Graph-Based Approaches to Clustering Network-Constrained Trajectory Data

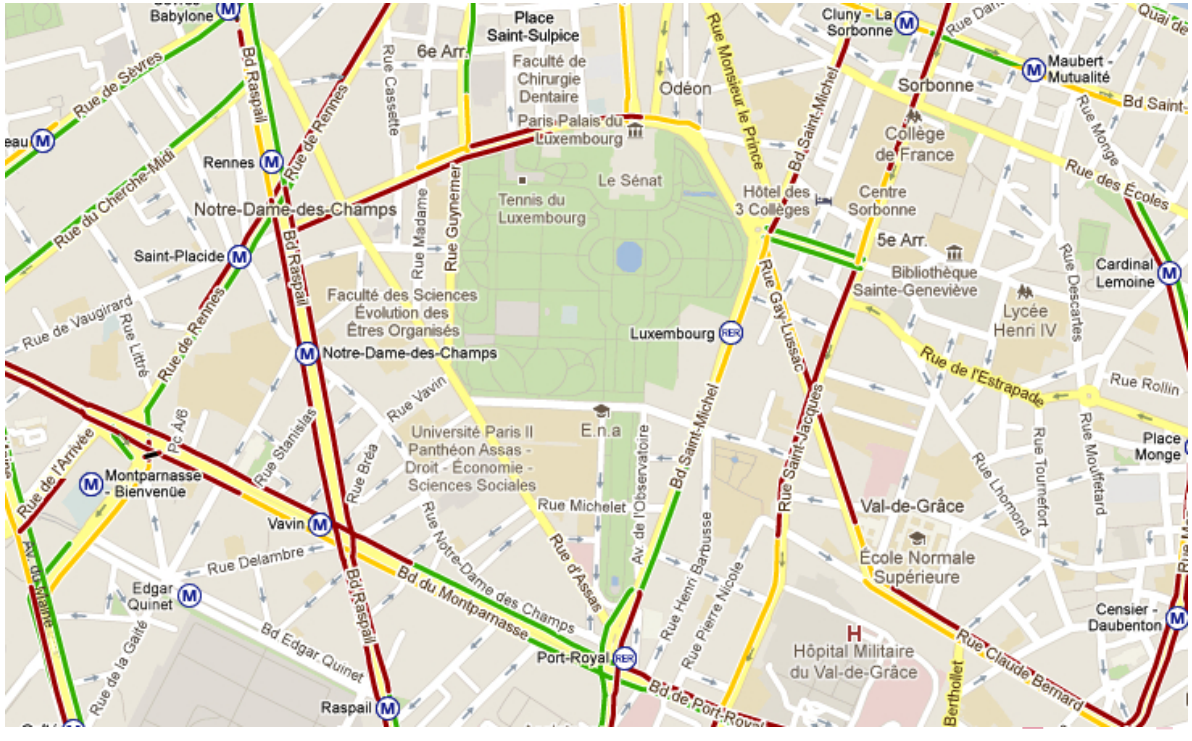
Mohamed K. El Mahrsi*, Fabrice Rossi**

* Télécom ParisTech - Département Informatique et Réseaux (Paris, France)
** Équipe SAMM EA 4543 - Université Paris I Panthéon-Sorbonne (Paris, France)

NFMCP Workshop
ECML-PKDD 2012 (Bristol, UK)

Context and Motivations

- Traffic congestions
  - Environmental damages
  - Economical losses
- Monitoring road traffic
  - Dedicated sensors
  - High deployment and maintenance costs
  - Incomplete data about the state of the road network



Context and Motivations

- Widespread of location-aware devices (GPS, smartphones, PDAs, etc.)
- Store and share MO trajectories (vehicles, pedestrians, etc.)
- Mine and analyze trajectory data to gain a better understanding of flow dynamics in the road network?
- Objective: cluster trajectory data in order to
  - Discover groups of trajectories that moved along the same parts of the network
  - Regroup roads that where visited by the same trajectories

Outline

Existing Approaches

- Trajectory similarity
  - Free movement in Euclidean space: DTW, LCSS, EDR, ERP, ...
  - Under network constraints [Hwang et al., 2005, Hwang et al., 2006, Zhao et al., 2009, Xia et al., 2010].
- Trajectory clustering
  - TraClus [Lee et al., 2007];
  - Moving clusters [Kalnis et al., 2005];
  - Flock patterns [Benkert et al., 2006, Vieira et al., 2009];
  - Convoy patterns [Jeung et al., 2008] ;
  - Under network constraints: NetScan, NNCluster, ...

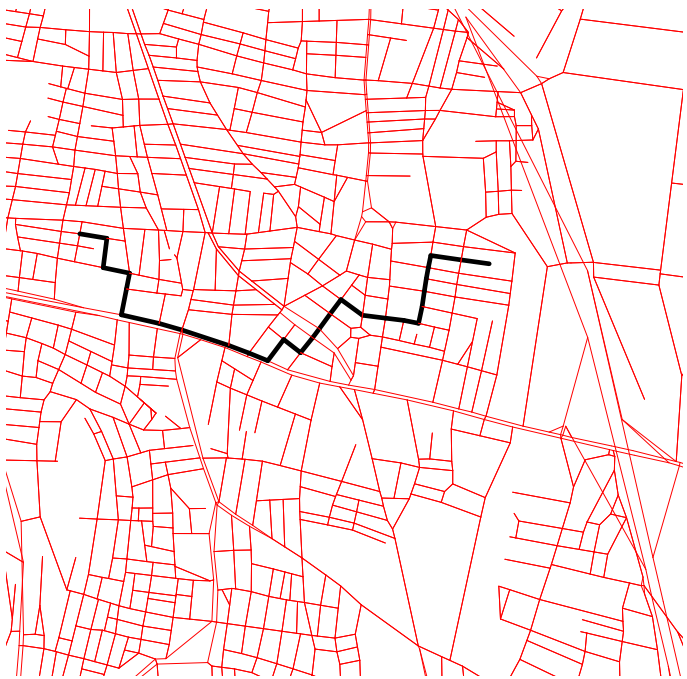
"Limitations" of Existing Approaches

- Focus on unconstrained trajectory clustering
- Use of density based clustering
  - no optimization of a quality criterion
  - very sensitive to the parameters
  - inefficient in cas of heterogeneous density
- Flat clustering $\rightarrow$ large number of clusters

Outline

Network-Constrained Trajectories Data Model[1]

- Road network : directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - vertices: intersections and terminal points of roads
  - edges: road segments (with travel direction)
- Trajectory : ordered sequence of visited segments
  - $T =< id, \{e_1, e_2, ..., e_i, ..., e_l\} >$
  - $\forall 1 \leq i \leq l-1$, e_i and e_{i+1} are connected



[1] [Brakatsoulas et al., 2005, Kharrat et al., 2008, Kharrat et al., 2009, Lou et al., 2009, Roh and Hwang, 2010]

Problem n°1: The Trajectory Clustering Problem

- Given a set of trajectories $\mathcal{T} = \{T_1, T_2, ..., T_n\}$
- Partition $\mathcal{T}$ into a set of clusters $\mathcal{C} = \{C_1, C_2, ..., C_k\}$, such as:
  - trajectories in a same cluster C_i share as much road segments as possible;
  - trajectories in to different clusters C_i and C_j share as few segments as possible.

Problem n°2: The Road Segment Clustering Problem

- Given a set of trajectories $\mathcal{T} = \{T_1, T_2, ..., T_n\}$
- Given the set of edges (segments) $\mathcal{E}$ travelled by $\mathcal{T}$
- Partition $\mathcal{E}$ into a set of clusters $\mathcal{C}' = \{C'_1, C'_2, ..., C'_{k'}\}$, such as:
  - segments in the same cluster are travelled by the same trajectories;
  - segments in two different clusters C'_i and C'_j are visited by different trajectories.

Solving the Trajectory Clustering Problem

- Data model? $\rightarrow$ Network-constrained trajectory model (symbolic)
- Distance/similarity measure?
- Clustering algorithm?

Trajectory similarity

- Trajectories are regarded as "bags of segments"
- Use a cosine similarity between trajectories

$$\text{Similarity}(T_i, T_j) = \frac{\sum_{e \in \mathcal{E}} \omega_{e,T_i} \cdot \omega_{e,T_j}}{\sqrt{\sum_{e \in \mathcal{E}} \omega_{e,T_i}^2} \cdot \sqrt{\sum_{e \in \mathcal{E}} \omega_{e,T_j}^2}}$$

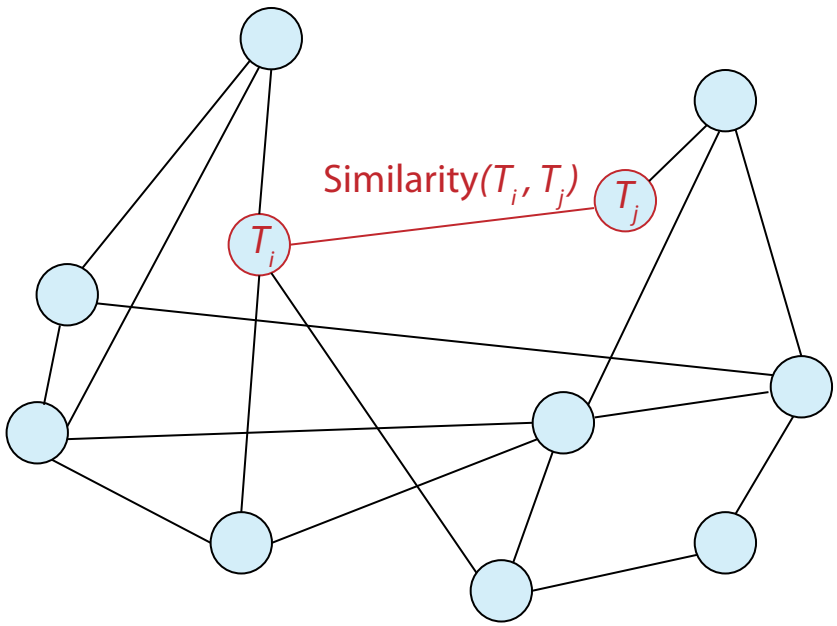
- With modified tf-idf[2] weighting

$$\omega_{e,T} = \frac{n_{e,T} \cdot \text{length}(e)}{\sum_{e' \in T} n_{e',T} \cdot \text{length}(e')} \cdot \log \frac{|\mathcal{T}|}{|\{T_i : e \in T_i\}|}$$

[2] tf-idf: term frequency - inverse document frequency

Clustering algorithm

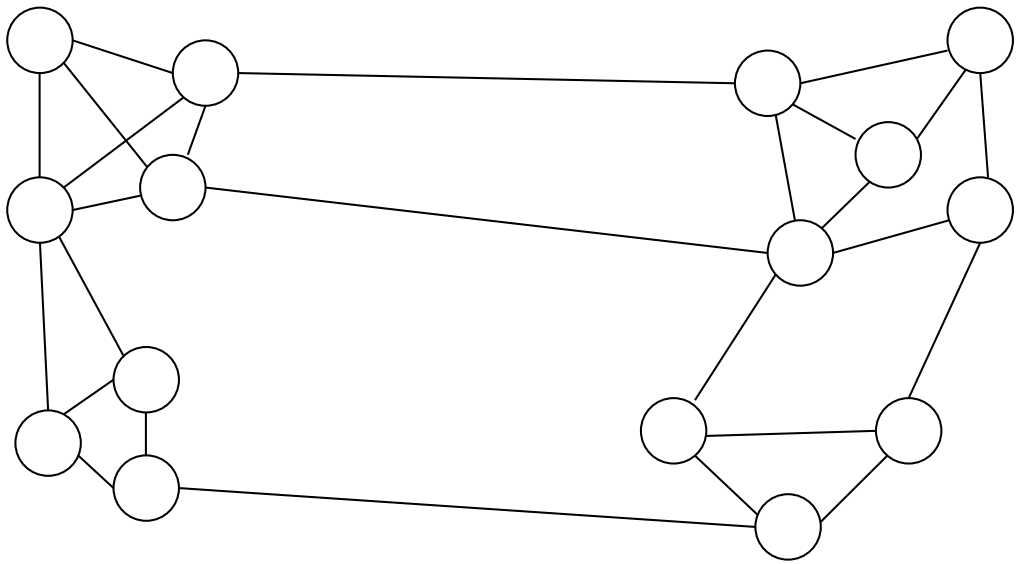
- Key idea: transpose the trajectory clustering algorithm into a graph clustering one
- Similarity graph



- Cluster the graph using a hierarchical modularity-based community detection algorithm [Noack and Rotta, 2009]

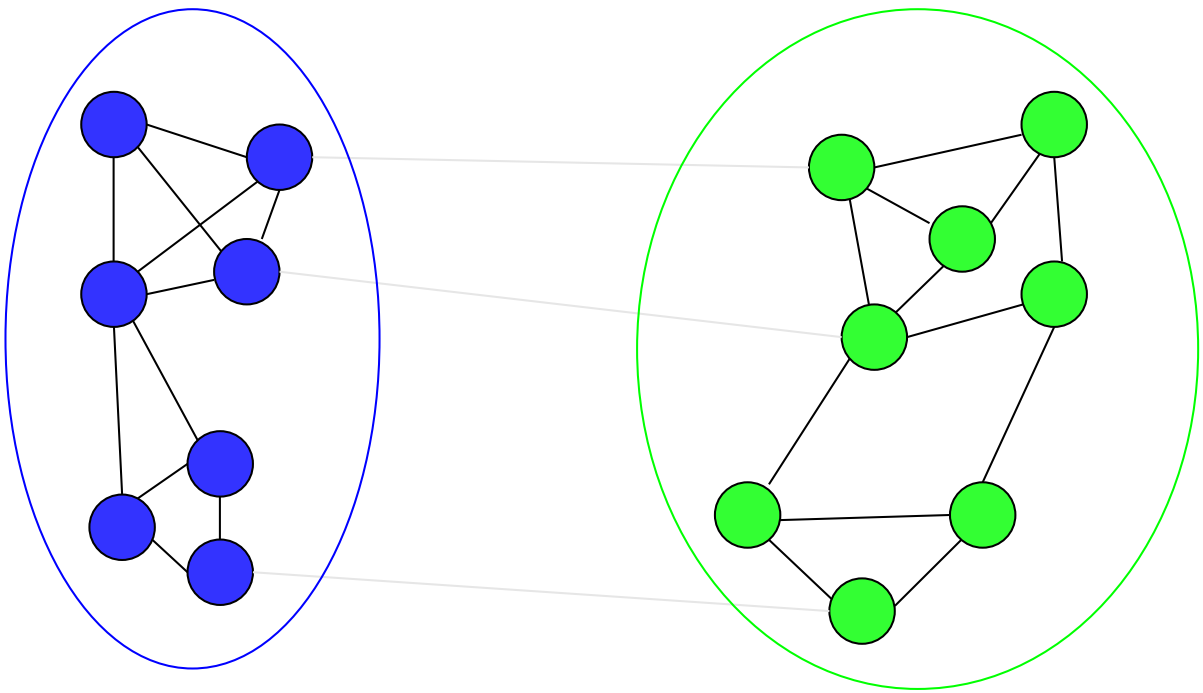
The Modularity-Based Clustering Algorithm

- The algorithm finds the best partitionning of the graph (the one maximizing the modularity measure)



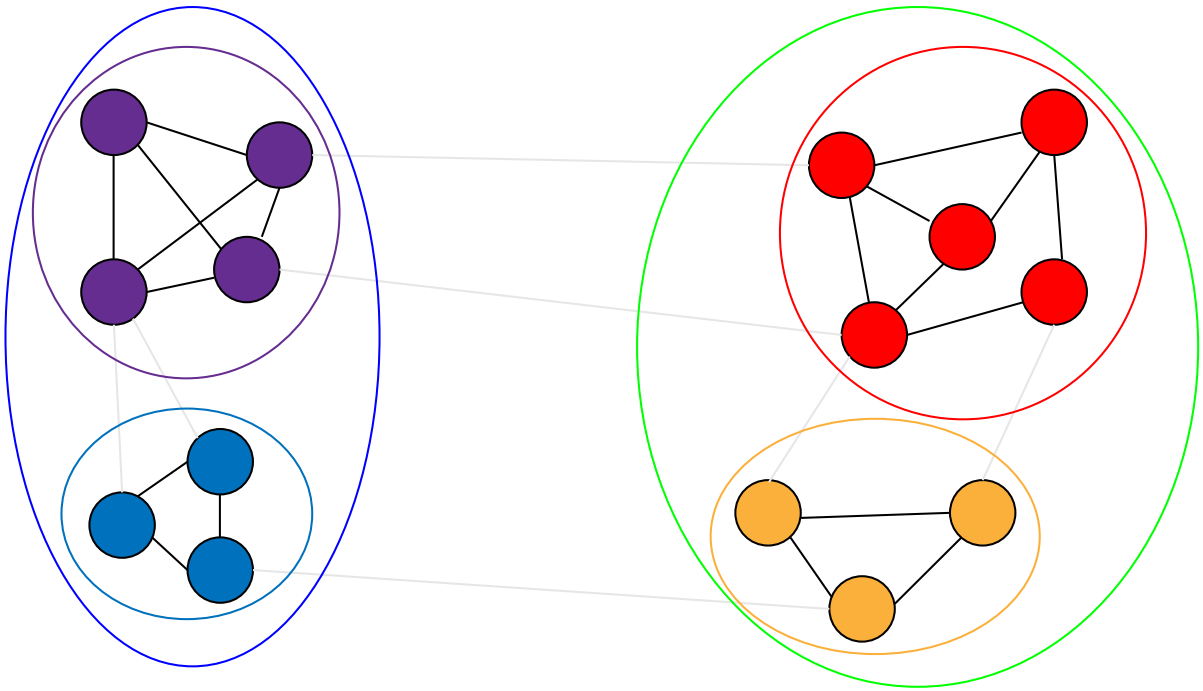
The Modularity-Based Clustering Algorithm

- Each community is isolated and clustered individually



The Modularity-Based Clustering Algorithm

- The recursion stops when the clustering does not yield a "significant" partition



Solving the Segment Clustering Problem

- Proceed by analogy to trajectory clustering
- Segments as bags-of-trajectories
- Again, use a cosine similarity with modified tf-idf
- Similarity graph
  - vertices: road segments
  - edges: similarity between the road segments
- Use the same community detection algorithm to discover segment clusters

Outline

Data Description

- Synthetic datasets
- The Oldenburg road network
  - 6105 vertices (road intersections)
  - 7035 undirected edges
- Two types
  - Datasets with random trajectories generated using the Brinkhoff generator
  - Datasets generated with labeled trajectories

Trajectory Clustering: Results

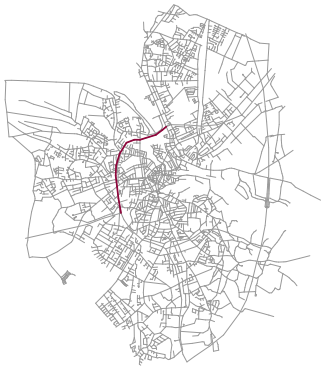
- Comparison to classic hierarchical agglomerative clustering and the NNCluster baseline
- Quality indexes: intra-cluster overlaps, ARI, purity and entropy

Dataset (clusters)	Discovered clusters	Adj. Rand Index		Purity		Entropy	
		NNCluster(B)	Mod.	NNCluster(B)	Mod.	NNCluster(B)	Mod.
1 (9)	9	0.902	1	0.924	1	0.062	0
2 (10)	10	0.881	1	0.902	1	0.059	0
3 (11)	11	0.764	0.872	0.823	0.915	0.113	0.064
4 (6)	6	1	1	1	1	0	0
5 (6)	6	1	1	1	1	0	0
6 (6)	6	1	1	1	1	0	0
7 (12)	14	0.618	0.960	0.712	1	0.185	0
8 (11)	12	0.921	0.971	0.942	1	0.038	0
9 (12)	10	0.752	0.889	0.778	0.872	0.136	0.075

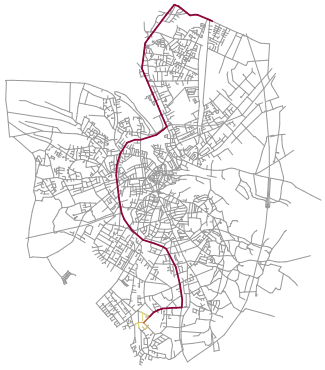
Segment Clustering: Results

- How to evaluate the discovered clusters? $\rightarrow$ not so subtle!
- Still looking for meaningful quality indexes
- "Explain" the segment clusters using the trajectory clusters

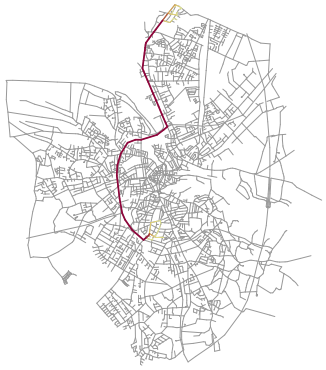
Segment Clustering: Results



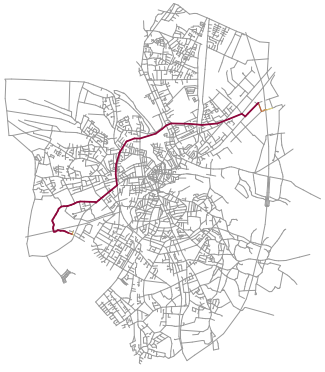
(a) 14 segments



(b) 12 trajectories



(c) 19 trajectories



(d) 7 trajectories

(e) 3 trajectories

(f) 4 trajectories

Figure: Example of a segment cluster (a) and the trajectory clusters that crossed it (b-f) detected in a small dataset (85 trajectories).

Conclusion

- New framework to clustering trajectory data
- Advantages:
  - integration of road network constraints
  - well defined quality criterion
  - non parametric
  - hierarchical clustering suitable for multi-level exploration
- Drawbacks:
  - sensitivity to noise
  - high complexity ($O(n^3)$ in theory, $O(n^2)$ in practice)
  - segment clusters are difficult to interpret and evaluate

Future Work

- Experiment on real datasets
  - availability?
  - map matching?
- Bi-clustering of trajectories and segments simultaneously
- Comparison to unconstrained clustering approaches

References I

Benkert, M., Gudmundsson, J., Hübner, F., and Wolle, T. (2006).

Reporting flock patterns.
In *ESA'06: Proceedings of the 14th conference on Annual European Symposium*, pages 660--671, London, UK. Springer-Verlag.

Brakatsoulas, S., Pfoser, D., Salas, R., and Wenk, C. (2005).

On map-matching vehicle tracking data.
In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 853--864. VLDB Endowment.

Hwang, J.-R., Kang, H.-Y., and Li, K.-J. (2005).

Spatio-temporal similarity analysis between trajectories on road networks.
In Akoka, J., Liddle, S. W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., Cherfi, S. S.-S., van den Heuvel, W.-J., Thalheim, B., Kolp, M., Bresciani, P., Trujillo, J., Kop, C., and Mayr, H. C., editors, *ER (Workshops)*, volume 3770 of *Lecture Notes in Computer Science*, pages 280--289. Springer.

Hwang, J.-r., Kang, H.-y., and Li, K.-j. (2006).

Searching for similar trajectories on road networks using spatio-temporal similarity.

Jeung, H., Yiu, M. L., Zhou, X., Jensen, C. S., and Shen, H. T. (2008).

Discovery of convoys in trajectory databases.
Proc. VLDB Endow., 1(1):1068--1080.

Kalnis, P., Kalnis, P., Mamoulis, N., and Bakiras, S. (2005).

On discovering moving clusters in spatio-temporal data.
IN SSTD, pages 364--381.

References II

Kharrat, A., Popa, I. S., Zeitouni, K., and Faiz, S. (2008).

Clustering algorithm for network constraint trajectories.
In Ruas, A. and Gold, C. M., editors, *SDH*, Lecture Notes in Geoinformation and Cartography, pages 631--647. Springer.

Kharrat, A., Popa, I. S., Zeitouni, K., and Faiz, S. (2009).

Caractérisation de la densité de trafic et de son évolution à partir de trajectoires d'objets mobiles.
In Menga, D. and Sedes, F., editors, *UbiMob*, volume 394 of *ACM International Conference Proceeding Series*, pages 33--40. ACM.

Lee, J.-G., Han, J., and Whang, K.-Y. (2007).

Trajectory clustering: a partition-and-group framework.
In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593--604, New York, NY, USA. ACM.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., and Huang, Y. (2009).

Map-matching for low-sampling-rate gps trajectories.
In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 352--361, New York, NY, USA. ACM.

Noack, A. and Rotta, R. (2009).

Multi-level algorithms for modularity clustering.
In *Proceedings of the 8th International Symposium on Experimental Algorithms*, SEA '09, pages 257--268, Berlin, Heidelberg. Springer-Verlag.

Roh, G.-P. and Hwang, S.-w. (2010).

Nncluster: An efficient clustering algorithm for road network trajectories.
In Kitagawa, H., Ishikawa, Y., Li, Q., and Watanabe, C., editors, *Database Systems for Advanced Applications*, volume 5982 of *Lecture Notes in Computer Science*, pages 47--61. Springer Berlin - Heidelberg.

References III

Vieira, M. R., Bakalov, P., and Tsotras, V. J. (2009).

On-line discovery of flock patterns in spatio-temporal data.
In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 286--295, New York, NY, USA. ACM.

Xia, Y., Wang, G.-Y., Zhang, X., Kim, G.-B., and Bae, H.-Y. (2010).

Research of spatio-temporal similarity measure on network constrained trajectory data.
In *Proceedings of the 5th international conference on Rough set and knowledge technology*, RSKT'10, pages 491--498, Berlin, Heidelberg. Springer-Verlag.

Zhao, H., Han, Q., Pan, H., and Yin, G. (2009).

Spatio-temporal similarity measure for trajectories on road networks.
In *Proceedings of the 2009 Fourth International Conference on Internet Computing for Science and Engineering*, ICICSE '09, pages 189--193, Washington, DC, USA. IEEE Computer Society.