. Nguyen, H. W. Lauw, P. Tsaparas, *Review Synthesis for Micro-Review Summarization*. ACM International Conference on Web Search and Data Mining (WSDM), 2015.]

Dynamic Processes on Networks

- Diffusion phenomena on networks have been studied extensively as models for real-world phenomena
  - Viral marketing
  - Epidemic Spreads
  - Public opinion formation.
- An important and well studied problem in this area is the problem of diffusion maximization
  - Find a small set of nodes in the network that will be the initiators such that they maximize the spread of the diffusion on the underlying network
  - Greedy algorithms have provable approximation guarantees and work well in practice

  [D. Kempe, J. Kleinberg, E. Tardos. *Maximizing the Spread of Influence through a Social Network*. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.]

Diffusion maximization

- Diffusion maximization over dynamic networks
  - When the network changes over time the Greedy algorithm no longer has provable guarantees
  - It is not only about who are the initiators, but also when they become active.

  [N. Gayraud, E. Pitoura, P. Tsaparas, *Diffusion Maximization on Evolving networks*, COSN 2015]

- Maximizing positive opinions on social networks
  - Opinions have continuous values as opposed to discrete adoption behavior.
  - Opinion formation models capture the process of opinion diffusion in a social network.
  - Find users to influence to change their opinion to positive so as to maximize the overall positive opinion.

  [A. Gionis, E. Terzi, P. Tsaparas. *Opinion Maximization in Social Networks*. SDM 2013]

Outline

- CineCubes: Aiding data workers gain insights from OLAP queries (Panos Vassiliadis, pvassil@cs.uoi.gr)
- Analysis of large social networks that evolve over time (Evaggelia Pitoura, pitoura@cs.uoi.gr)
- Extracting Knowledge from Social Networks and Their Content (Panayiotis Tsaparas, tsap@cs.uoi.gr)
- Search and Analysis of spatially-enriched data (Nikos Mamoulis, nikos@cs.uoi.gr)

Search and Analysis of spatially-enriched data

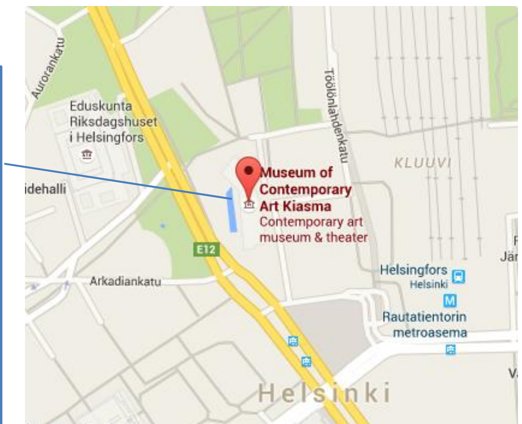
Location information in RDF knowledge bases

subject	*property*	*object*
Dresden	cityOf	Germany
Prague	cityOf	CzechRepublic
Leipzig	cityOf	Germany
Dresden	hosted	Wagner
Leipzig	hosted	Bach
Wagner	hasName	"Richard Wagner"
Wagner	performedIn	Leipzig
Dresden	hasGeometry	"POINT (...)"
Prague	hasGeometry	"POINT (...)"
Leipzig	hasGeometry	"POINT (...)"
...	...	...

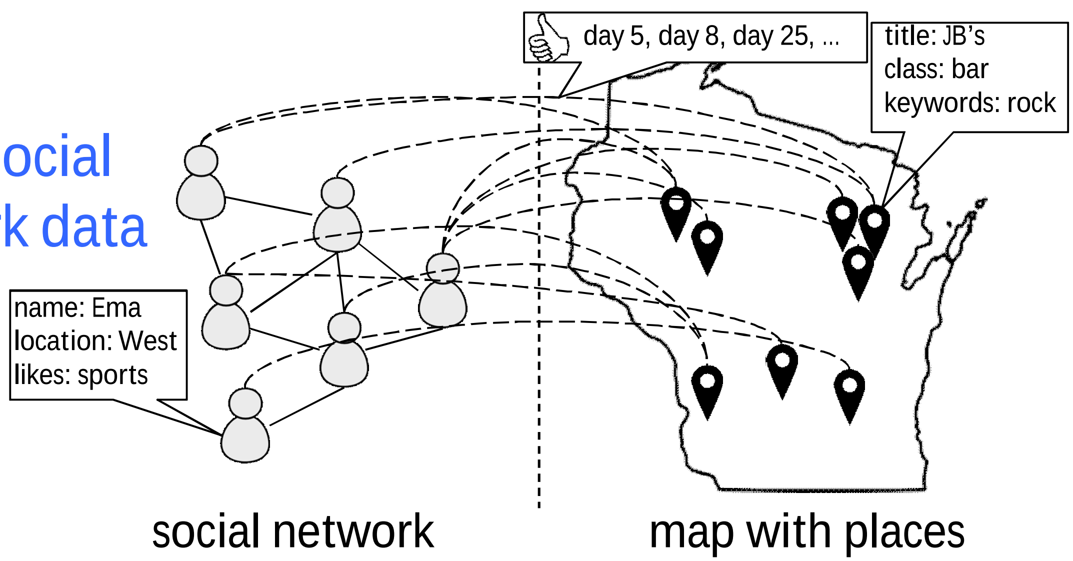
spatial entities

Geo-tagged documents or textually annotated POIs

"Contemporary art museum & cultural center with thematic annual exhibitions, a theater and events"

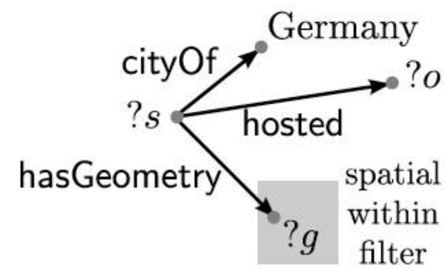


Geo-social network data

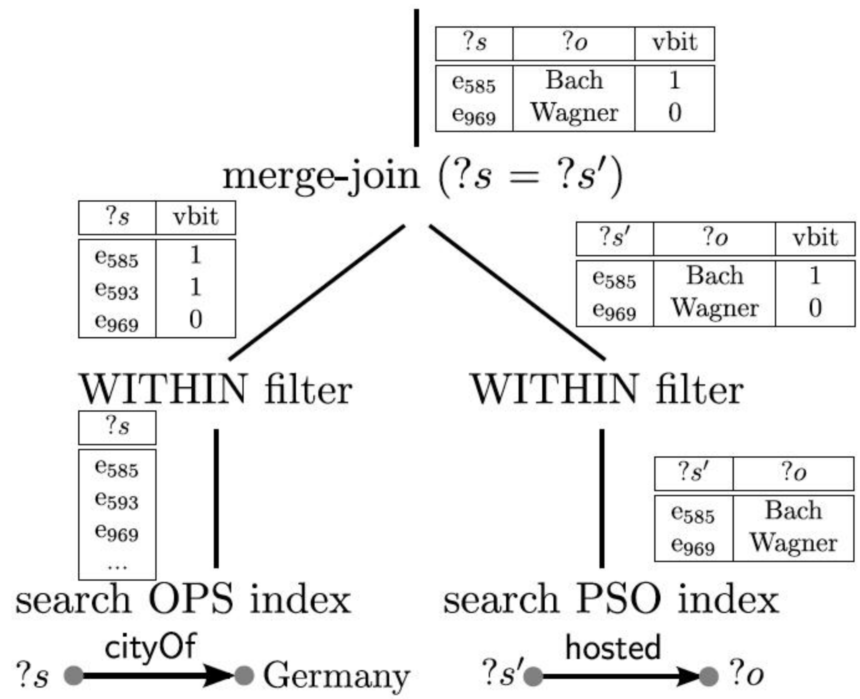


Indexing Locations in large RDF data

1. *Encode locations of spatial entities* into their IDs
   1. Handle points/polygons
   2. Handle spatial skew
2. *Extend RDF-3X* to support Spatial RDF queries
   1. On-the-fly spatial filters using entity IDs
   2. Spatial join algorithm that operate on IDs
   3. Extended query optimizer



spatial selection query



query evaluation plan

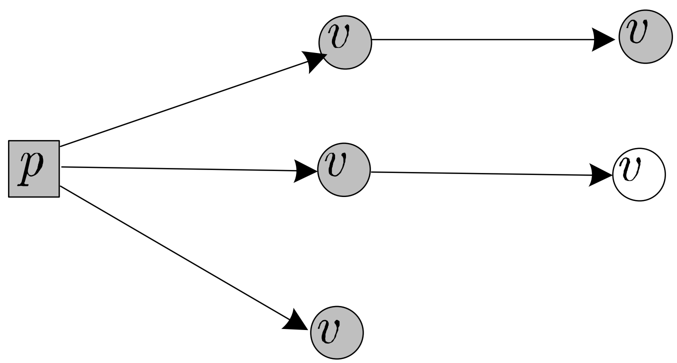
J. Liagouris, N. Mamoulis, P. Bouros, and M. Terrovitis, *"An Effective Encoding Scheme for Spatial RDF Data,"* PVLDB 2014

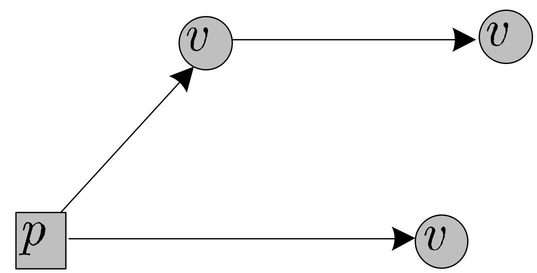
Spatial keyword queries on RDF data

Problem: Find places in RDF data which are close to a query location and they are related to a set of given query keywords.

Key concepts

- Search for RDF subgraphs (1) rooted at place entities near query location (2) containing query keywords
- Measure relevance to keywords by aggregated graph distance of keyword appearance to root
- Spatial-first search, with the help of keyword reachability index / preprocessing





places related to {ancient,roman,catholic,history}

J. Shi, D. Wu, and N. Mamoulis *"Top-k Relevant Semantic Place Retrieval on Spatial RDF Data,"* SIGMOD 2016.

Location Aware Keyword Query Suggestion

Problem: Suggest alternative keyword queries based on:

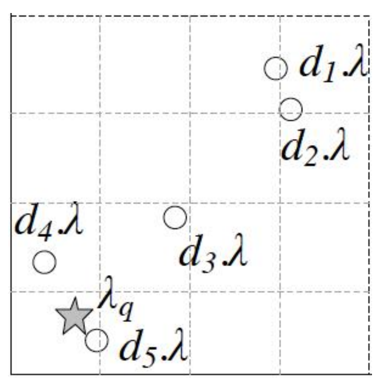
1. semantic relevance to original query and
2. proximity of results to query location

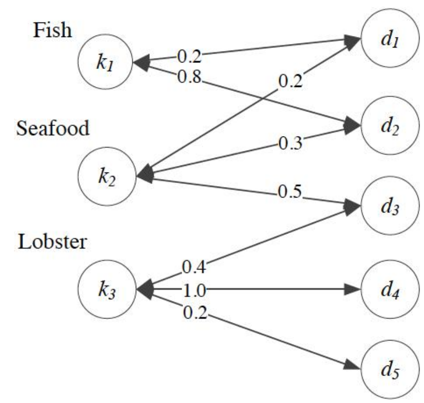
Key concepts

- A Keyword-Document graph connects past keyword queries to documents
- Weights in the KD graph model query-document relevance (from click data)
- Given a keyword query k_q and a location λ_q, edge weights are adjusted
- Location-aware relevant queries to k_q are modeled by their RWR distance to k_q
- The graph is partitioned for more efficient RWR-based query suggestion

d_1	Fish and Seafood
d_2	Fish Seafood
d_3	Lobster Seafood
d_4	Lobster Restaurant
d_5	Lobster House

k_1	Fish
k_2	Seafood
k_3	Lobster





S. Qi, D. Wu, and N. Mamoulis, *"Location Aware Keyword Query Suggestion Based on Document Proximity,"* IEEE TKDE + ICDE 2016 poster

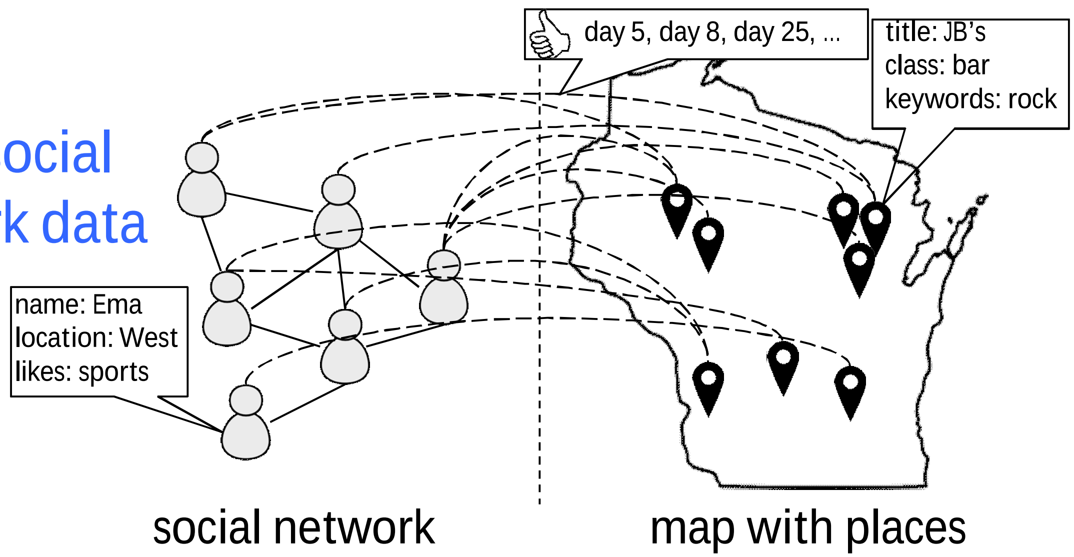
Location Recommendation using Check-in Data

Problem: Recommend places to a geo-social network user

Our recommender system considers

- locations of recommended venues (faraway venues are disregarded)
- similarity between check-in histories of users (artificial weighted edges are added to the social graph connecting users with similar profiles)
- social relationships between users (RWR graph distance is used to model similarities between users and then CF is applied)

Geo-social network data



H. Wang, M. Terrovitis, and N. Mamoulis, *"Location Recommendation in Location-based Social Networks using User Check-in Data,"* GIS 2013

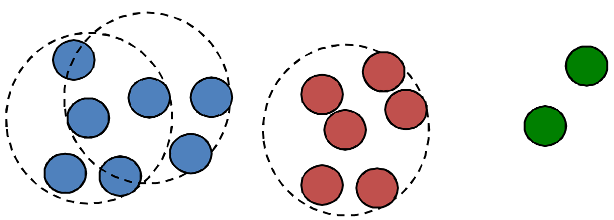
Clustering Places in Geo-Social Networks

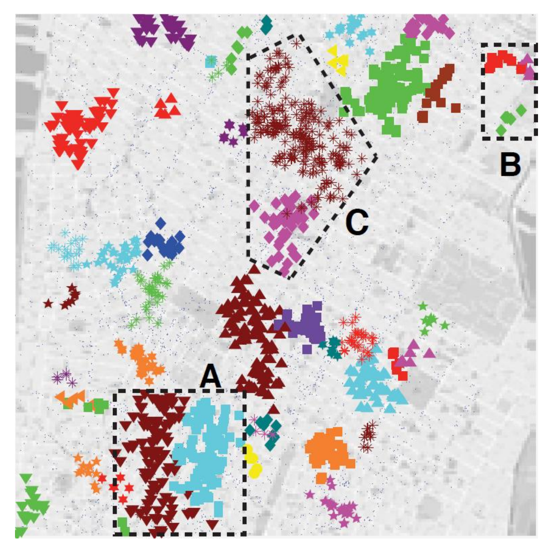
Problem: Consider the geo-social network when clustering places in an urban map

Key concepts

$$D_S(p_i, p_j) = 1 - \frac{|CU_{ij}|}{|U_{p_i} \cup U_{p_j}|}$$

- § Social distance between places
  CU_{ij}: *Contributing users* for (p_i, p_j)
  - users who visited both places
  - users who visited one place AND have a friend who visited the other
- § Geo-social distance between places
  - weighted sum of social and spatial distance
- § Clustering extends DBSCAN to use geo-social distance





J. Shi, N. Mamoulis, D. Wu, and D. W. Cheung, "*Density-based Place Clustering in Geo-Social Networks*," SIGMOD 2014.

Joint Search by Social and Spatial Proximity

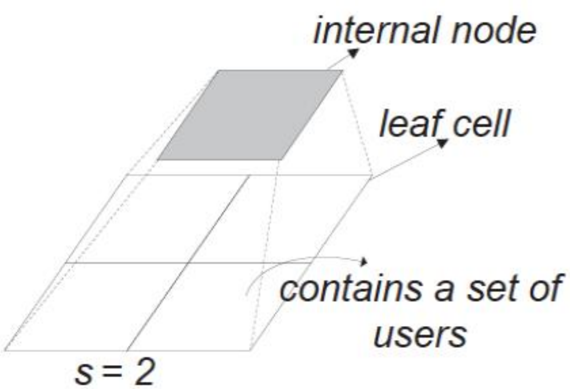
Problem: Find mobile users who are geographically close and socially near the target user

Key concepts

- a hybrid ranking function, which combines both distances

$$f(u_i, u_j) = \alpha \cdot E(u_i, u_j) + (1-\alpha) \cdot D_S(u_i, u_j)$$

- A multi-level spatial grid index with social summaries (aggregated distances to landmarks in the social graph)

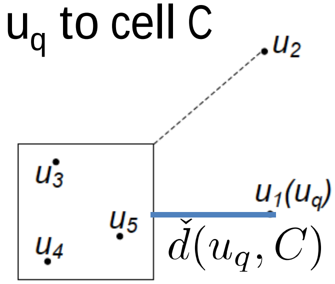


m_{ij} : distance between node v_i and j-th landmark

for a cell C: $\hat{m}[j] = \max_{v_i \in C} m_{ij}$, $\check{m}[j] = \min_{v_i \in C} m_{ij}$

define lower distance bounds from query user u_q to cell C

$$\check{p}(u_q, C) = \max_{1 \le j \le M} \begin{cases} \check{m}[j] - m_{qj} & \text{if } m_{qj} < \check{m}[j] \\ m_{qj} - \hat{m}[j] & \text{if } m_{qj} > \hat{m}[j] \\ 0 & \text{otherwise} \end{cases}$$



prioritize examination of cells and their contents based on

$$MINF(u_q, C) = \alpha \cdot \check{p}(v_q, C) + (1-\alpha) \cdot \check{d}(u_q, C)$$

K. Mouratidis, J. Li, Y. Tang, and N. Mamoulis, "*Joint Search by Social and Spatial Proximity*," IEEE TKDE + ICDE 2016 poster.

More Info

http://dmod.eu/data_web/