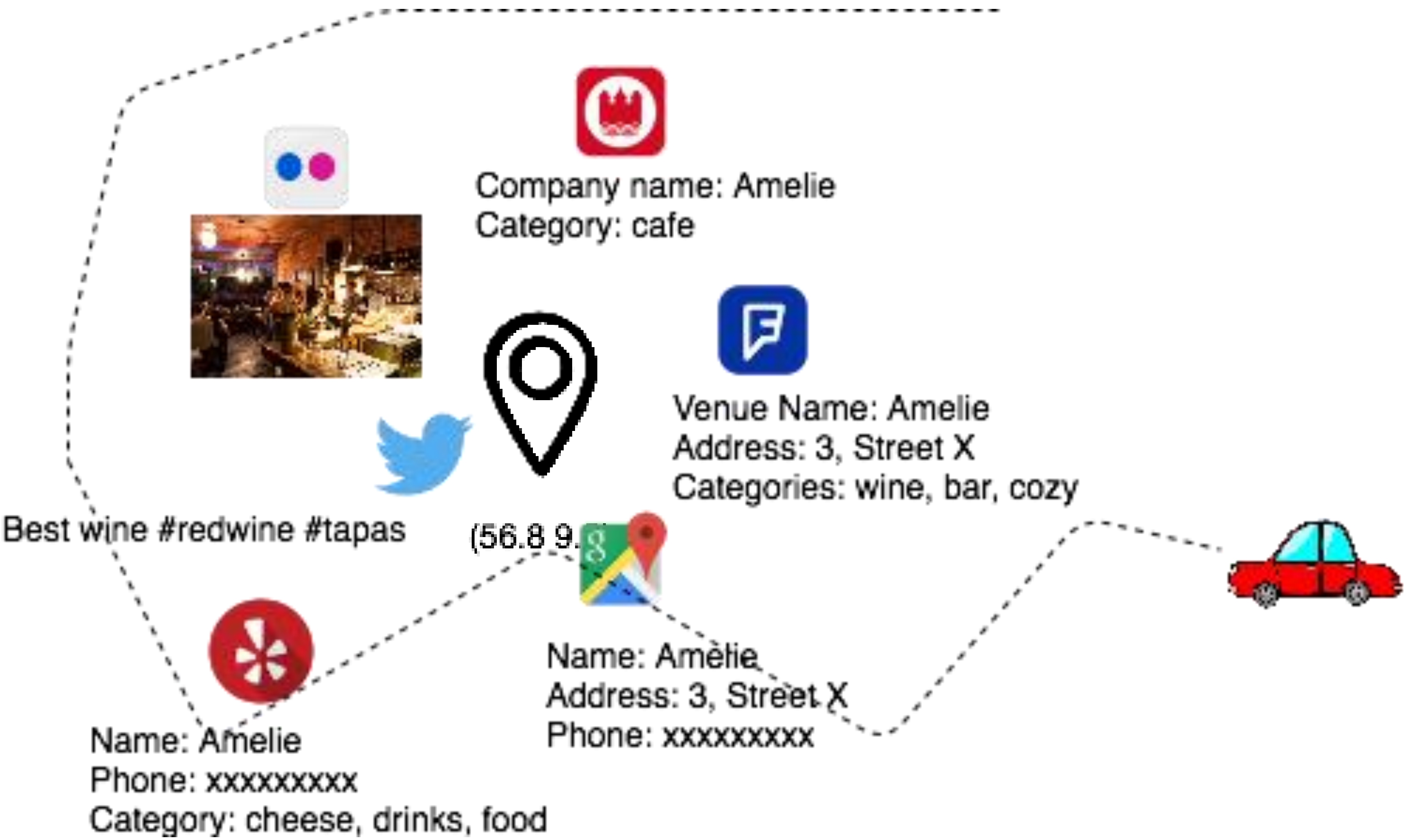
Multi-Source Spatial Entity Linkage

Suela Isaj

Supervisor: Torben Bach Pedersen (AAU)

Co-supervisor: Esteban Zimányi (ULB)

Multi-Source Spatial Entities



Company name: Amelie
Category: cafe

Venue Name: Amelie
Address: 3, Street X
Categories: wine, bar, cozy

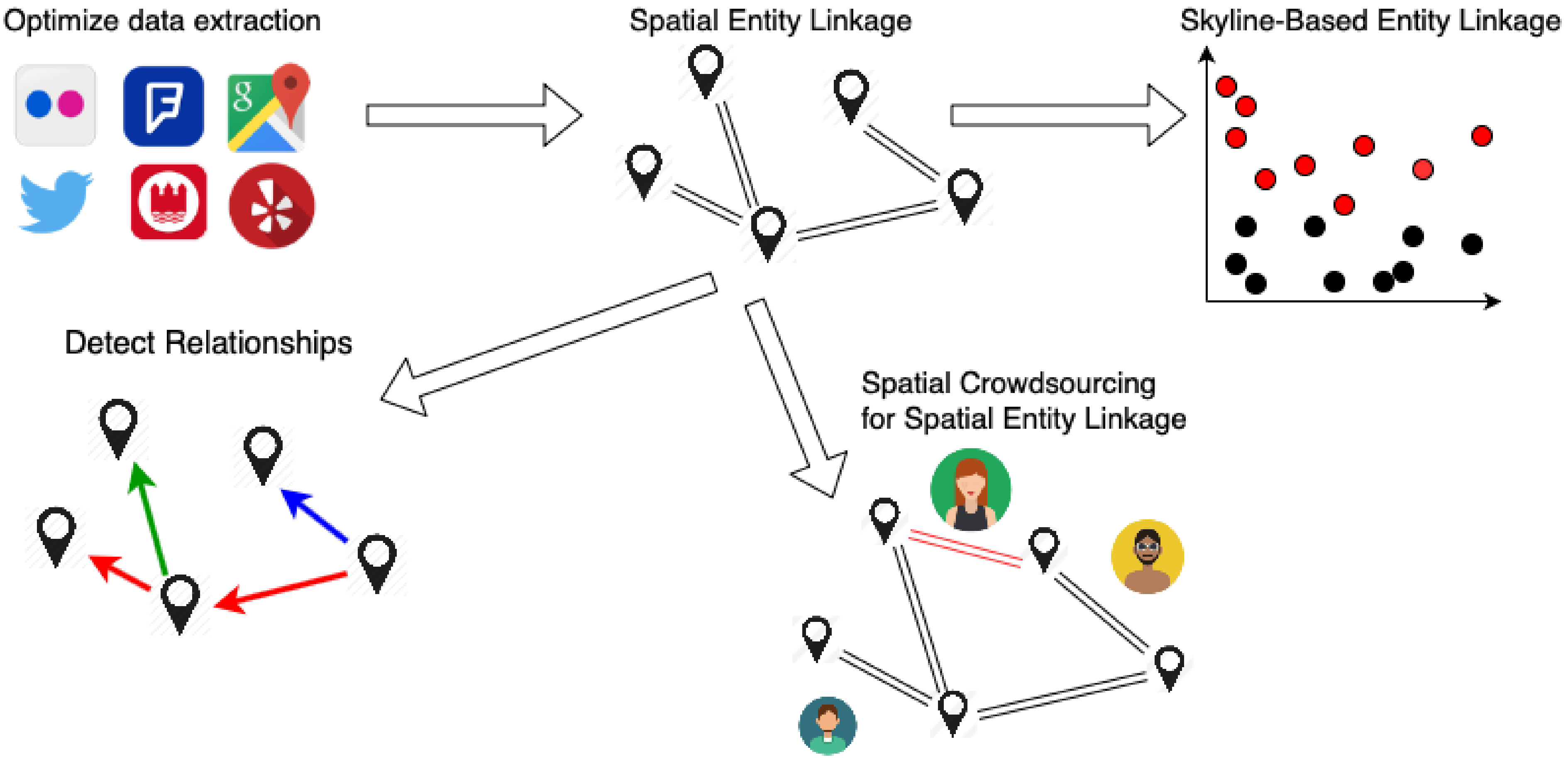
Best wine #redwine #tapas

(56.8 9.

Name: Amèlie
Address: 3, Street X
Phone: xxxxxxxxx

Name: Amelie
Phone: xxxxxxxxx
Category: cheese, drinks, food

Overall PhD study



Optimize data extraction

Spatial Entity Linkage

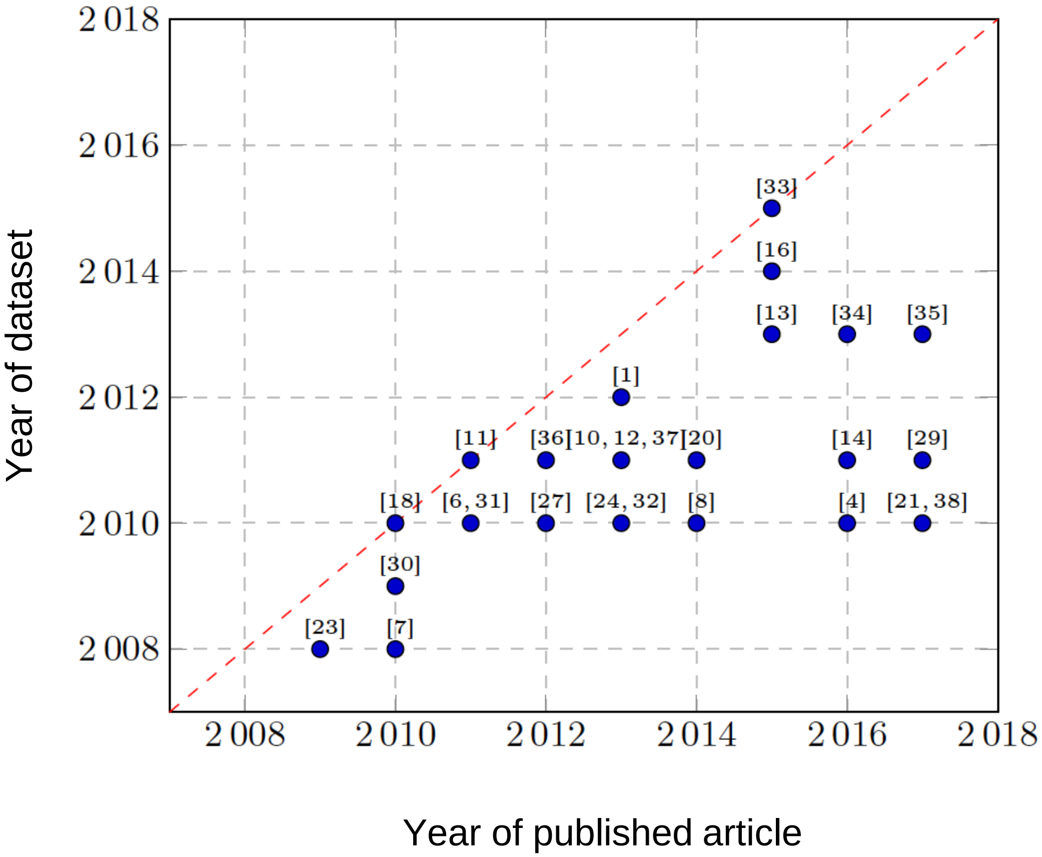
Skyline-Based Entity Linkage

Detect Relationships

Spatial Crowdsourcing for Spatial Entity Linkage

Geo-social related work

- ❑ Old datasets
- ❑ Non-operational social networks
- ❑ Limited locations
- ❑ Missing reference to current systems
- ❑ Simulated user activity instead of real data



API limitations

- Bandwidth

  Number of requests within a time frame
- Result size

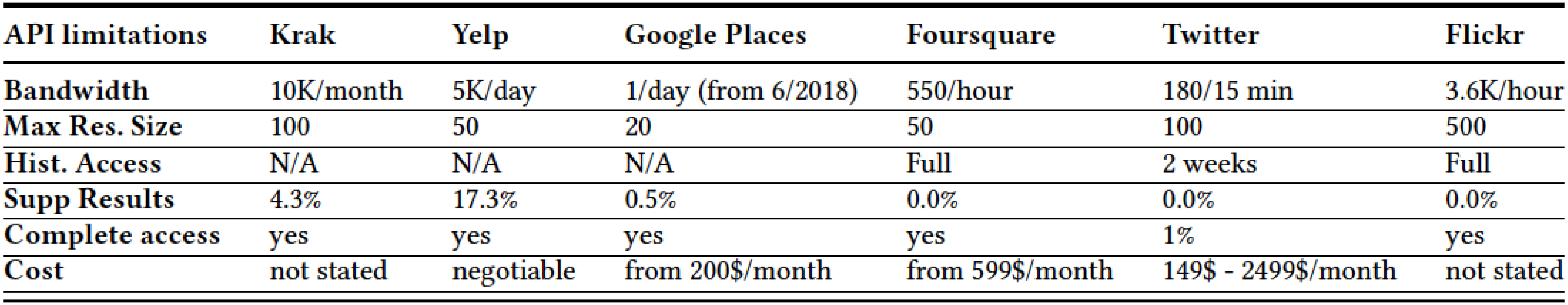
  Number of locations/data for a single request
- Historical access

  Is the API able to retrieve old data?
- Supplemental results

  Does the API give data outside $\mathit{Circle}\ (p, r)$*?*
- Costs

  Premium services / Pay as you go
- Access to the complete dataset

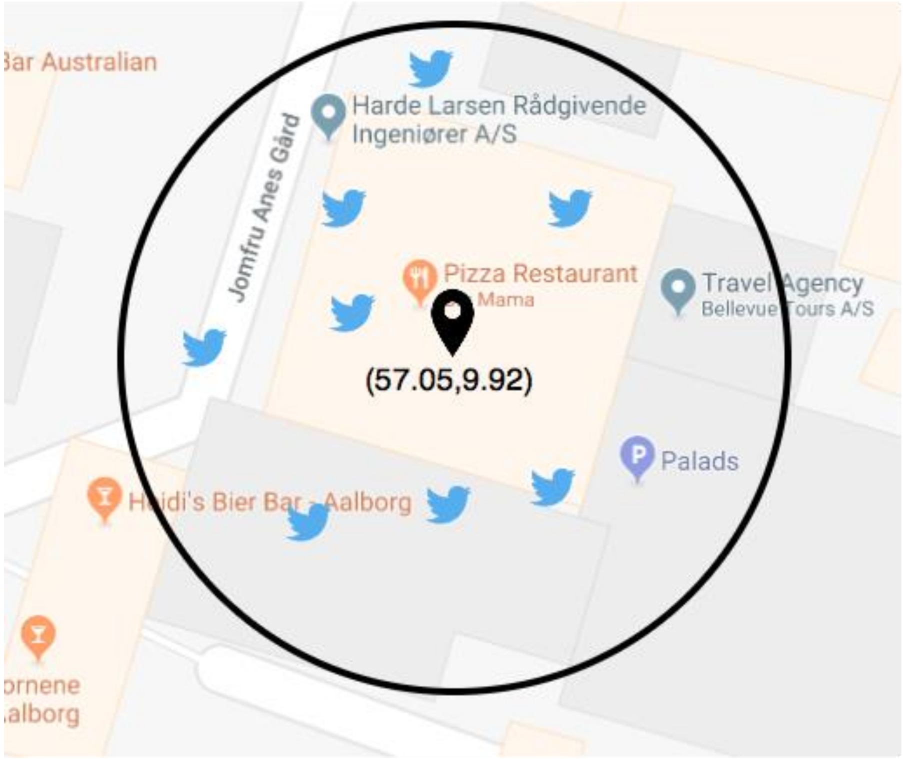
  Sample vs whole access

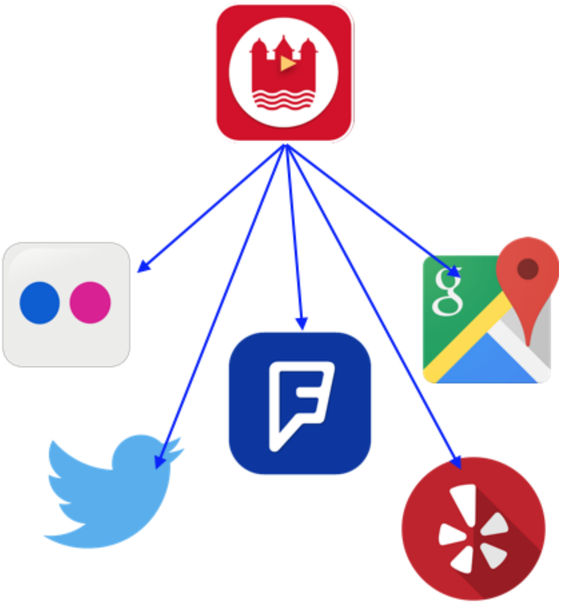


API limitations	Krak	Yelp	Google Places	Foursquare	Twitter	Flickr
Bandwidth	10K/month	5K/day	1/day (from 6/2018)	550/hour	180/15 min	3.6K/hour
Max Res. Size	100	50	20	50	100	500
Hist. Access	N/A	N/A	N/A	Full	2 weeks	Full
Supp Results	4.3%	17.3%	0.5%	0.0%	0.0%	0.0%
Complete access	yes	yes	yes	yes	1%	yes
Cost	not stated	negotiable	from 200$/month	from 599$/month	149$ - 2499$/month	not stated

Data extraction

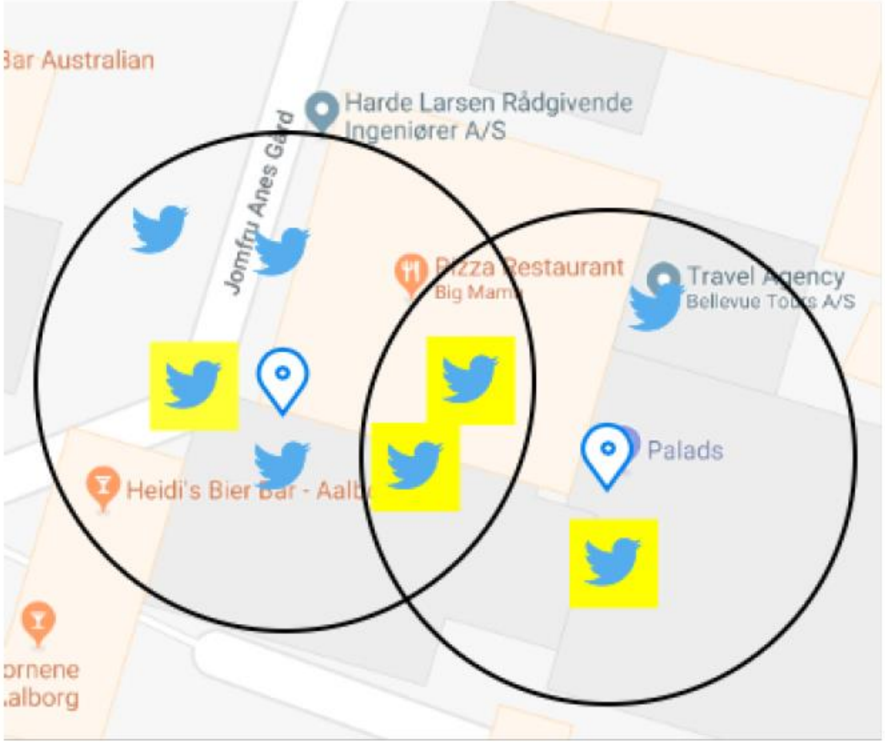
- Location-based queries - $API\ call\ (p, r)$
- Well-selected points
- Use the points of one source (seed) to query the others

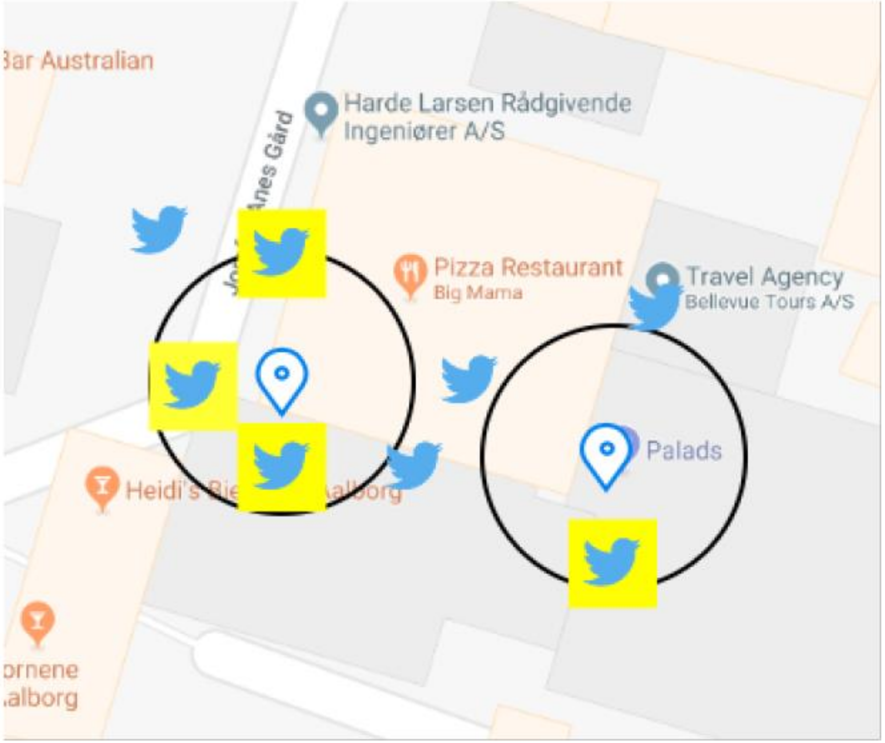




Radius selection

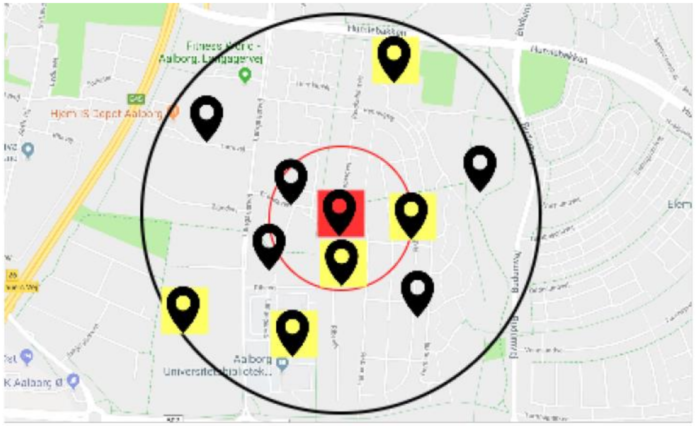
Limited by maximal result size!



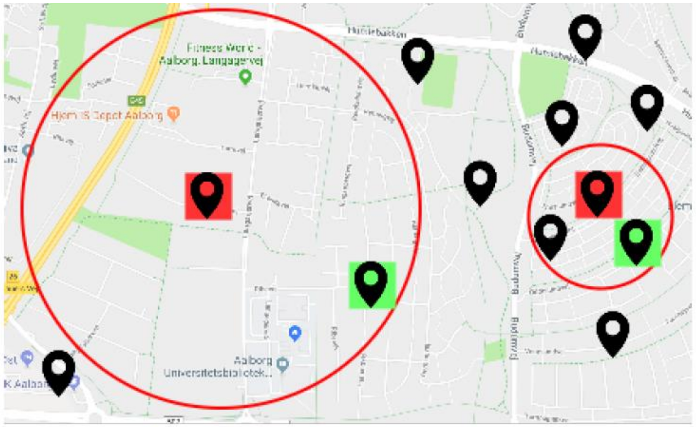


Multi-Source Seed-Driven Algorithms

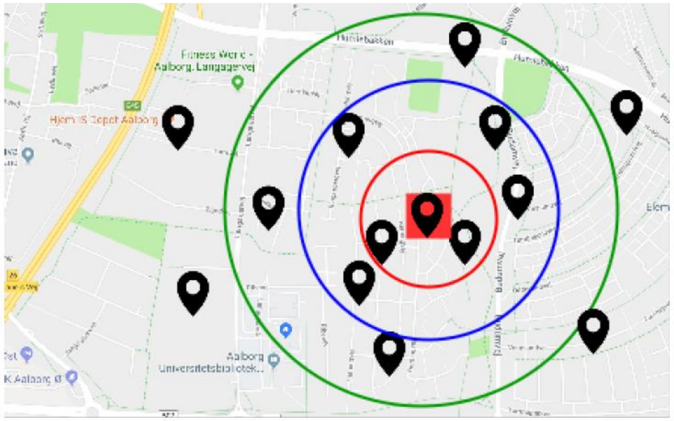
- $MSSD - F$ – Fixed 2 km
- $MSSD - D$ – Seed density-based
- $MSSD - N$ – Seed nearest neighbor
- $MSSD - R$ – Recursively adapted to the source


Fitness World - Aalborg, Langagervej
Hjem IS Depot Aalborg
Aalborg Universitetsbibliotek

(a) *MSSD-D* radius


Fitness World - Aalborg, Langagervej
Hjem IS Depot Aalborg
Aalborg Universitetsbibliotek

(b) *MSSD-N* radius

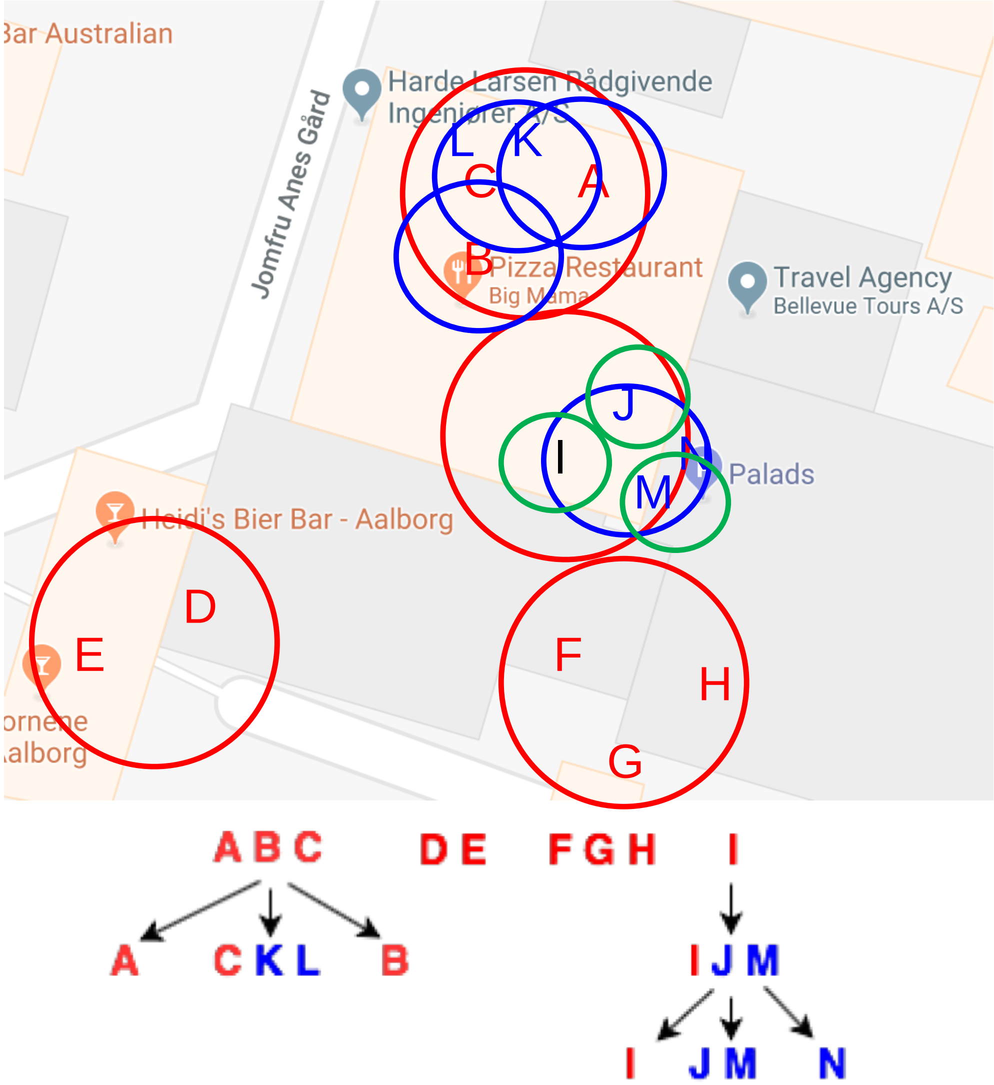

Fitness World - Aalborg, Langagervej
Hjem IS Depot Aalborg
Aalborg Universitetsbibliotek

(c) *MSSD-R* radius

MSSD*

- Red – seed locations
- Blue – source locations

- Cluster points with DBSCAN
- Query with the centroid
- If the maximal result size is reached, split the cluster and query with smaller radius

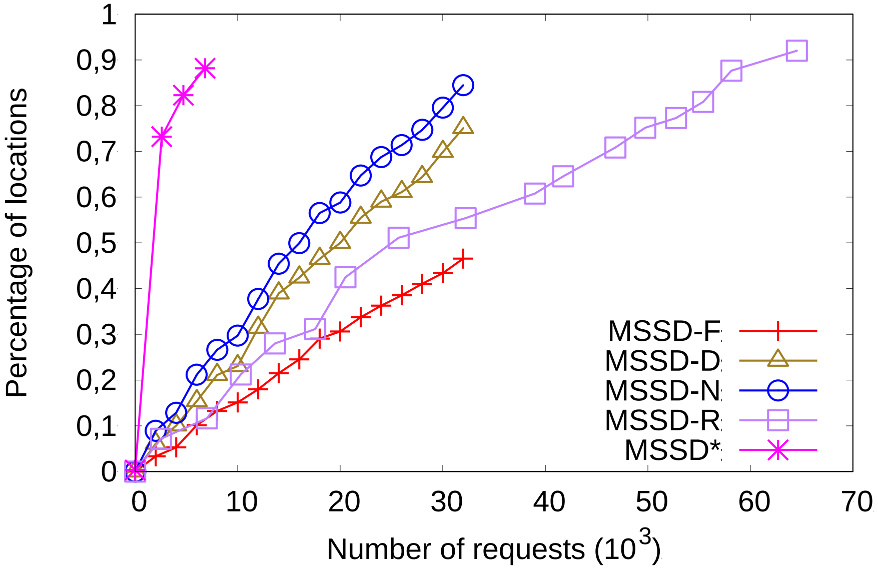


Experiments

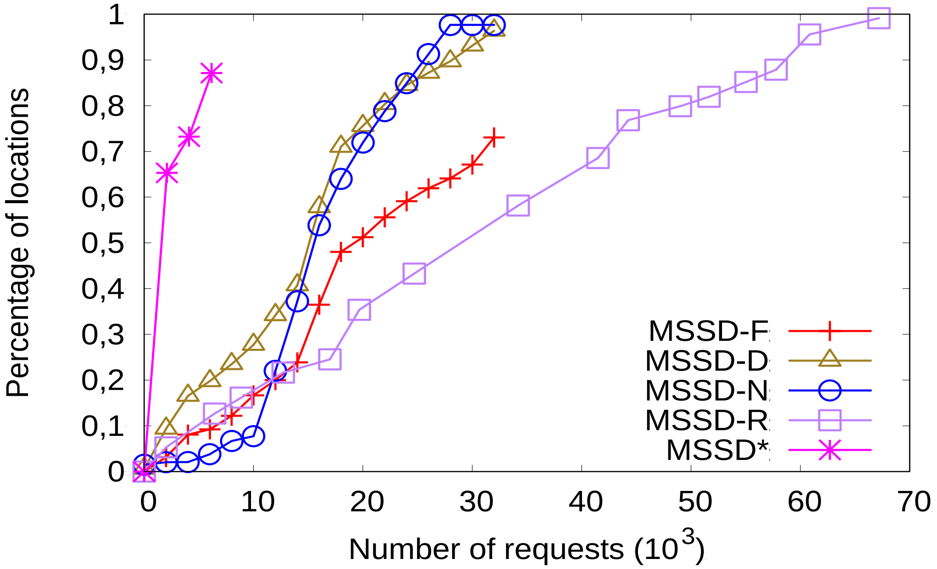
- Requests versus number of locations
- $MSSD - N$ - the best from the fixed request versions
- $MSSD - R$ - the best for number of locations but expensive

$\boldsymbol{MSSD} *$

- 90% of the locations of $MSSD - R$
- with 25% of the requests of $MSSD - F, MSSD - D, MSSD - N$
- 12%-15% of $MSSD - R$ requests for Flickr, Yelp and Foursquare, 8.5% for Google Places and 2.7% for Twitter.



(a) Flickr

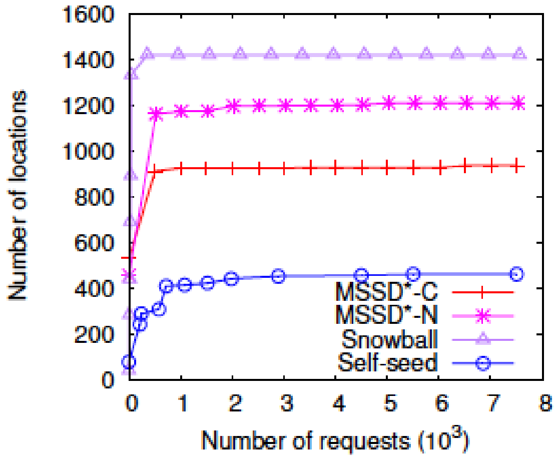


(b) Foursquare

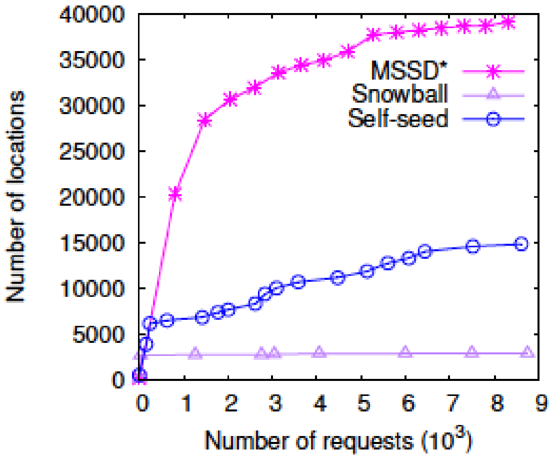
Comparison to other methods

- Snowball (Scellato et al in WOSN’10, Gao et al in AAAI’15)
  - Only applicable to social networks, not directories
  - Proved to be biased
  - Does not guarantee that the activity is within the searched area
- Linked accounts (Armenatzoglou et al in PVLDB’13, Preotiuc-Pietro et al in WebSci’13, Hristova et al in WWW’16)
  - Only applicable to social networks, not directories
  - Does not guarantee that the activity is within the searched area
  - Rare to find:
    - 0.27 % of users in Flickr with linked accounts to Twitter
    - 0.003 % of users in Twitter with linked accounts to Foursquare.
- Self-seed (Lee at al in GIS-LBSN’10)
  - Similar to ours
  - Limited within a social network

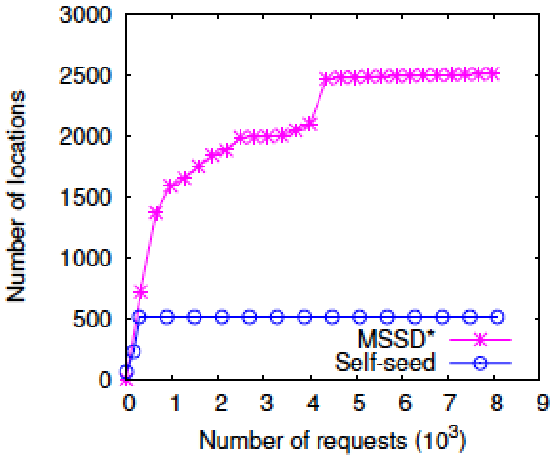
Comparison to other approaches



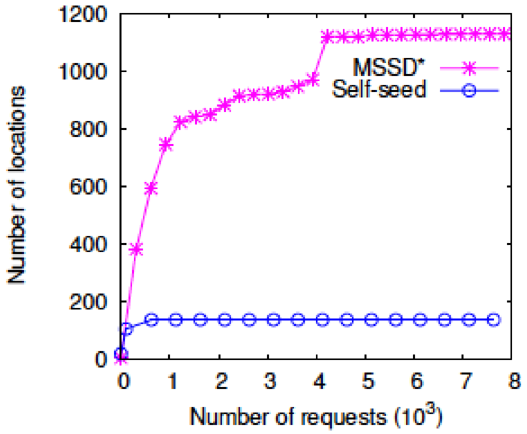
(a) Twitter



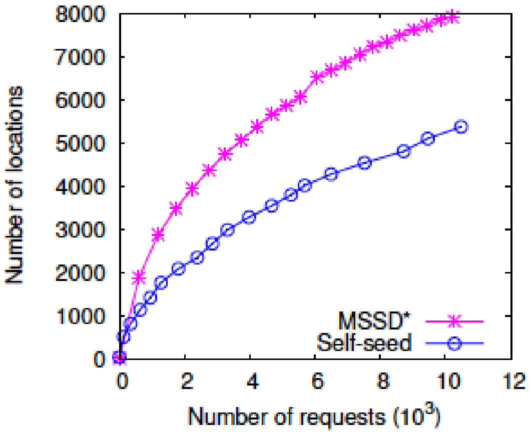
(b) Flickr



(c) Foursquare

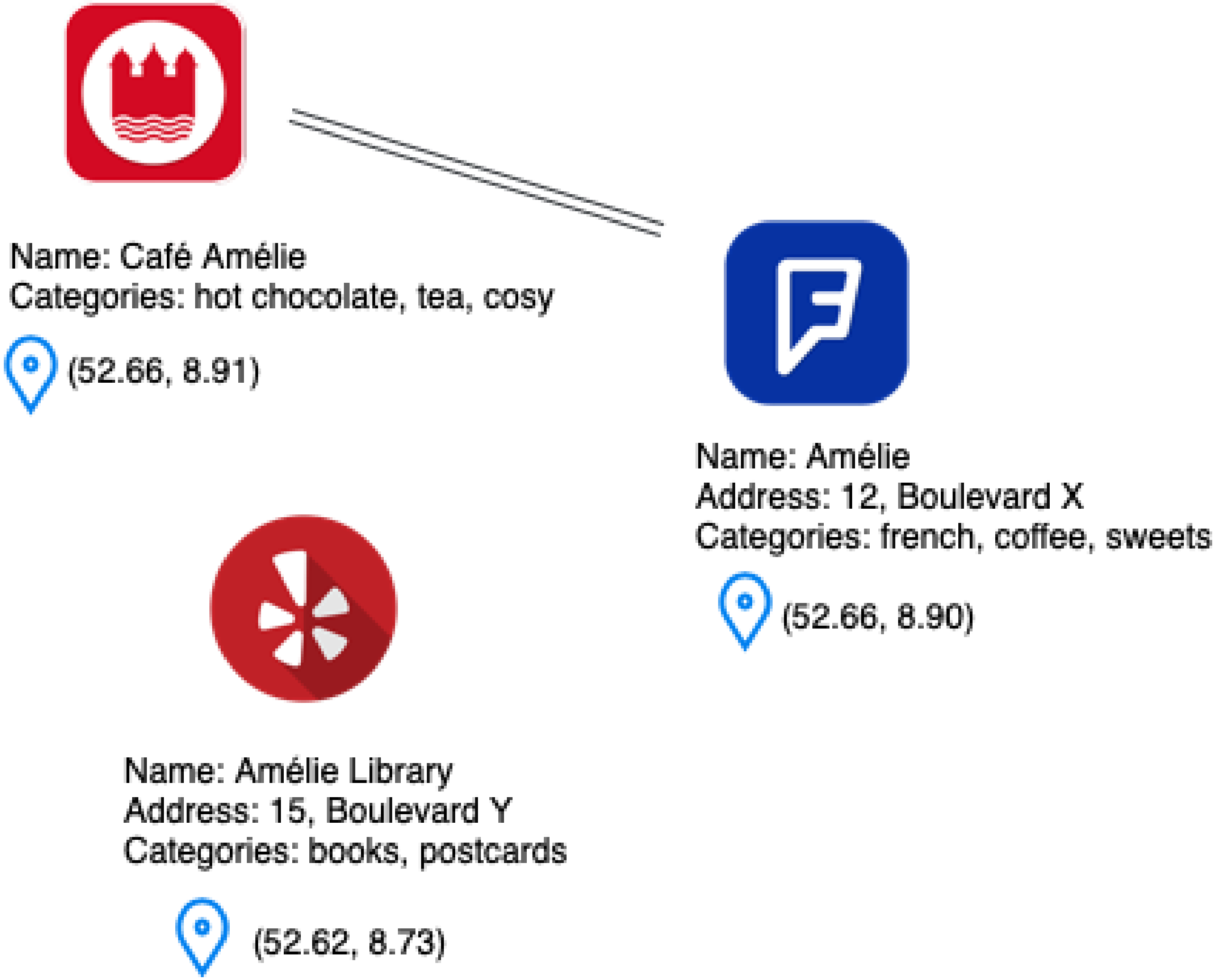


(d) Yelp



(e) Google Places

Spatial Entity Linkage



Name: Café Amélie
Categories: hot chocolate, tea, cosy

(52.66, 8.91)

Name: Amélie
Address: 12, Boulevard X
Categories: french, coffee, sweets

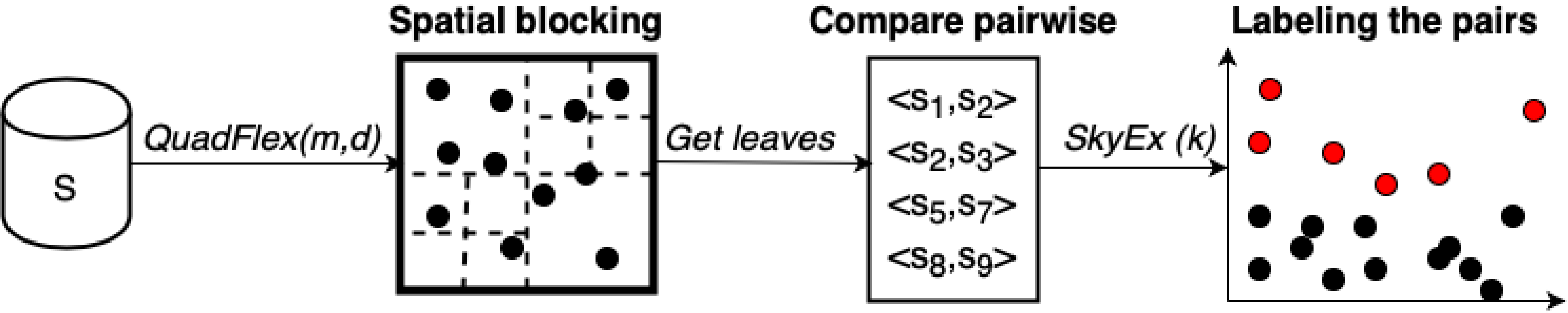
(52.66, 8.90)

Name: Amélie Library
Address: 15, Boulevard Y
Categories: books, postcards

(52.62, 8.73)

QuadSky solution

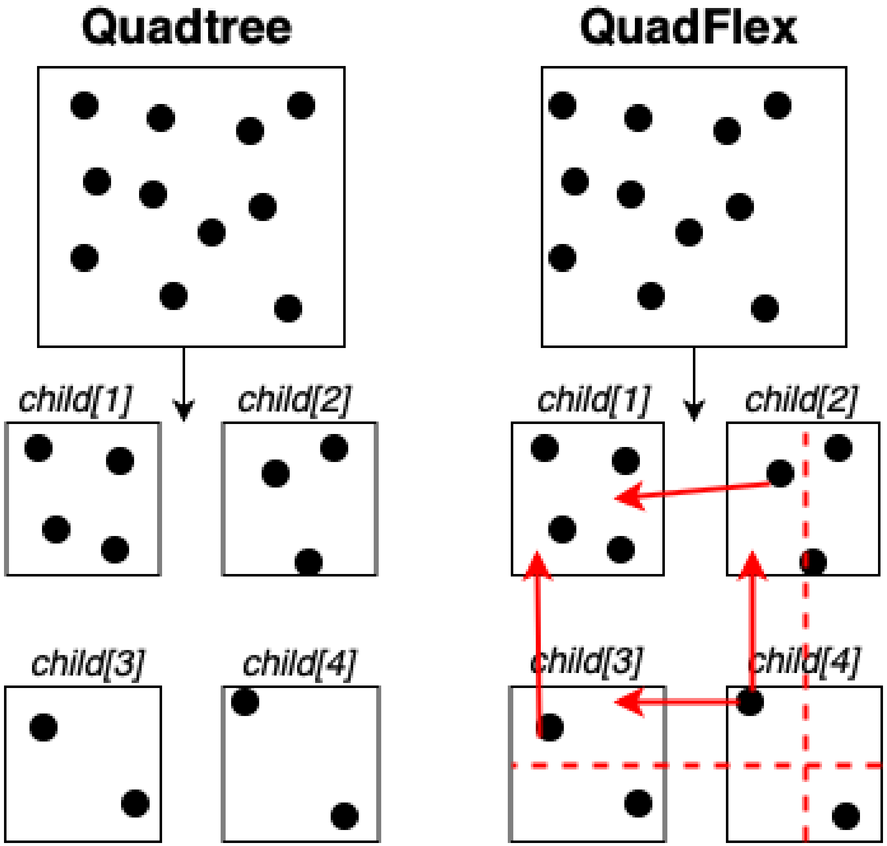
- Spatial Blocking (QuadFlex) + Labelling the pairs (SkyEx)
- Input: A set of spatial entities
- Output: Labelled pairs (Yes/No)



Spatial Blocking

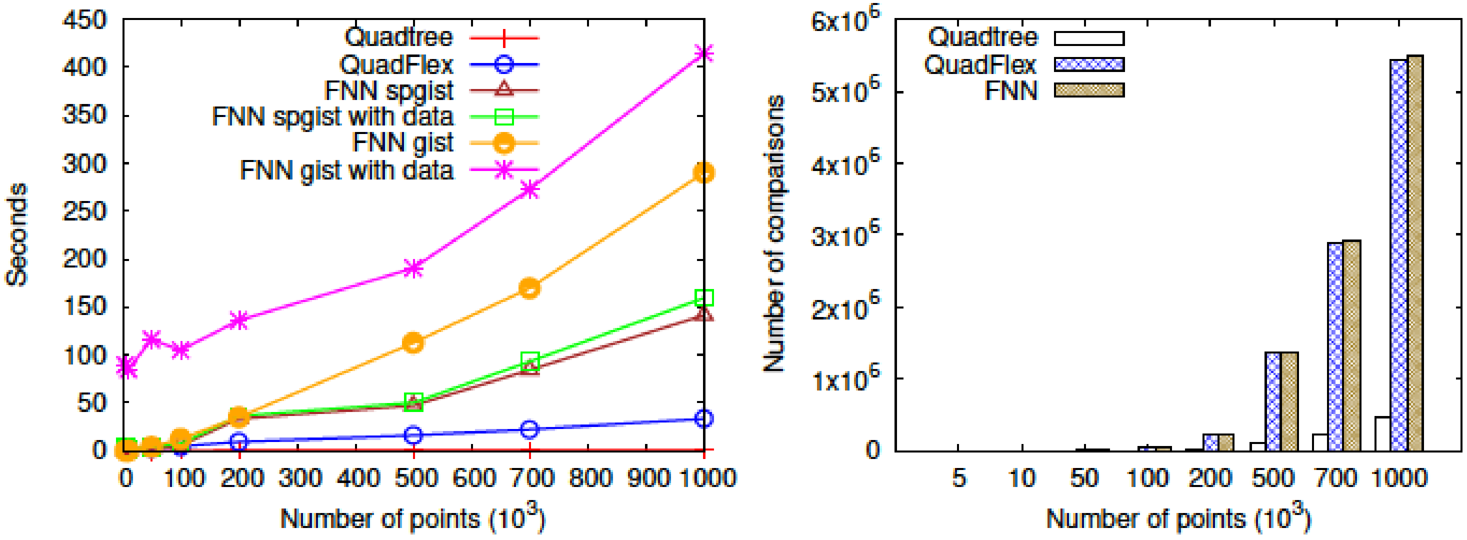
- Avoid exhaustive comparisons

- QuadFlex solution
  - Diagonal and Density instead of Capacity
  - Allow point assignment in multiple children



Spatial Blocking (QuadFlex)

- Runtime of QuadTree, Comparisons as FNN
- GiST and SP-GiST(postgres)
- QuadFlex has 99.99% of the comparisons of FNN, Quadtree only 10%



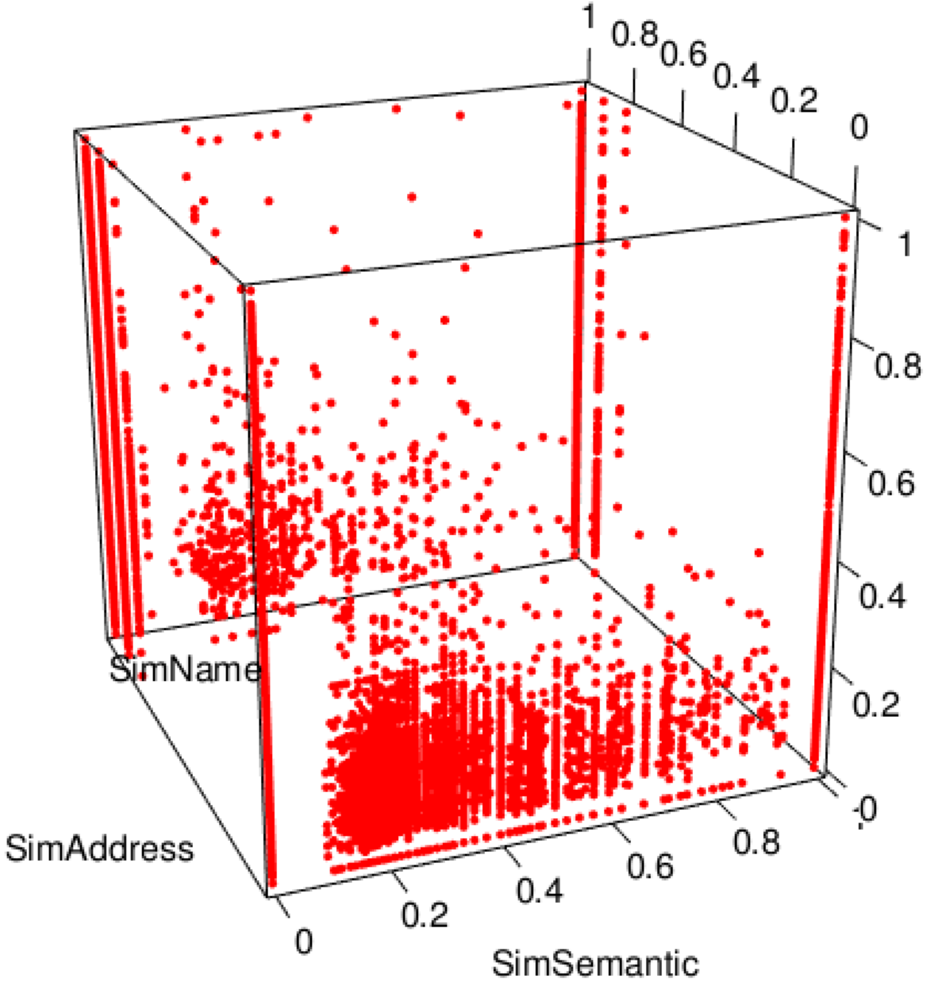
(a) Execution time

(b) Number of comparisons

Comparing quadtree, QuadFlex and FNN

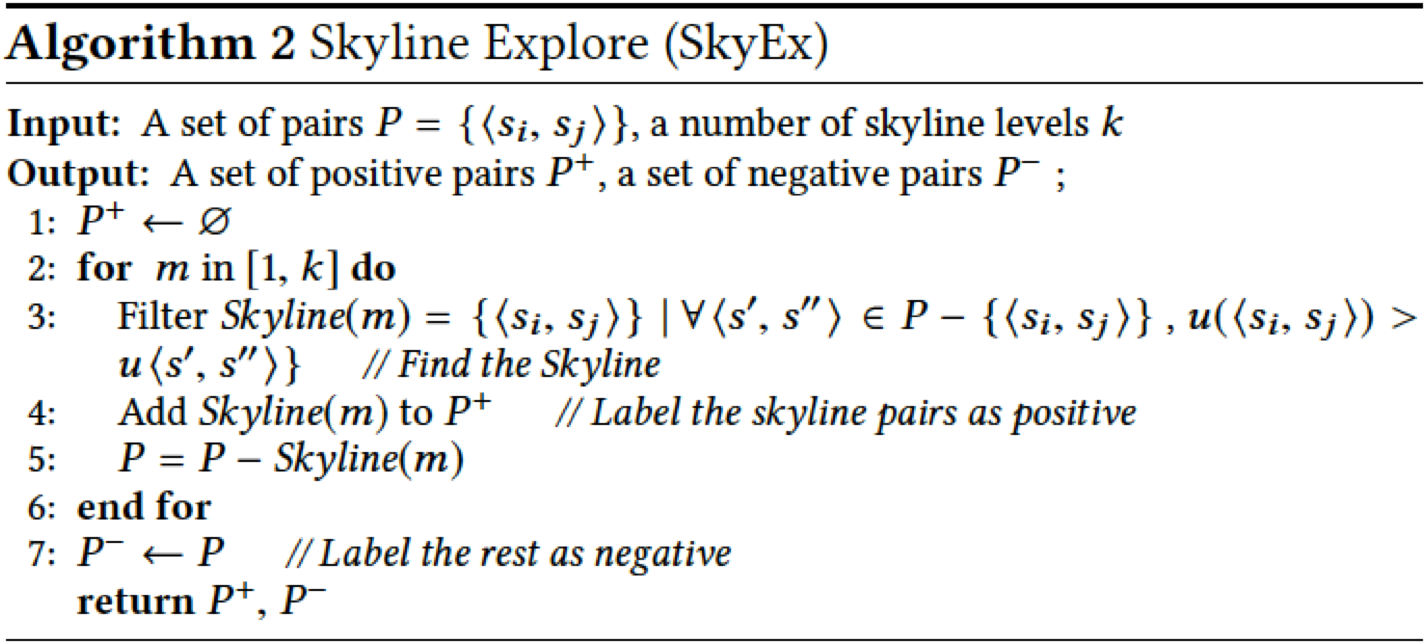
Pairwise Comparison

- Comparing the attributes
- Name: Levenshtein
- Address: Custom
- Categories: Wu&Palmer Wordnet



SkyEx (Skyline Explore)

- No training set, no overfitting, no extensive experiments
- Pareto Optimality – abstraction of a similarity function (utility)
- The best candidates are in the first skylines



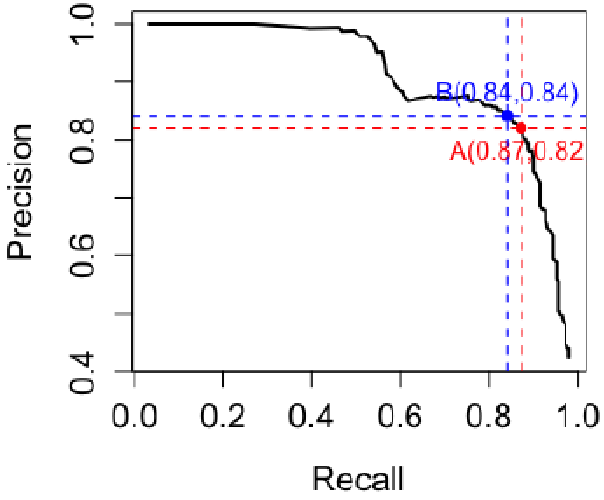
Algorithm 2 Skyline Explore (SkyEx)

```
Input: A set of pairs $P = \{\langle s_i, s_j \rangle\}$, a number of skyline levels $k$
Output: A set of positive pairs $P^+$, a set of negative pairs $P^-$ ;
$P^+ \leftarrow \varnothing$
for $m$ in $[1, k]$ do
    Filter $Skyline(m) = \{\langle s_i, s_j \rangle\} \mid \forall \langle s', s'' \rangle \in P - \{\langle s_i, s_j \rangle\}, u(\langle s_i, s_j \rangle) > u\langle s', s'' \rangle\}$    // Find the Skyline
    Add $Skyline(m)$ to $P^+$    // Label the skyline pairs as positive
    $P = P - Skyline(m)$
end for
$P^- \leftarrow P$    // Label the rest as negative
return $P^+, P^-$
```

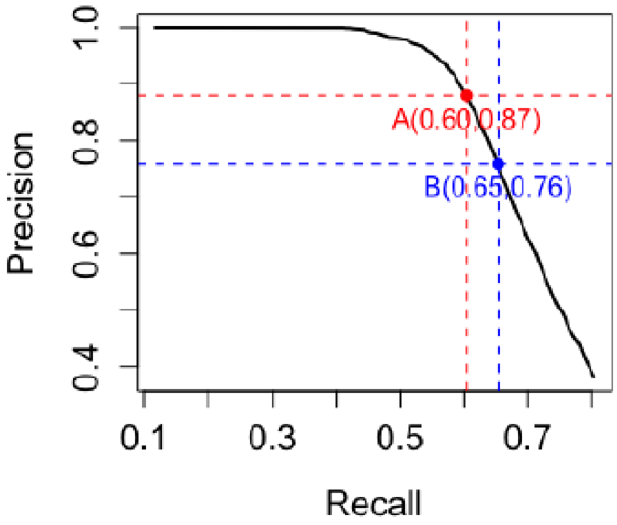
SkyEx results

- Precision / Recall/ F-measure
- Automatic labeling (Phone or Website) – 777,452 pairs
  - F-measure = 0.72
- Manual labeling – 1,500 pairs
  - F-measure = 0.85

Sample –manual labeling

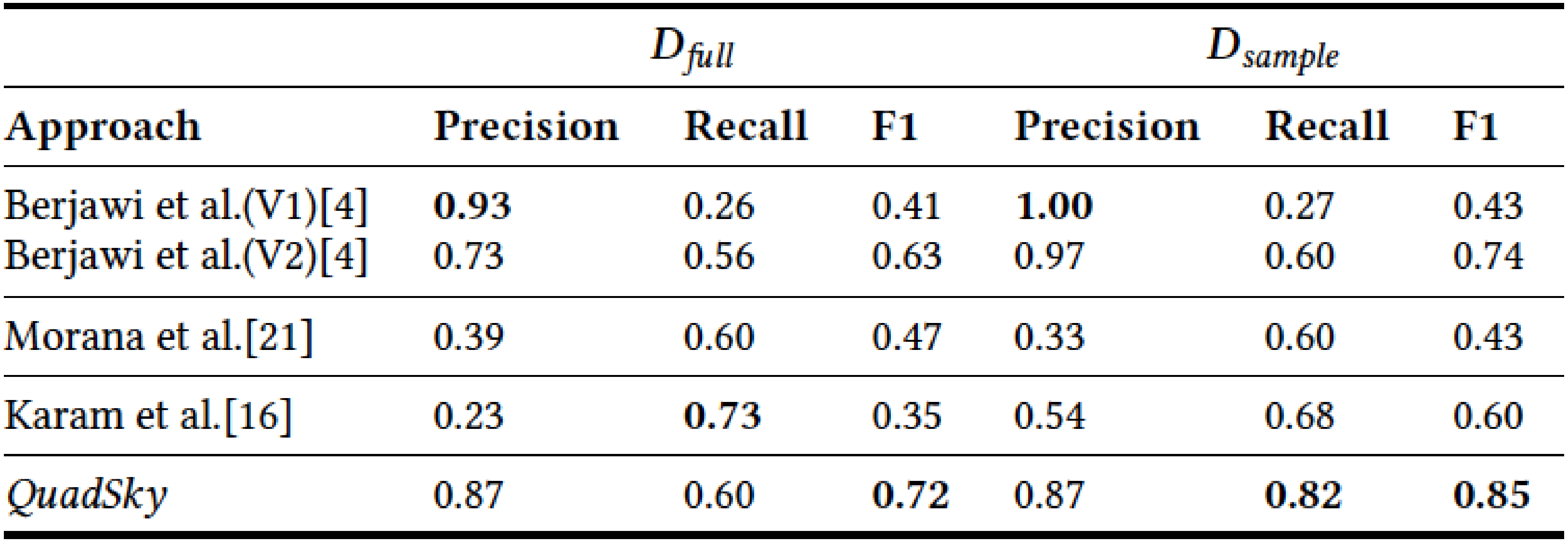


Whole dataset –automatic labeling



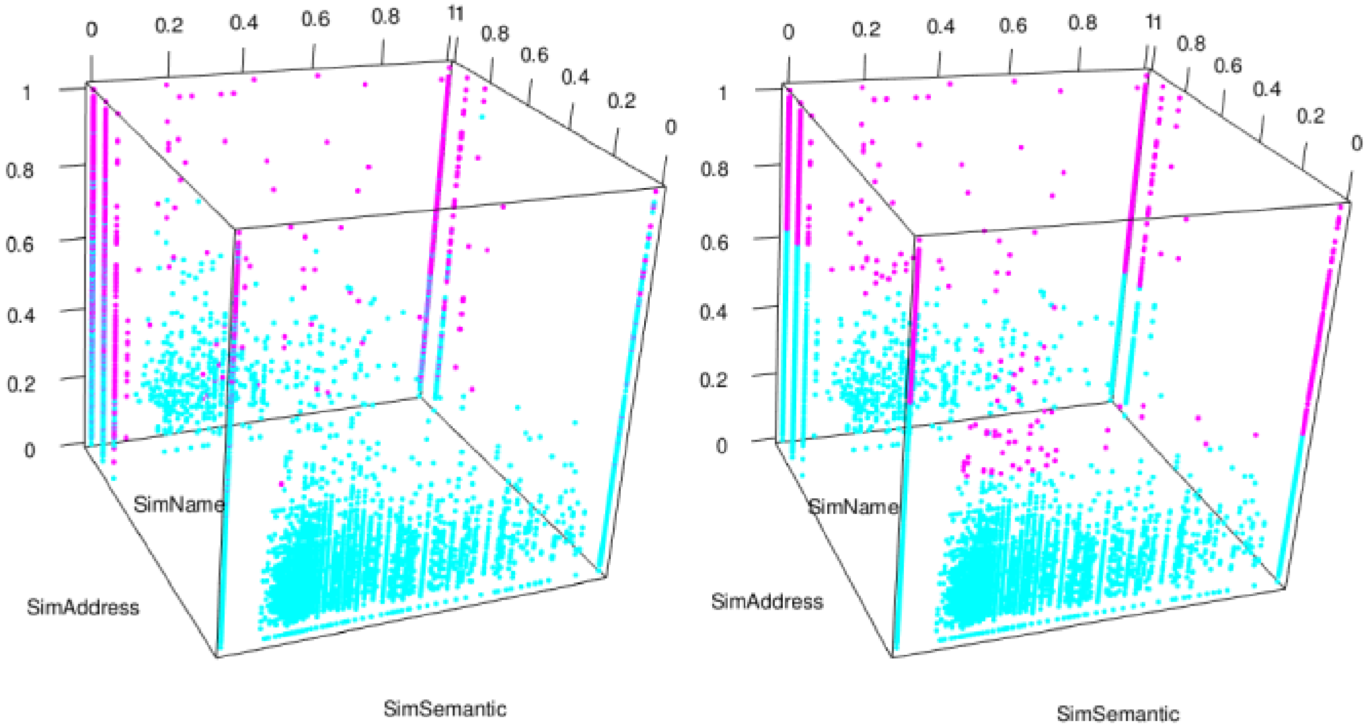
Comparison to other approaches

- Berjawi et al. – 50 m apart
  - Euclidean for geo, Levenshtein for name & address
  - Name + address + geo (V1)
  - Name + geo (V2)
- Morana et al – blocks of same category or name
  - Euclidean for geo, Levenshtein for address and name, Resnik (Wordnet) for categories
  - 2/3 (name + geo + categories) + 1/3 address
- Karam et al – 5m apart
  - Levenshtein for name, Euclidean for geo, Keywords semantically
  - Belief theory



	D_{full}			D_{sample}		
Approach	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Berjawi et al.(V1)[4]	**0.93**	0.26	0.41	**1.00**	0.27	0.43
Berjawi et al.(V2)[4]	0.73	0.56	0.63	0.97	0.60	0.74
Morana et al.[21]	0.39	0.60	0.47	0.33	0.60	0.43
Karam et al.[16]	0.23	**0.73**	0.35	0.54	0.68	0.60
QuadSky	0.87	0.60	**0.72**	0.87	**0.82**	**0.85**

SkyEx labeling



(a) Actual classes

(b) SkyEx classes

Next steps

- Data extraction
  - *“Seed-Driven Geo-Social Data Extraction” S.Isaj, T.B. Perdersen*– Accepted in SSTD 2019
- Spatial entity linkage
  - *"Multi-Source Spatial Entity Linkage” S.Isaj, E. Zimanyi, T.B. Perdersen* – Accepted in SSTD 2019
  - *”Spatial Entity Linkage with the aid of Spatial Crowdsourcing” S.Gummidi, S.Isaj, T.B. Perdersen, E. Zimanyi* – Expected submission in WWW, November 2019
  - *“Discovering relationships between multi-source spatial entities” – Expected submission VLDB-J or Geoinformatica (February 2020)*
- Skyline-based approach
  - *"Skyline-based approach for Entity Resolution”* - Expected submission ICDE, October 2019
  - *”SkyEx – Skyline Exploration for Classifying Pairs”- Demo paper (R package) Expected Submission CIKM (May 2020)*

Work and Time plans

- **Teaching hours (completed 700 hours):**
  - **Fall 2017**
    - 294 group supervision of 2 SW3 + 1 DAT5 + censoring in Web Intelligence course
    - 50 hours as Social Media Manager of Daisy group
  - **Spring 2018**
    - 205 group supervision of 2 BAIT4 + 1 ITVEST master project
    - 50 hours as Social Media Manager of Daisy group
  - **Fall 2018**
    - 50 hours as Social Media Manager of Daisy group
  - **Spring 2019**
    - 50 hours as Social Media Manager of Daisy group
  - 50 hours left – Social Media Manager of Daisy group

- **ECTS (completed 30,25 ECTS)**
  - 14,25 ECTS on General Courses and 16 ECTS on Project courses = 23,75 ECTS
  - Conference presentations

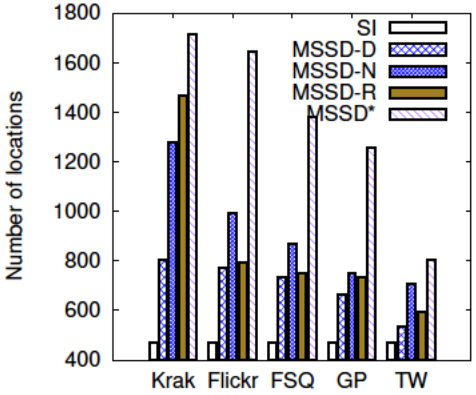
Thank you

Next steps

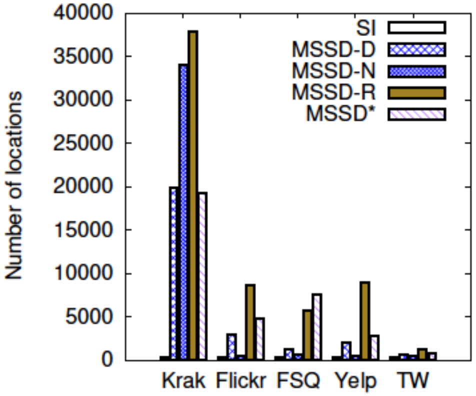
- Data extraction
  - *“Seed-Driven Geo-Social Data Extraction” S.Isaj, T.B. Perdersen*– Accepted in SSTD 2019
- Spatial entity linkage
  - *"Multi-Source Spatial Entity Linkage” S.Isaj, E. Zimanyi, T.B. Perdersen* – Accepted in SSTD 2019
  - *”Spatial Entity Linkage with the aid of Spatial Crowdsourcing” S.Gummidi, S.Isaj, T.B. Perdersen, E. Zimanyi* – Expected submission in WWW, November 2019
  - *“Discovering relationships between multi-source spatial entities” – Expected submission VLDB-J or Geoinformatica (February 2020)*
- Skyline-based approach
  - *"Skyline-based approach for Entity Resolution”* - Expected submission ICDE, October 2019
  - *”SkyEx – Skyline Exploration for Classifying Pairs”- Demo paper (R package) Expected Submission CIKM (May 2020)*

Multi-Seed

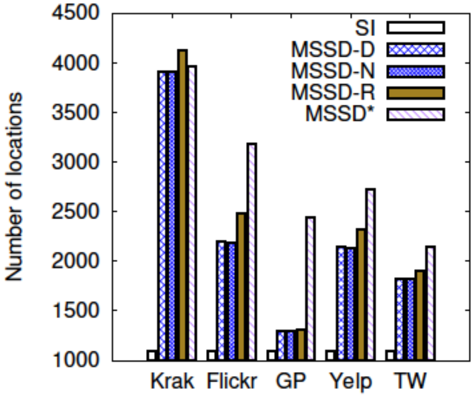
- Krak performs the best for Flickr, Yelp, and Foursquare.
- MSSD* sometimes performs better than MSSD-R



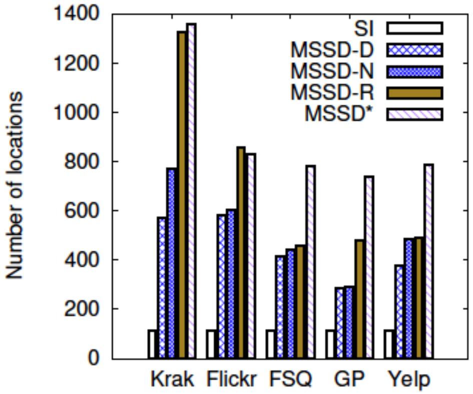
(a) Yelp



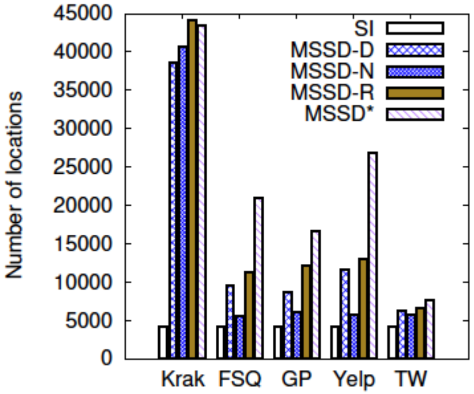
(b) GooglePlaces



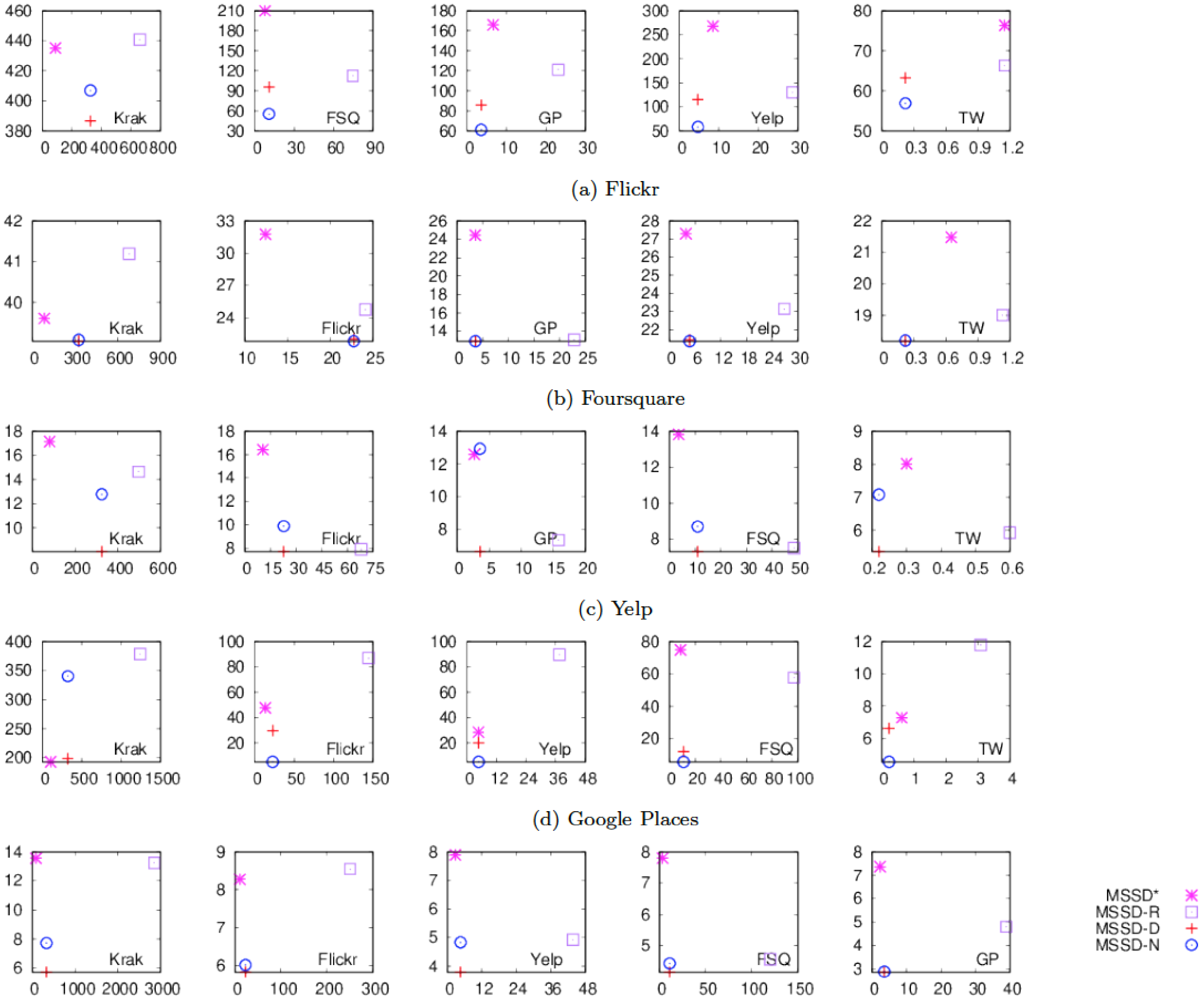
(c) Foursquare



(d) Twitter



(e) Flickr



(a) Flickr

(b) Foursquare

(c) Yelp

(d) Google Places

(e) Twitter

Keyword-based querying

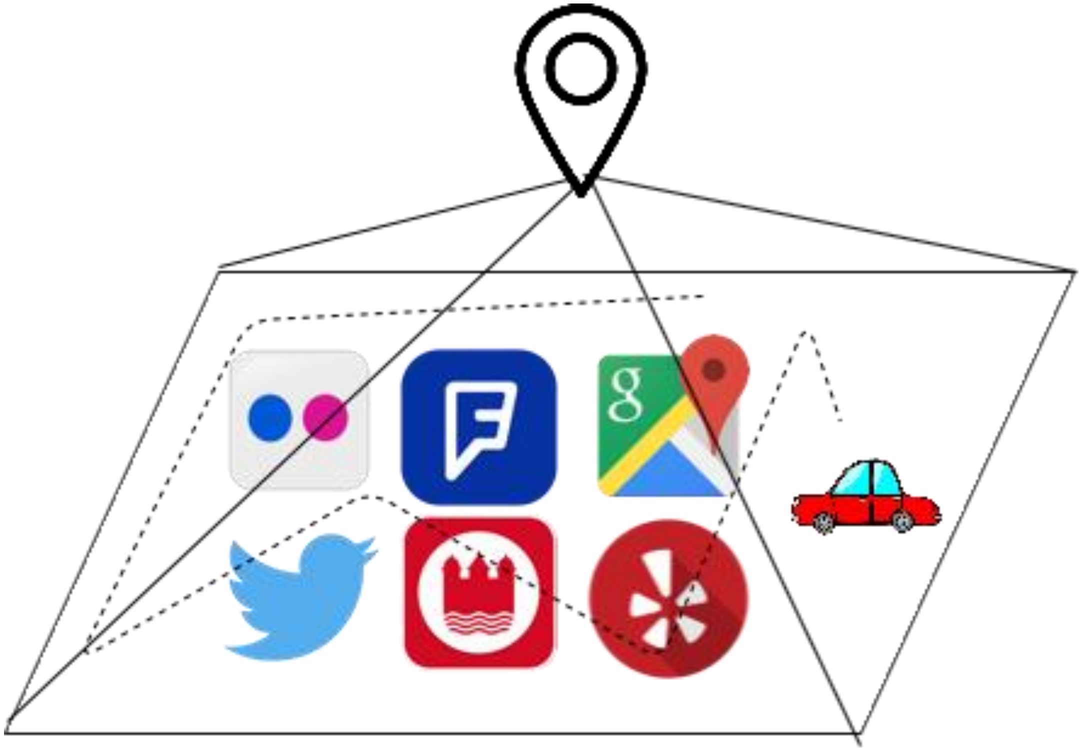
- Query with “Brussels” and getting “brussels sprouts”

- Names of cities and towns in North Denmark as keywords
- Flickr - precision 31.6% recall 5%
- Twitter - precision 0.85% recall 3%
- Foursquare – query by location: precision 93% recall 17%
- Yelp – query by location: precision 85% recall 19%
- Google Places – precision 100% recall 0.07%

Multi-Source Heterogeneous Locations

- Various scopes -> more locations (all)
- Richer context behind locations (directories)
- Crowd-sourced context (social networks)

- Maps / Yellow pages
- User preferences
- Influential locations



daisy