Multi-Source Spatial Entity Linkage

Suela Isaj[1], Torben Bach Pedersen[1], Esteban Zimányi[2]
[1] Aalborg University, Denmark [2] Université libre de Bruxelles, Belgium

Optimizing Data Extraction

Geo-social data extraction problem: Maximizing the extracted data from social networks while minimizing the performed API requests.

API limitations: Bandwidth, maximal result size, cost, historical access, supplemental results, limited access to the full dataset, etc.

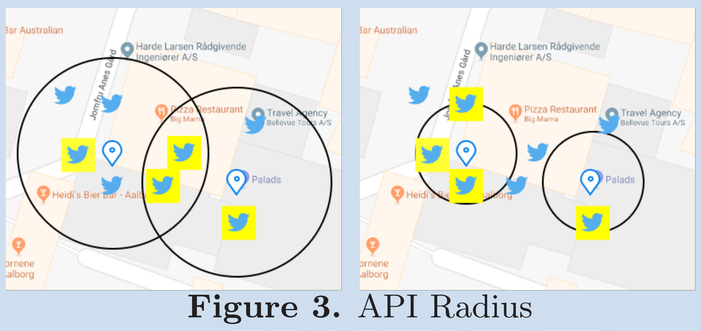

Figure 3. API Radius

Versions of Multi-Source Seed Driven Algorithm *MSSD* :

- *SI* : default API settings
- *MSSD-F* : p of *seed* and a fixed r
- *MSSD-D* : p of *seed* and $r_d = \frac{r}{N}$ $(N = \{q | q \in Circle(p, r)\})$
- *MSSD-N* : p *seed* and a $r_n = |p - q|$ (q - nearest neighbor of p)
- *MSSD-R* : p *seed* and adapt r recursively
- *MSSD** : Split DBSCAN clusters of *seed* points and adapt r recursively depending on the source (center changes)

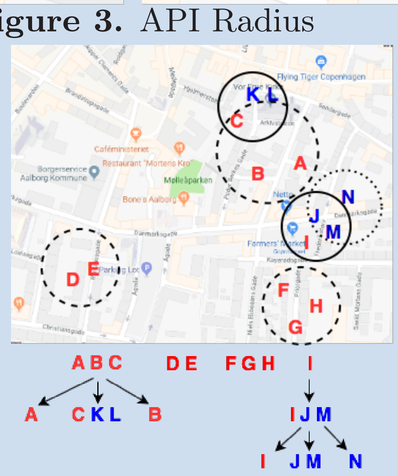

Figure 4. *MSSD**

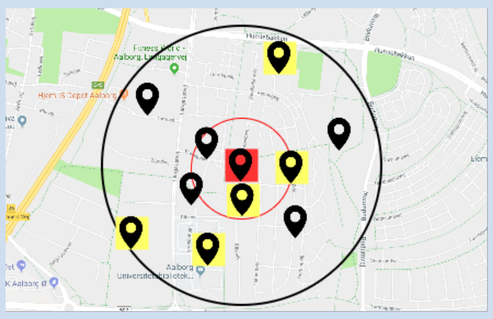

Figure 5(a). *MSSD-D*

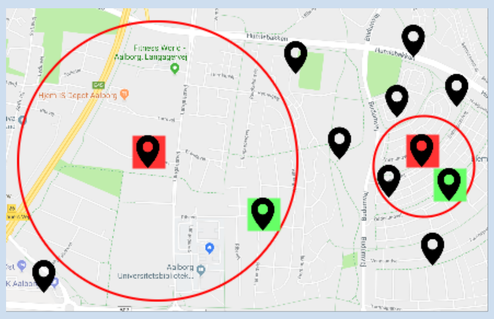

Figure 5(b). *MSSD-N*

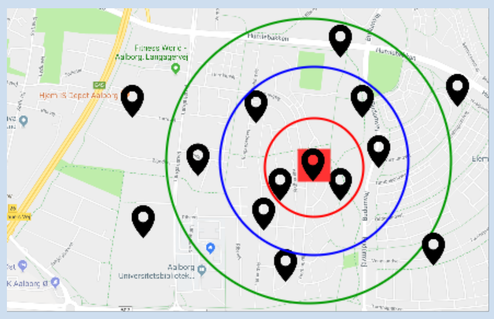

Figure 5(c). *MSSD-R*

PhD Contributions

Motivation:

- Rich location-based data and activity
- Lack of recent datasets
- Limited accessibility
- Real-world user activity

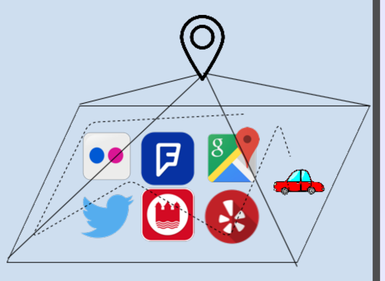

Figure 1. Locations

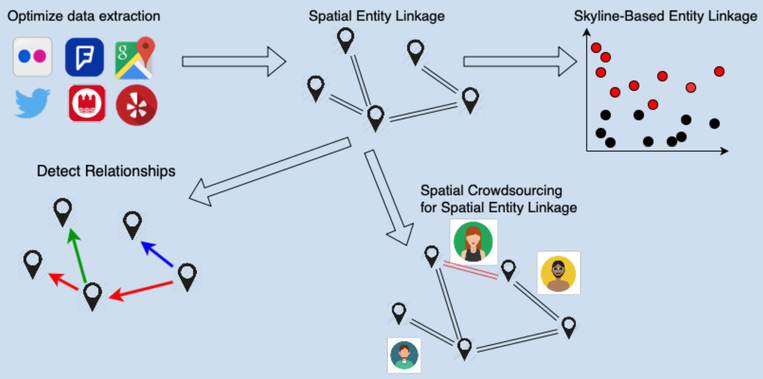

Figure 2. Overall PhD plan

Spatial Entity Linkage

Spatial Entity Linkage problem: Discover whether a pair of spatial entities $\langle s_i, s_j \rangle$ refers to the same physical real-world entity.

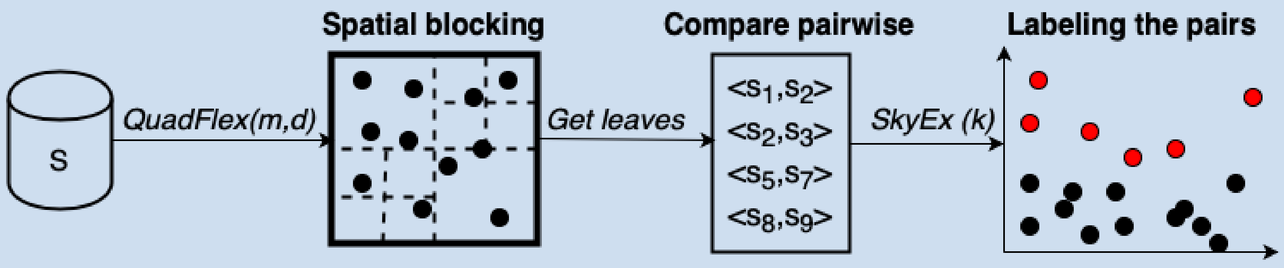

Figure 6. QuadSky approach

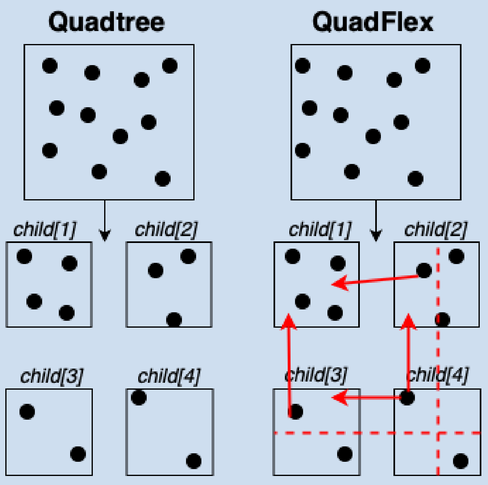

Figure 7. QuadFlex approach

Algorithm 1 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+ ← ∅
for m in [1, k] do
    Filter Skyline(m) = {⟨s_i, s_j⟩} | ∀⟨s', s''⟩ ∈ P − {⟨s_i, s_j⟩} ,
    u(⟨s_i, s_j⟩) > u⟨s', s''⟩}    // Find the Skyline
    Add Skyline(m) to P+    // Label the skyline pairs as positive
    P = P − Skyline(m)
end for
P− ← P    // Label the rest as negative
return P+, P−
```

QuadSky Results

QuadFlex vs. FNN

- *QuadFlex* maintains an execution time that is 8 times less than FNN GiST and 3 times less than FNN SP-GiST.
- *QuadFlex* enumerates 12 times more comparisons than quadtree and 99.99% comparisons of FNN.

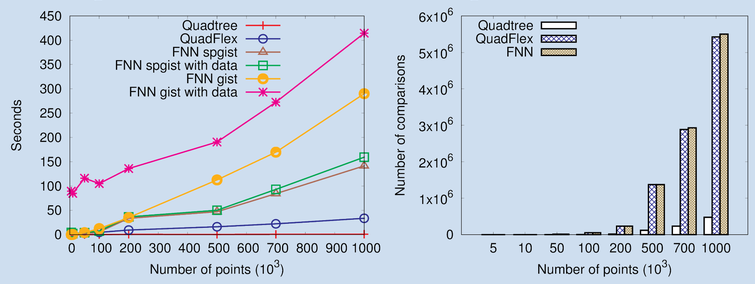

Figure 8. Execution time and nr of comparisons

SkyEx labeling

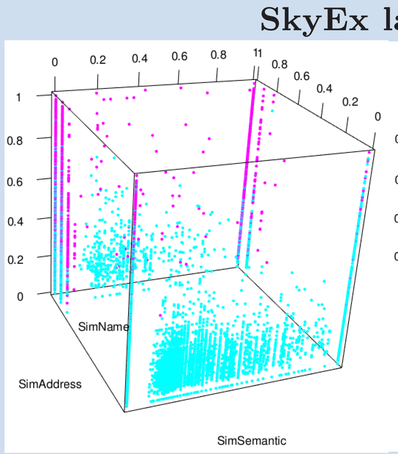

Figure 9(a). Actual

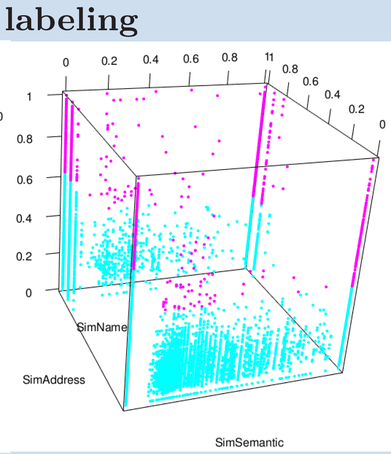

Figure 9(b). SkyEx

Comparison to other approaches

Approach	Precision	Recall	F1
Berjawi et al.(V1)	**0.93**	0.26	0.41
Berjawi et al.(V2)	0.73	0.56	0.63
Morana et al.	0.39	0.60	0.47
Karam et al.	0.23	**0.73**	0.35
QuadSky	0.87	0.60	**0.72**

MSSD Requests vs. Number of Locations results

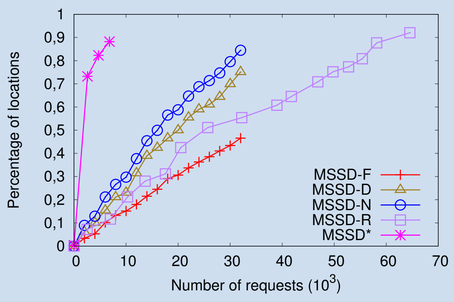

Figure 10(a). Flickr

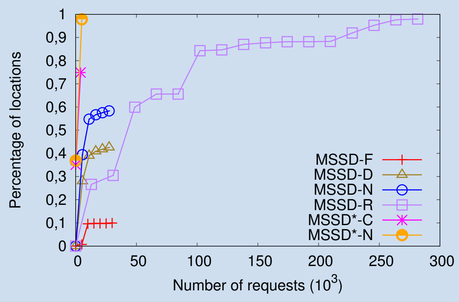

Figure 10(b). Twitter

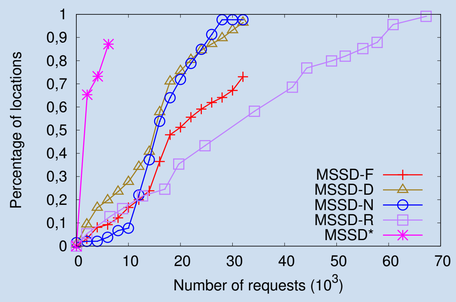

Figure 10(c). Foursquare

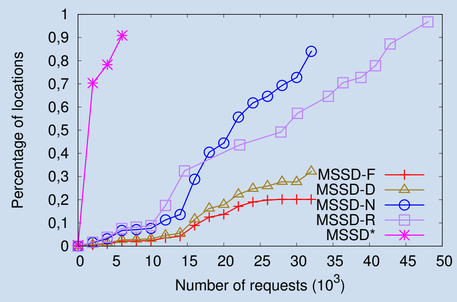

Figure 10(d). Yelp

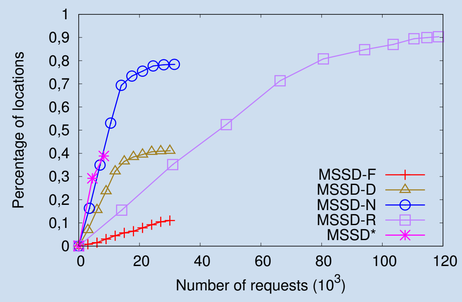

Figure 10(e). Google

References

1. Suela Isaj, Torben Bach Pedersen: *Seed-Driven Geo-Social Data Extraction*, in SSTD 2019
2. Suela Isaj, Esteban Zimányi, Torben Bach Pedersen: *Multi-Source Spatial Entity Linkage*, in SSTD 2019