Privacy for Spatial Queries and Data

Christian S. Jensen

www.cs.aau.dk/~csj

joint work with Man Lung Yiu, Hua Lu,
Jesper Møller, Gabriel Ghinita, and Panos Kalnis

Motivation

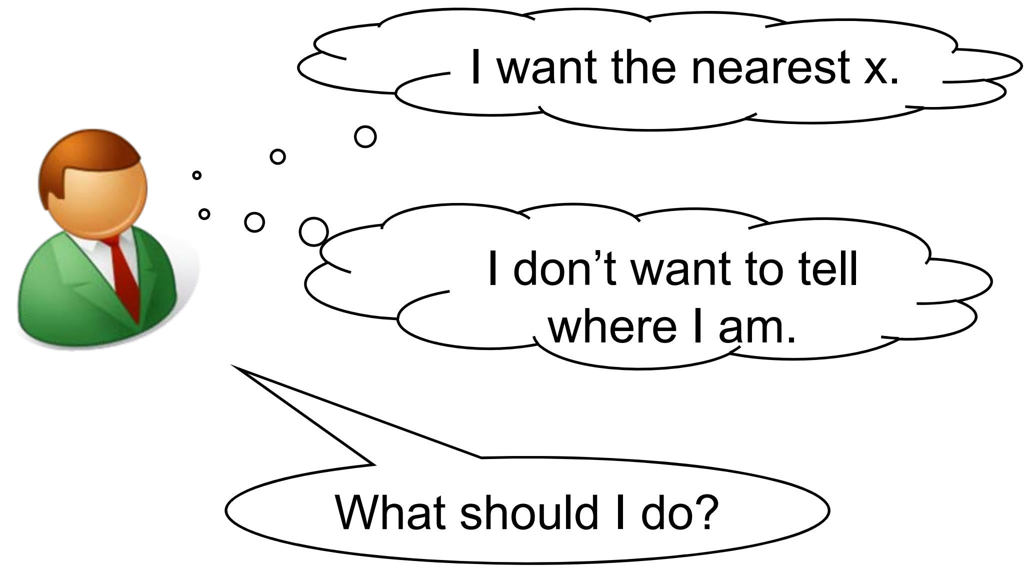
- Outsourcing and cloud computing are on the rise.
- Big and growing mobile Internet
  - 2.7 B mobile phone users (cf. 850 MM PCs)
  - 1.1 B Internet users, 750 MM access the Internet from phones
  - This year, 1.2 B mobile phones will be sold, 200 MM high-end (cf. 200 MM PCs); 13 MM new users in China and India monthly
  - Africa has surpassed North America in numbers of users
- The mobile Internet will be location aware.
  - GPS, Wi-Fi-based, cell-id-based, Bluetooth-based, other
  - A very important signal in a mobile setting!
- Privacy is an enabling technology.

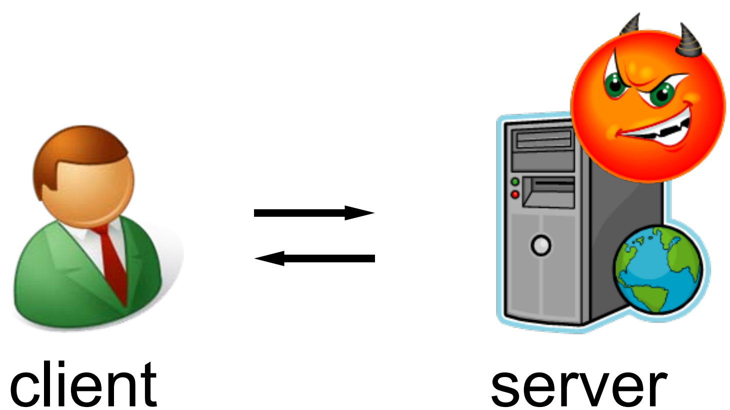
Outline

- Query Location Privacy
  - Motivation and related work
  - Solution: SpaceTwist
  - Granular search in SpaceTwist
  - Empirical study
  - Summary
- Spatial Data Privacy
- Closing remarks

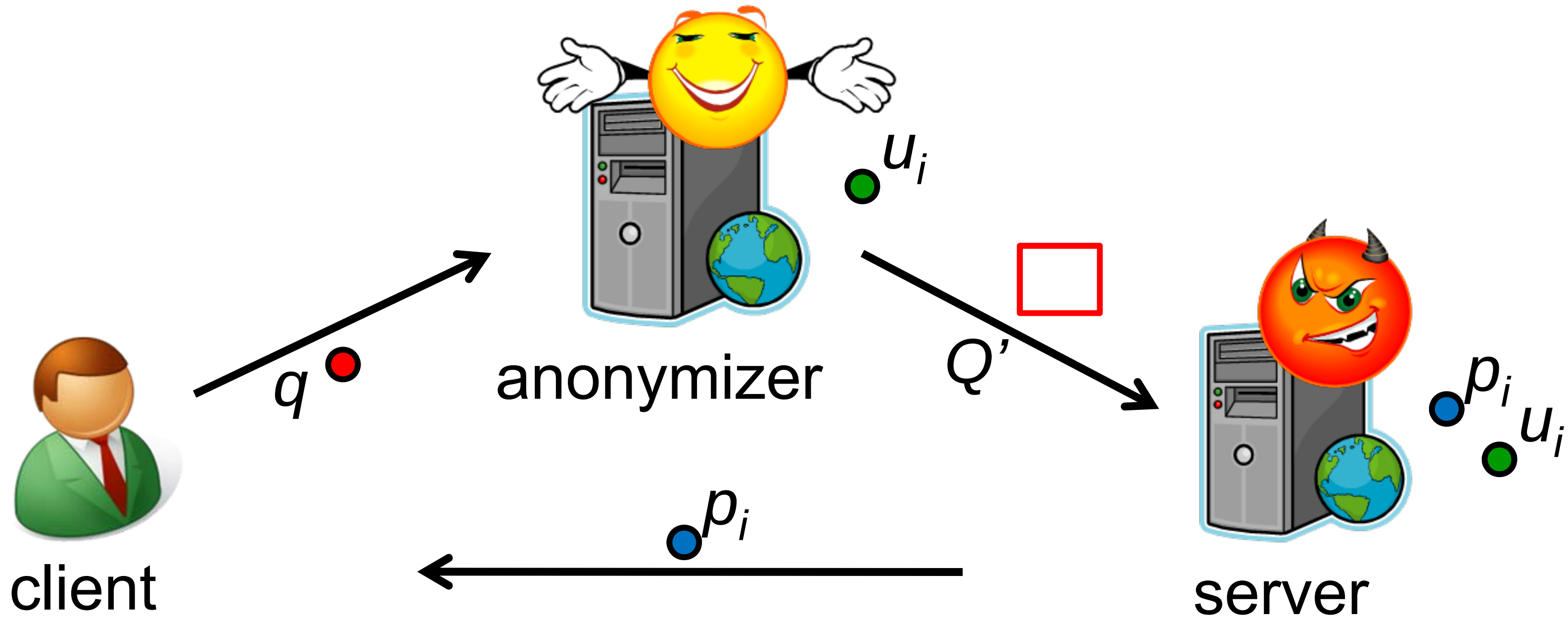
Query Location Privacy

- A mobile user wants nearby points of interest.
- A service provider offers this functionality.
  - Requires an account and login
- The user does not trust the service provider.
  - The user wants location privacy.

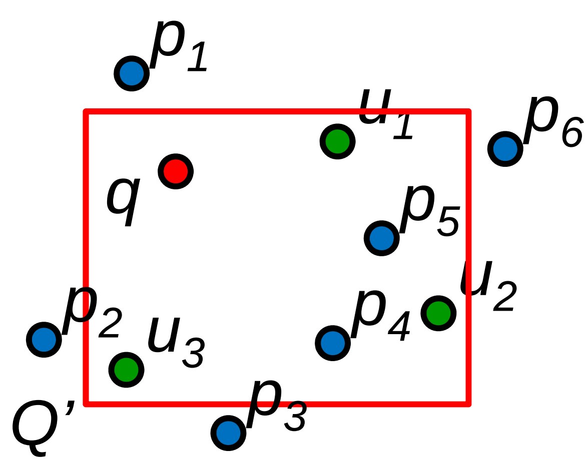




Spatial Cloaking

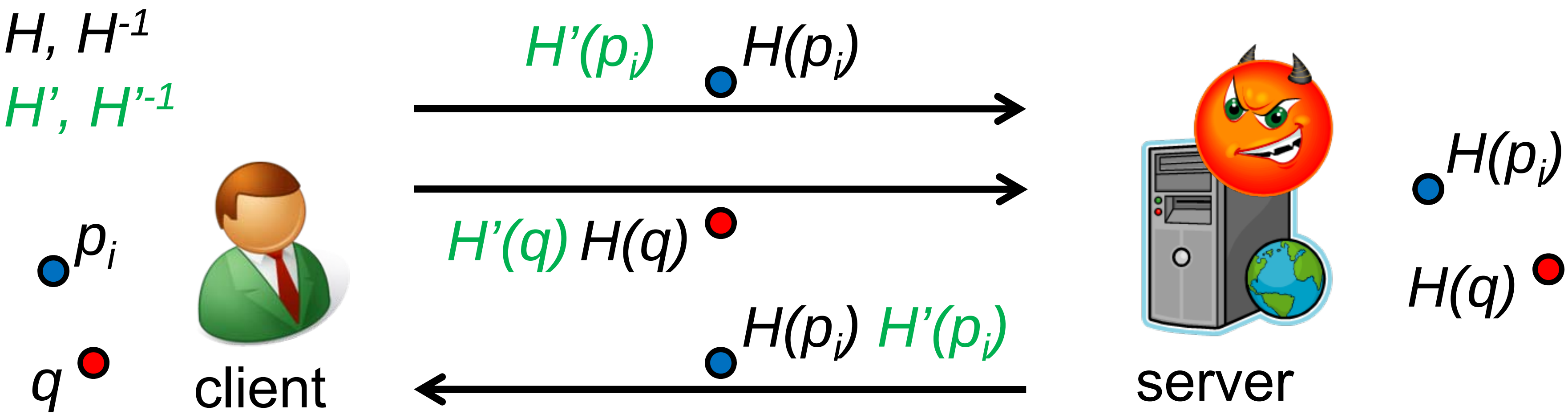


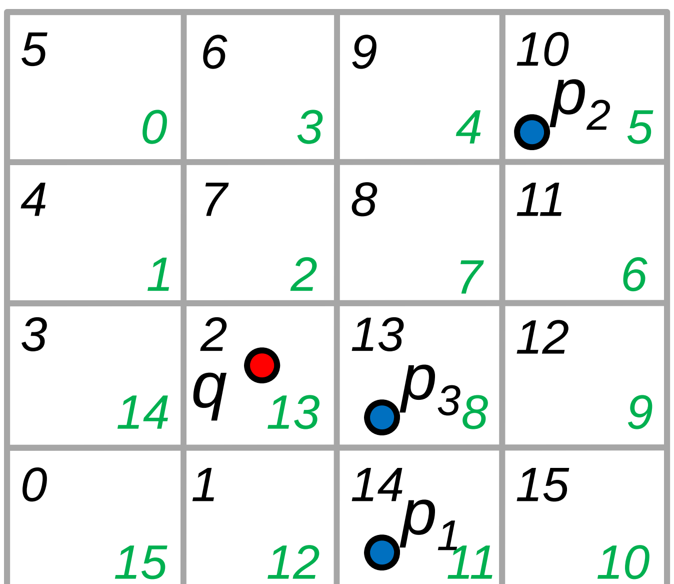
- kNN query (k=1)
- K anonymity
- Range kNN query
- Candidate set is $\{p_1, \ldots, p_6\}$
- Result is p_1



- Identity vs. location privacy
- p-2-p or only client
- Cloaking wo. K anonymity
- Q' may be other shapes, dummies.

Transformation-Based Privacy





{10,13,14}

$H(q) = 2$

10

{5,8,11}

$H'(q) = 13$

11

Definitions of Privacy

- *K*-anonymity: The user cannot be distinguished from *K-1* other users.
- The area of the region within which the user's position can be.
- The average distance between the true position and all possible positions.

Solution Requirements

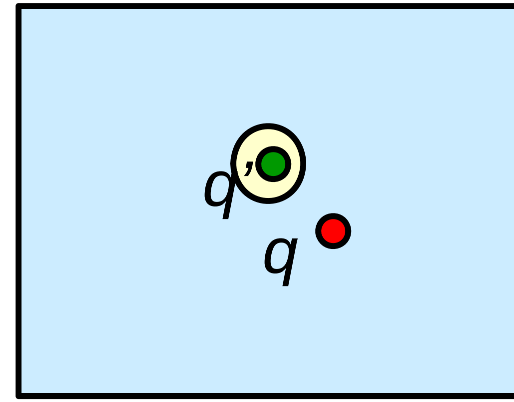
- The solution must enable the user to retrieve the nearest points of interest while affording the user location privacy.
  - Should offer flexibility in the degree of privacy guaranteed, so that the user can decide
    - Settings should be meaningful to the user
    - Like browser security settings or a slider
  - Should work with a standard client-server architecture
    - The user trusts only the mobile client
  - Should assume a typical setting where the user must log in to use the service
  - Should provide privacy at low performance overhead
    - Server-side costs – workload and complexity
    - Communication costs – bits transferred
    - Client-side costs – workload, complexity, power
  - Should enable better performance by reducing the result accuracy

SpaceTwist Concepts

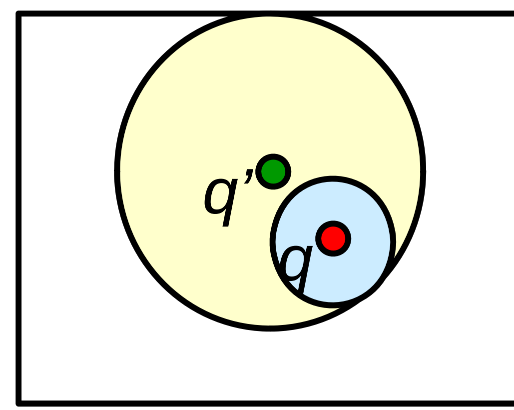
- Anchor location *q'* (*fake* client location)
  - Defines an ordering on the data points
- Client fetches points from server incrementally
- Supply space

  supply space
  - The part to space explored by the client so far
  - Known by both server and client
  - Grows as more data points are retrieved
- Demand space

  demand space
  - Guaranteed to cover the actual result
  - Known only by the client
  - Shrinks when a "better" result is found
- Termination when the supply space contains the demand space

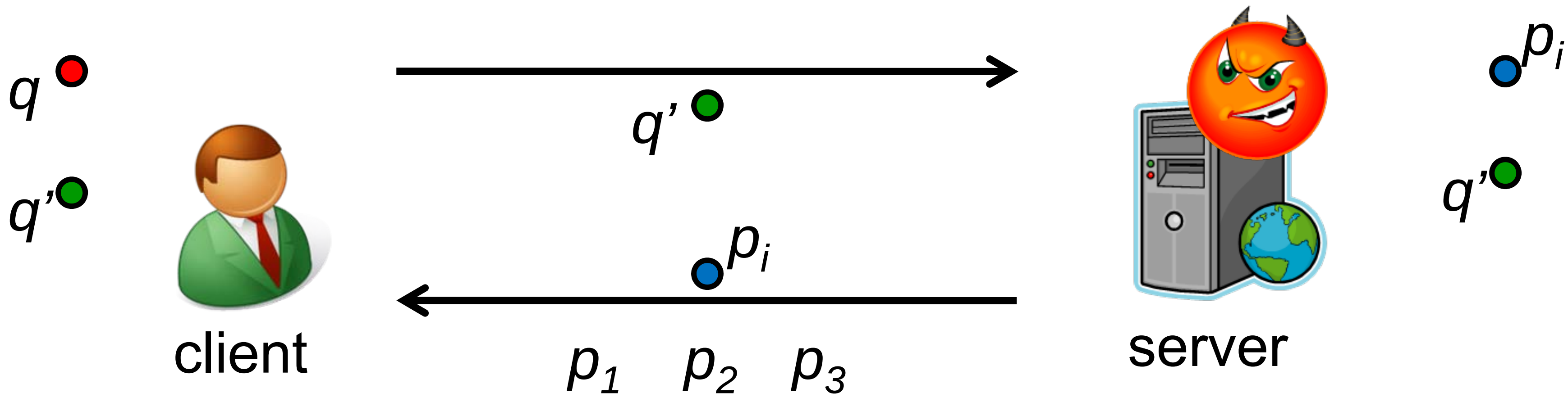


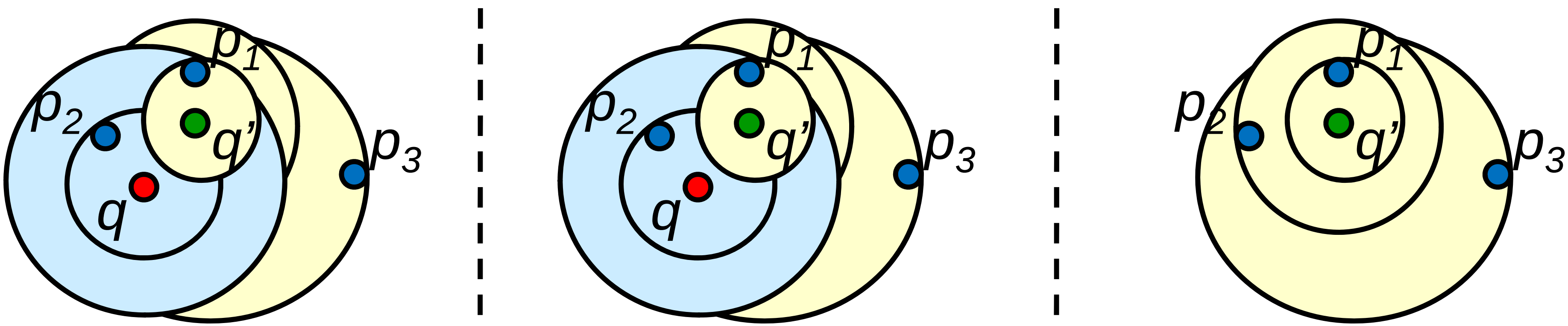
the beginning



the end

SpaceTwist Example





SpaceTwist Properties

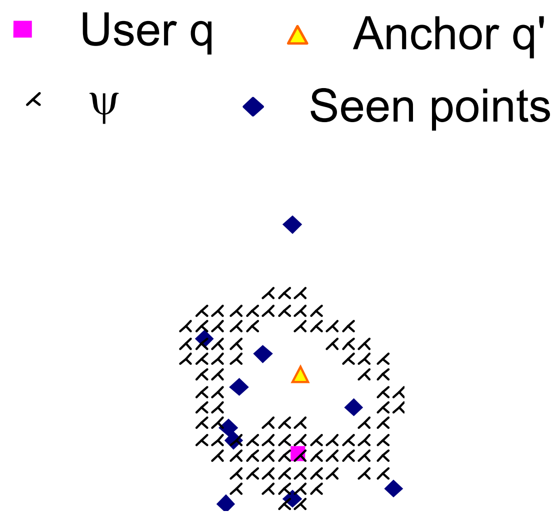
- Retrieves data points from the server incrementally until the client can produce the exact result
- Fundamentally different from previous approaches
  - No cloaking region
  - Queries are evaluated in the original space.
- Offers privacy guarantees
- Relatively easy to support in existing systems
  - Simple client-server architecture (no trusted components, peers)
  - Simple server-side query processing: incremental NN search
- Granular search (improved server-side performance)
  - Reduced communication cost for results with guaranteed accuracy

Privacy Analysis

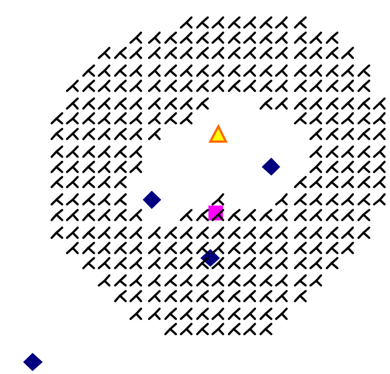
- What does the server know?
  - The anchor location q'
  - The reported points (in reporting order): $p_1, p_2, \ldots, p_{m\beta}$
  - Termination condition: $\text{dist}(q,q') + \text{dist}(q,\text{NN}) \le \text{dist}(q', p_{m\beta})$
- Possible query location q_c
  - The client did not stop at point $p_{(m-1)\beta}$
    - $\text{dist}(q_c, q') + \min\{ \text{dist}(q_c, p_i) : i \in [1,(m-1)\beta] \} > \text{dist}(q', p_{(m-1)\beta})$
  - Client stoped at point $p_{m\beta}$
    - $\text{dist}(q_c, q') + \min\{ \text{dist}(q_c, p_i) : i \in [1,m\beta] \} \le \text{dist}(q', p_{m\beta})$
- *Inferred* privacy region Ψ: the set of all possible q_c
- Quantification of privacy
  - Privacy value: $\Gamma(q, \Psi)$ = the average dist. of location in Ψ from q

Visualization of Ψ

- Visualization with different types of points
- Characteristics of Ψ (i.e., possible locations q_c)
  - Roughly an irregular ring shape centered at *q'*
  - Radius approx. dist(*q*,*q'*)



β=4



coarser granularity

Privacy Analysis

- By carefully selecting the distance between *q* and *q'*, it is possible to guarantee a privacy setting specified by the user.
- SpaceTwist extension: Instead of terminating when possible, request additional query points.
  - This makes the problem harder for the adversary.
  - It makes it easier (and more practical) to guarantee a privacy setting.

Communication Cost

- The communication cost is the number of (TCP/IP) packets transmitted.
- It is inefficient to use a packet for each point.
- Rather packets are filled before transmission.
  - The packet capacity β is the number of points in a packet.
- Actual value of β?
  - Depends on the Maximum Transmission Unit (MTU)
  - In empirical studies, we use MTU = 576 bytes and β = 67.
- The cost has been characterized analytically.
- Empirical studies have been conducted.

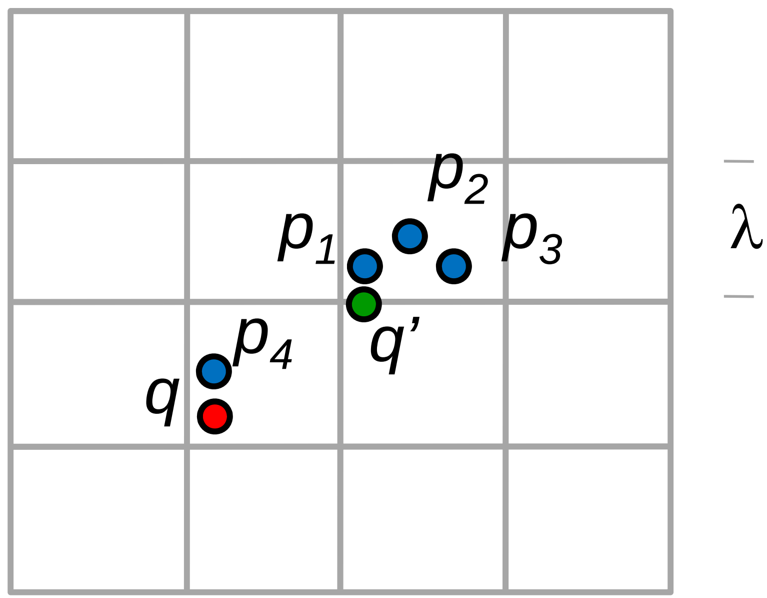
Granular Search

- What if the server considers searching on *a small sample* of the data points instead of all?
  - Lower communication cost
  - Ψ becomes large at low data density
  - But less accurate results

- Accuracy requirement: the user specifies an error bound ε
  - A point $p \in P$ is a relaxed NN of q iff
    $\mathrm{dist}(q, p) \leq \varepsilon + \min \{\mathrm{dist}(q, p') : p' \in P\}$

- A grid with cell length $\lambda = \varepsilon / \sqrt{2}$ is applied.
- As before, the server reports points in ascending distance from *q'*, but it never reports more than one data point *p* from the same cell.

Granular Search Example



Experimental Study

- Our solution GST (Granular SpaceTwist)
  - Without delayed termination
- Spatial datasets (domain: $[0,10,000]^2$)
  - Two real datasets: SC (172,188 pts), TG (556,696 pts)
  - Synthetic uniform random UI datasets
- Performance metrics (workload size = 100)
  - Communication cost (in number of packets; 1 packet = 67 points)
  - Result error (result NN distance – actual NN distance)
  - Privacy value of *inferred* privacy region Ψ
- Default parameter values
  - Anchor distance dist(q,q'): 200
  - Error bound ε: 200
  - Data size N: 500,000

Transformation-Based Privacy Vs. GST

- Hilbert transformation [Khoshgozaran and Shahabi, 2007]
  - SHB: single Hilbert curve
  - DHB: two orthogonal Hilbert curves
- GST computes result with low error
  - Very low error on real (skewed) data
  - Stable error across different data distributions

	Error (meter)								
	UI, N=0.5M			*SC*			*TG*		
k	**SHB**	**DHB**	**GST**	**SHB**	**DHB**	**GST**	**SHB**	**DHB**	**GST**
1	7.1	2.2	51.3	1269.3	753.7	2.5	1013.9	405.8	16.1
2	9.3	4.0	49.0	1634.3	736.2	2.6	1154.6	548.7	16.7
4	13.2	6.0	47.6	1878.5	810.9	2.6	1182.3	596.5	17.0
8	19.0	7.3	42.0	2075.6	864.5	2.6	1196.2	599.7	16.3
16	27.0	10.3	36.3	2039.6	985.7	2.6	1199.6	603.2	14.5

Spatial Cloaking Vs. GST

- Our problem setting: no trusted middleware
- Competitor: client-side spatial cloaking (CLK)
  - CLK: enlarge *q* into a square with side length 2*dist(*q*,*q'*)
    - Extent comparable to inferred privacy region Ψ of GST
  - GST produces result at low communication cost
    - Low cost even at high privacy
    - Cost independent of N

varying dist(*q*,*q'*)

	SC		*TG*	
dist(q, q')	**CLK**	**GST**	**CLK**	**GST**
50	1.3	1.0	1.9	1.0
100	2.0	1.0	4.6	1.0
200	6.2	1.0	15.0	1.0
500	33.5	1.1	72.8	1.3
1000	107.0	1.4	282.0	2.6

	UI	
N(million)	**CLK**	**GST**
0.1	3.0	1.0
0.2	5.1	1.0
0.5	12.2	1.0
1	23.9	1.0
2	47.5	1.0

varying data size N

communication cost (# of packets)

Summary

- SpaceTwist is a novel solution for query location privacy of mobile users
  - Granular search at the server
- Advantages
  - Guaranteed, flexible privacy settings
  - Assumes only a simple client-server setting
  - Low processing and communication cost
  - Enables trading of (guaranteed) accuracy for performance
- Extensions
  - Ring-based server-side retrieval order, spatial networks
- Future work
  - Additional query types

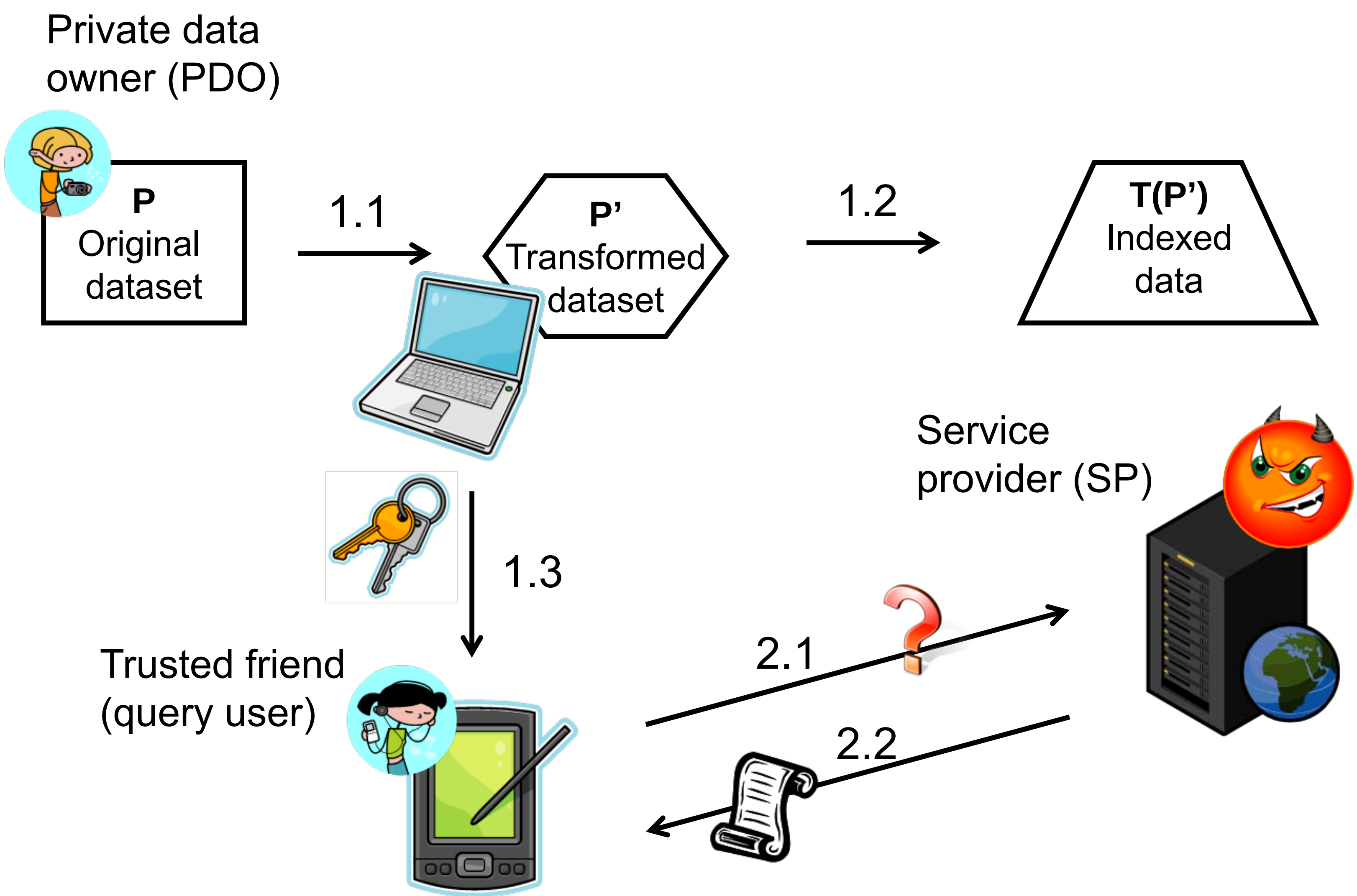
Outline

- Query Location Privacy
  - Motivation and related work
  - Solution: SpaceTwist
  - Granular search in SpaceTwist
  - Empirical study
  - Summary
- Spatial Data Privacy
  - Problem setting, solution framework, and objectives
  - Tailored and general attack models
  - Solution overview
  - Summary
- Closing remarks

Problem Setting

- On a trip to Paris, Alice takes photos with her GPS phone camera
  - Private spatial data: each photo tagged with its GPS location (automatically)
  - Example of user-generated content
- Alice wants to outsource spatial search on the above data to a service provider, e.g., *Flickr*, *Facebook, Picasa*
- Trusted query users: Alice’s friends
  - Nobody else (including the service provider) can be trusted

Solution Framework



Objectives

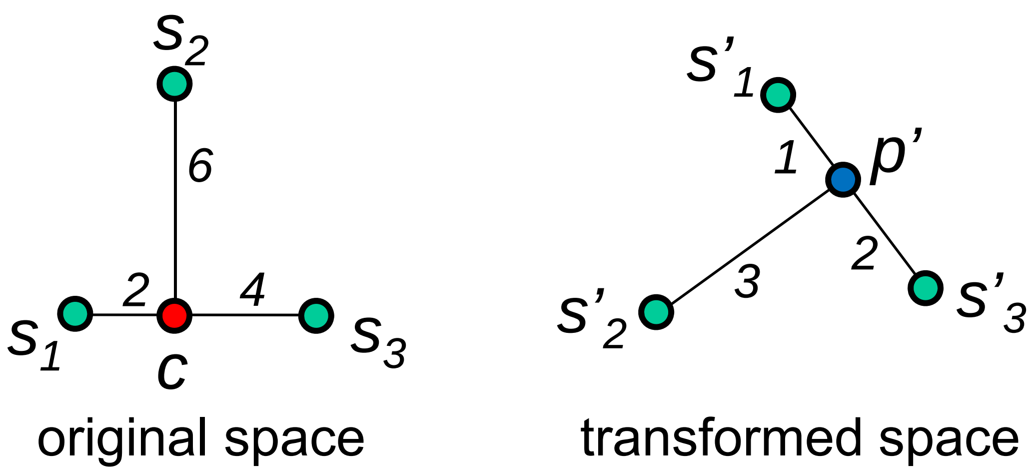
- Objectives of the solution
  - Support *efficient* and *accurate* processing of range queries
  - *Make it hard to reconstruct* the original points in P from the transformed points in P’
- Orthogonal aspects
  - Verifying the correctness of the query results
  - Protecting the identities of the data owner and query users

Attack Models

- What does the attacker know?
  - The set P' of the transformed points
  - Background information: a subset S of points in P and their corresponding points S' in P'
    - But no other points in P
    - Cannot choose an S (S')
- Tailored attack
  - Specific to the *known* transformation method
  - Goal: determine the exact location of each point
  - Formulate a system of equations, solve for the key parameters by using the values in S and S'
- Tailored attacks can be computationally infeasible

Attack models

- General attack
  - Independent of the (unknown) transformation method
  - Goal: estimate a location c, such that the feature vector of c (wrt. S) is the most similar to the feature vector of p’ (wrt. S’)



V(c, S) = <2, 6, 4>

V(p’, S’) = <1, 3, 2>

$$\Phi(c, p') = L_1\left(\frac{V(p', S')}{|V(p', S')|}, \frac{V(c, S)}{|V(c, S)|}\right)$$

Overview of Solutions

Method	*Tailored attack*	*General attack*	*Transferred data cost*	*Round trips*
HSD	2 known points in same partition	High distortion	Low	1
ERB	N/A	Low distortion	High grows with ε	1
HSD*	N/A	High distortion	Moderate	1
CRT	N/A	N/A	Moderate	Tree height

- See papers (listed at the end) for details!

Summary

- Contributions
  - A framework that enables service providers to process range queries without knowing actual data
  - Spatial transformations: HSD, ERB, HSD*
  - Cryptographic transformation: CRT
  - Proposals for tailored and general attacks
- Future work
  - Support other spatial queries, e.g., nearest neighbors, spatial joins

Concluding Remarks

- The contributions to spatial query and data privacy presented here are part of a trend.

Data Management infrastructure for cloud computing

- Many other challenges, e.g., relating to
  - Privacy for historical data
  - Trust
  - Authentication (e.g., “does the server produce ‘correct’ results”?)

References

- M. F. Mokbel, C.-Y. Chow, and W. G. Aref. The New Casper: Query Processing for Location Services without Compromising Privacy. In *VLDB*, 2006.
- P. Indyk and D.Woodruff. Polylogarithmic Private Approximations and Efficient Matching. In *Theory of Cryptography Conference*, 2006.
- A. Khoshgozaran and C. Shahabi. Blind Evaluation of Nearest Neighbor Queries Using Space Transformation to Preserve Location Privacy. In *SSTD*, 2007.
- G. R. Hjaltason and H. Samet. Distance Browsing in Spatial Databases. *TODS*, 24(2):265–318, 1999.
- R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, Order-Preserving Encryption for Numeric Data. In SIGMOD, 2004.
- E. Damiani, S. D. C. Vimercati, S. Jajodia, S. Paraboschi, and P. Samarati, Balancing Confidentiality and Efficiency in Untrusted Relational DBMSs. In Proc. of Computer and Communications Security, 2003.
- H. Hacigümüs, B. R. Iyer, C. Li, and S. Mehrotra, Executing SQL over Encrypted Data in the Database-Service-Provider Model. In SIGMOD, 2002.
- R. Agrawal and R. Srikant, Privacy-Preserving Data Mining. In SIGMOD, 2000.

Readings

- C. S. Jensen: When the Internet Hits the Road. Proc. BTW, pp. 2-16, 2007.
- C. S. Jensen, C. R. Vicente, and R. Wind. User-Generated Content – The case for Mobile Services. IEEE Computer, 41(12):116–118, December 2008.

- http://daisy.aau.dk
- http://streamspin.com

- To obtain permission to use these slides and to obtain copies of papers in submission, send e-mail to csj@cs.aau.dk

- M. L. Yiu, C. S. Jensen, H. Lu. SpaceTwist: Managing the Trade-Offs Among Location Privacy, Query Performance, and Query Accuracy in Mobile Services. Proc. ICDE, April 2008.
- M. L. Yiu, C. S. Jensen, J. Møller, and H. Lu. Design and Analysis of an Incremental Approach to Location Privacy for Location-Based Services. In submission.
- M. L. Yiu, G. Ghinita, C. S. Jensen, and P. Kalnis. Outsourcing Search Services on Private Spatial Data. In Proc. ICDE, 2009, to appear.
- M. L. Yiu, G. Ghinita, C. S. Jensen, and P. Kalnis. Enabling Search Services on Outsourced Private Spatial Data. In submission.

