Clustering and Modeling of Network level Traffic States based on Locality Preservative Non-negative Matrix Factorization

Yufei Han
Post-doctoral research fellow, Centre de Robotique, MINES-ParisTech
60 Saint Michel Boulevard, 75272, Paris
Yufei.Han@mines-paristech.fr
Fabien Moutarde
Associate Professor, Centre de Robotique, MINES-ParisTech
60 Saint Michel Boulevard, 75272, Paris
Fabien.Moutarde@mines-paristech.fr

ABSTRACT

In this paper, we propose to cluster and model network-level traffic states based on a geometrical weighted similarity measure of network-level traffic states and locality preservative non-negative matrix factorization. The geometrical weighted similarity measure makes use of correlation between neighboring roads to describe spatial configurations of global traffic patterns. Based on it, we project original high-dimensional network-level traffic information into a feature space of much less dimensionality through the matrix factorization method. With the obtained low-dimensional representation of global traffic information, we can describe global traffic patterns and the evolution of global traffic states in a flexible way. The experiments prove validity of our method for the case of large-scale traffic network.

INTRODUCTION

Most of previous studies on intelligent traffic focus on mining temporal patterns of traffic data measured on individual links [1][2][3]. These works only analyze temporal properties of local link level traffic states. In fact, in a typical urban traffic scenario, traffic states of one link are correlated with neighboring areas. Network-level traffic states can be regarded as complementary knowledge and constraints in predicting or analyzing link level traffic patterns. Therefore, in recent years, with improvement of intelligent transportation systems, it becomes necessary to unveil global traffic patterns at network level. Global traffic information provides overall descriptions of spatial configurations of traffic states over the whole road network, which can improve performances of traffic guidance or control systems [3].

In large-scale traffic networks, like urban traffic systems, network level traffic information is often represented in a high-dimensional feature space, which makes it

difficult to extract characteristics of global traffic states. In our work, we firstly adopt a geometrical weighted distance to evaluate similarity between network-level traffic patterns, which is described in the second section. Then, we make use of a matrix factorization method with a topological reguliarization term to obtain a low-dimensional representation model of global traffic states, as described in the third section. In a further step, we perform clustering of global traffic states based on the learned low dimensional representation, in order to extract typical spatial patterns of network-level traffic states. In the final part of the paper, we present clustering structures of network-level traffic patterns with respect to a large-scale link network and make conclusions of the whole paper.

GLOBAL TRAFFIC STATUS SIMILARITY MEASURE

A network level traffic state is defined by a sequence of link level traffic states with respect to each individual link in the road network, which is normally represented in a n-dimensional vector, with n being the number of links in the network. Besides simply concatenating all link level traffic information into the high dimensional representation, network-level traffic states also contain spatial configurations of link level traffic states in the network, which form different global traffic state patterns. In a typical traffic network, the traffic state of one specific link is closely correlated with its up-stream or down-stream nearest neighbors in most cases. For example, in Figure.1, the links u_i^j and d_i^m are up-stream and down-steam nearest neighbors of the link i respectively. Assuming link i fell into a heavy traffic jam, its neighboring links u_i^j and d_i^m are more likely to be congested with vehicles than those which are located far from the link i and vice versa. Motivated by the property, we propose to adopt a weighted fusion scheme among traffic states with respect to geometrical neighborhoods of links during evaluating similarity between network level traffic states observed at different time. As we can see in Figure.1, we firstly calculate differences between traffic states of corresponding links. For each link i, we then obtain a weighted sum of the link-wise difference values with respect to the link i and its up-stream and down-stream neighbors, which is defined to be local variation v_i of traffic states around the current link, as denoted in Figure.1 and Eq.1:

$$v_i = \sum_j w_j^u a(u_i^j) + \sum_m w_m^d a(d_i^m) + w^i a(i) \tag{1}$$

a is the link-wise difference between traffic states of the corresponding link. w_j^u, w_m^d and w^i is the weights attached to up-stream neighbors, down-stream neighbors and the current link i respectively. After that, we map the L1 norm of $\{v_i\}$ into [0, 1] using a Gaussian kernel in Eq.2 as the final similarity measure between two network-level traffic states:

$$s = e^{-\frac{\sum_i v(i)}{2\sigma^2}} \tag{2}$$

To normalize range of the weighted sum, the sum of all weights is required to be 1. The weight w^j corresponding to the link j should be the largest one. Weights of the neighboring links can be designed to be proportional to degrees of traffic state correlation between one specific neighboring link and the current link i. In this paper, we just treat all neighboring links are of equal importance. Therefore, all neighboring links take the same weight value. By performing weighted fusion among local neighborhoods in the network, the derived similarity measure not only represents traffic state variations between corresponding links but also indicates the spatial correlations between local neighborhoods.

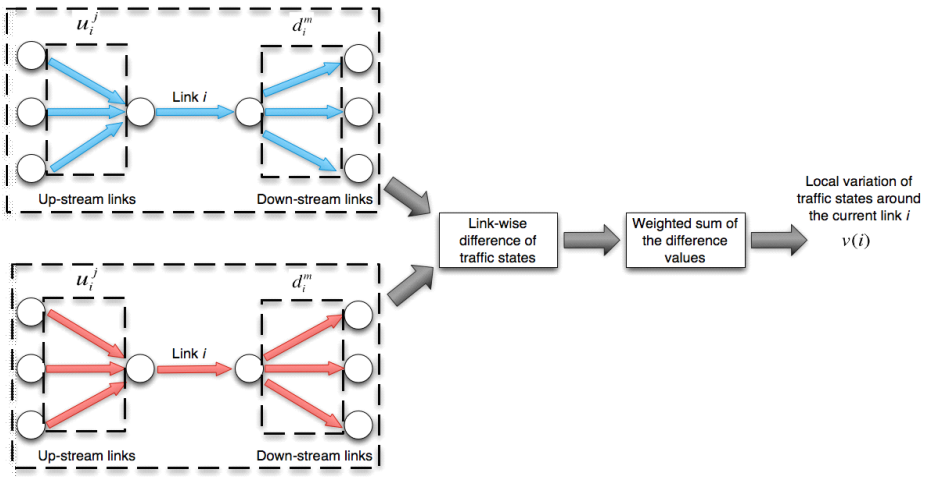

Figure.1 Geometrical weighted similarity measure

MODELING LOW-DIMENSIONAL REPRESENTATION OF NETWORK-LEVEL TRAFFIC STATUS

Dimensionality of network-level traffic status representation is directly proportional to the number of links in the network. Given a large-scale network, which is common in applications of urban traffic control, the resultant high-dimensional traffic state representation is difficult to store or use for traffic prediction / classification due to the curse of dimensionality. To attack this issue, we propose to use locality preserving non-negative matrix factorization (LPNMF) [4][5] to obtain low-dimensional representation of global traffic states. Assuming that k samples of n-dimensional global traffic states are stored as n*k matrix X, LPNMF factorize X into the non-negative n*s matrix M and s*k matrix V^T, which minimizes the following objective function:

$$O = \left\| X - MV^T \right\|_F^2 + \lambda Tr(V^T LV) \tag{3}$$

The first term is the Frobenius reconstruction error with respect to M and V. Each sample is approximated by a linear combination of the columns of M, weighted by the rows of V. Therefore, M can be regarded as containing a learned basis of the global traffic states, while V are s-dimensional representations of original samples in the

given basis. s is set to be much less than the original dimensionality n most of time. Therefore, we actually obtain a much lower dimensional representation of network-level traffic state after factorization, which removes redundancy in the original high-dimensional space and makes it flexible to implement statistical analysis on the manifold V. In contrast with SVD decomposition, derived manifold space is not necessarily orthogonal in NMF. It is also required that each data sample takes positive coordinates in the low-dimensional feature space. The above two properties makes NMF more suitable to describe the latent distribution structures, especially when overlap exists among different clusters of data samples. In the second term of the object function, L is called Graph Laplacian [6], defined as D-W. In the matrix W, w_{ij} is the pair-wise geometrical weighted similarity measure matrix between the i-th and j-th k global traffic state sample. Due to symmetry of the distance measure, W is also a symmetric matrix. D is a diagonal matrix whose entries are column sums of W, defined as Eq.4:

$$D_{ii} = \sum_{j} w_{ij} \tag{4}$$

By adding the Graph Laplacian based constraints, the obtained low-dimensional representation V are calibrated to have similar topological structures as original samples X, which means that two close samples x_i and x_j should also be close in the low-dimensional manifold V. With this property, we can analyze global traffic states easily in the low-dimensional manifold V instead without loss of intrinsic data distribution of original samples X.

Each element v_{ij} of matrix V represents to which degree the i-th original sample is associated with the j-th expanding basis in matrix M. If the i-th sample could be represented solely using the j-th basis, then v_{ij} will take the largest value in the i-th row of V [7]. Therefore, we simply use V to determine the cluster labels of the network-level traffic states. For each x_i, we examine the i-th row of V and assign xi to the j-th cluster, $j = \arg\max_{j} v_{ij}$. Simply as it is, we can still find out the intrinsic distributional properties based on the NMF factorization.

EXPERIMENTAL RESULTS

Experimental settings

To verify validity of the proposed method in clustering and modeling network-level traffic states, we firstly simulate real traffic scenes of the large-scale traffic network of Paris and its suburb regions using Metropolis [8], in order to generate a benchmark traffic database. Metropolis is a planning software [8] that is designed to model

transportation systems. It contains a complete environment to handle dynamic simulations of daily traffic in one specific traffic network, which allows the user to study impacts of transportation management policies in a large-scale urban traffic network in a time-dependent manner. The built traffic database is composed by 4660 road intersections and 13627 links in the network, as we can see in the Figure.2. Each simulated traffic scene is generated to cover 8 hours of traffic data observations, including congestion in morning rush hours. Different traffic situations are obtained by adding random events and fluctuation in the O-D matrix (Origin-Destination) and capacity of network flow. There are totally 108 simulated traffic scenarios in our benchmark data set. Each one contains 48 time steps, corresponding to 15-minute bins over which the network traffic flow are aggregated. To represent traffic states, we propose to use traffic index [9][10] in each link at a specific time, as in Eq.5.

$$x_{lt} = \frac{\Delta t_l^0}{\Delta t_{lt}} \in [0,1] \tag{5}$$

The denominator is the observed travel time in link l at time t, the nominator is the free-flow travel time on this link. The smaller the traffic index is, the corresponding link is more congested. To perform clustering analysis, we concatenate all observations of traffic states into a 13627*5184 matrix. Each column corresponds to a network-level traffic status obtained at each time step, which is a 13627-dimensional vector. In the experiment, the number of centroids in clustering is set to be 3 and 5 respectively. For the convenience of visualization, we project all the column vectors into 3-dimensional PCA space to illustrate structures of the obtained clusters.

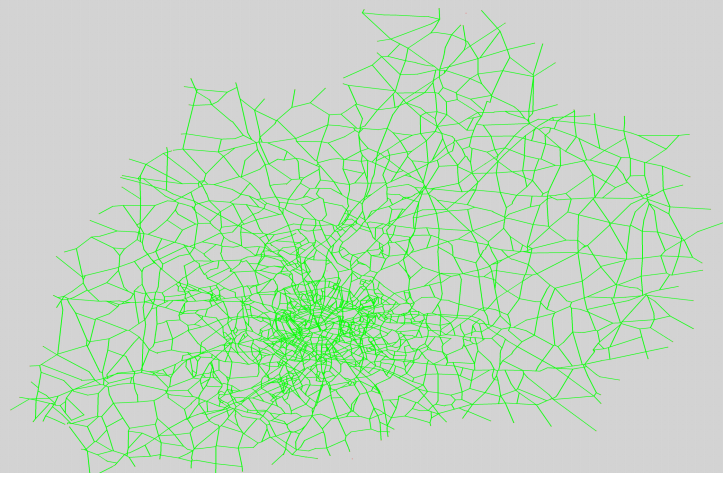

Figure.2 Traffic network of Paris and suburb regions

Results of network-level traffic state clustering

In the 3D PCA space, as shown in three different viewpoints in Figure.3, the samples corresponding to the free-flowing network-level states are concentrated within a small region in the PCA space. By contrast, samples corresponding to network-level congestion are distributed sparsely and far from the region of the free-flowing state. Notably, with increasing degrees of traffic jam in the network, variations of

network-level traffic patterns become larger and larger. In fact, spatial configurations of global traffic states keep the same if the whole network is free-flowing everywhere. On the contrary, congestion occurred at different parts of the network change the spatial configurations in different ways, which introduces variations in global traffic patterns. Figure.4 illustrates the three clusters derived according to the proposed clustering algorithm. The cluster labeled by blue legends represents that almost all links are fluid in the network. Both red and dark green clusters indicate that traffic jam occurs in the network. Figure.4 also shows spatial configurations of the most congested network-level traffic states in each cluster, which have the least average value of traffic indices among the corresponding clusters. They are used here as representative exemplars of global traffic patterns contained in each cluster. In the exemplars, red color is used to label jammed links whose traffic indices are less than a specified threshold, while green color used for fluid links. It can be seen that the exemplar of the red cluster contains much less busy links than the dark green one. It denotes that network-level traffic states of the two clusters contain different degrees of traffic congestions. Traffic network states in the green cluster contain much heavier congestions. Furthermore, according to the exemplars, most of jammed links are concentrate in the central region of the network. It denotes that most traffic congestion occurs inside Paris. Suburb regions are free-flowing most of the times.

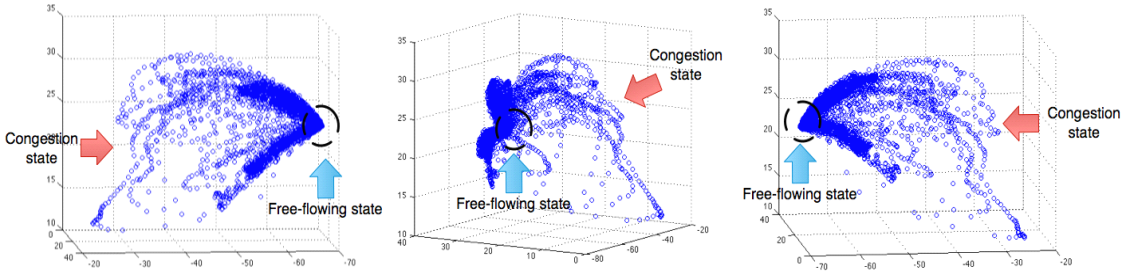

Figure.3 A three-views diagram of network-level traffic state representation in 3D PCA space

By increasing the number of clustering centrioids to 5, we can find more detailed structures of network-level traffic states, as shown in Figure.5. Figure.6 illustrates exemplars of clusters following the same settings in Figure.4. In Figure.5, we compare structures of the three clusters obtained above and those obtained after increasing of clusters. The cluster corresponding to light traffic congestion, labeled by red legends in Figure.4, is further split into two parts that are labeled by pink and purple legends respectively in Figure.5. These two sub-clusters form elongated shapes oriented to different directions in 3D-PCA space, which implies different distribution settings of congestion in the network. Exemplars of these two clusters make the difference more clear, as shown in the Figure.6(a) and (b). In the exemplar of the sub-cluster labeled by pink legends, illustrated in Figure.6(a), busy links tend to be more close to the central region than the exemplar of the sub-cluster labeled by purple legends, as shown in Figure.6(b). Despite of similar degrees of network-level congestion in both two exemplars, they indicate different spatial configurations of

traffic states in the network, which is consistent with the difference of orientations of the two elongated sub-clusters.

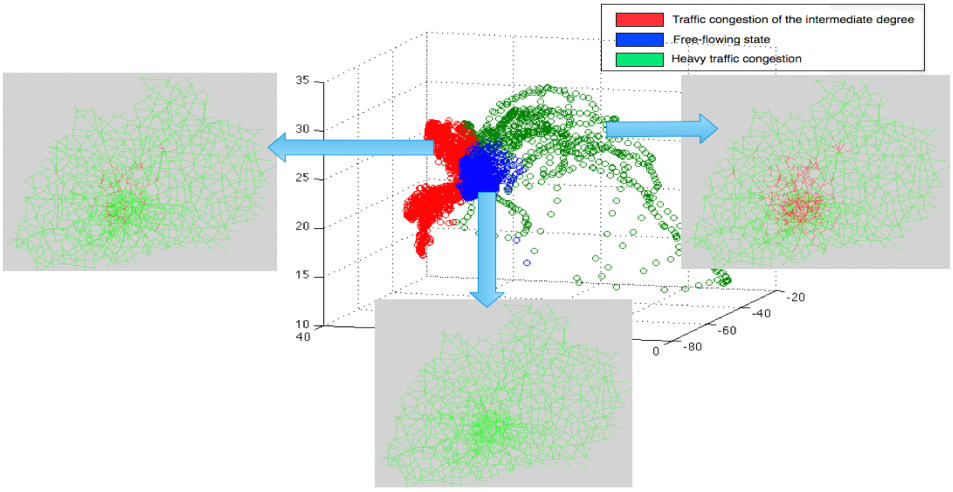

Figure.4 Three clusters and exemplars of network-level traffic states

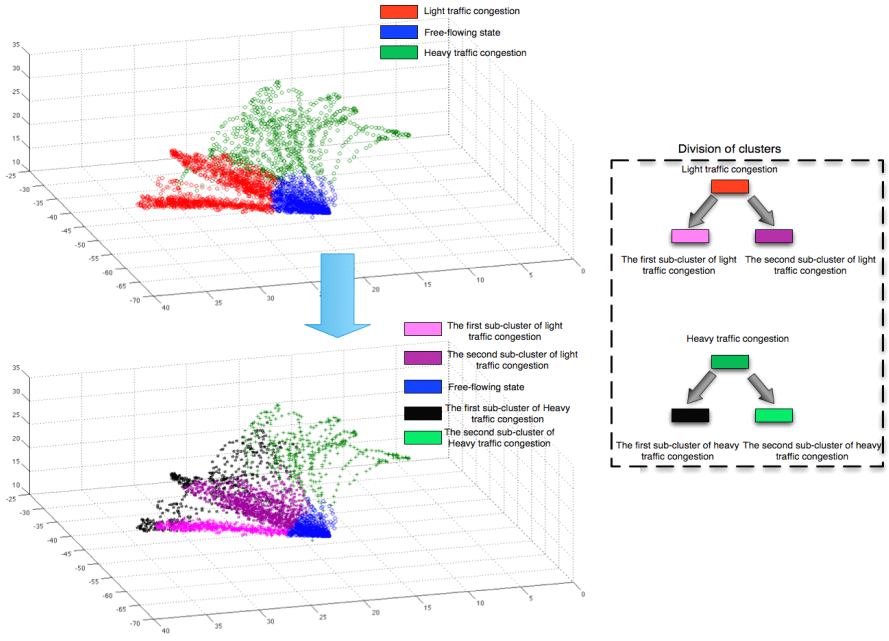

Figure.5 Division of clusters after increasing the number of clusters

Similar manner of split of the cluster can also be observed in the cluster corresponding to heavy traffic congestion in Figure.4, labeled by dark green legends. As we can see in Figure.5, this cluster is split to two sub-clusters labeled light green and black legends. Due to large variations of spatial configurations of traffic congestions, both of two sub-clusters have sparse structures in the 3D-PCA space. However, they differ in degrees and spatial layout of network-level traffic congestion. In Figure.6(c) and (d), we compare the exemplars of the two sub-clusters labeled by black and light green legends in Figure.5 respectively. Generally, the exemplar in

Figure.6(d) contains more busy links. Furthermore, although the central region of the network are highly congested in both exemplars, the area to which network-level traffic congestion extend is more wide in the exemplar shown in the Figure.6(d), especially in suburb regions. This implies a different setting of traffic scenarios during simulation. Notably, we should notice that a small part of samples in the cluster of light traffic congestion in Figure.4 are assigned into the sub-cluster labeled by black legends in Figure.5. This may be partly caused by existence of sparsely distributed data points that can't get stable cluster assignment. More importantly, it is actually smooth and continuous variations for network-level traffic status to evolve from the free-flowing state to congestion. There is no clear boundary between the two states. Therefore, cluster assignments are fuzzy for the samples of network-level status lying around the boundary between the two clusters.

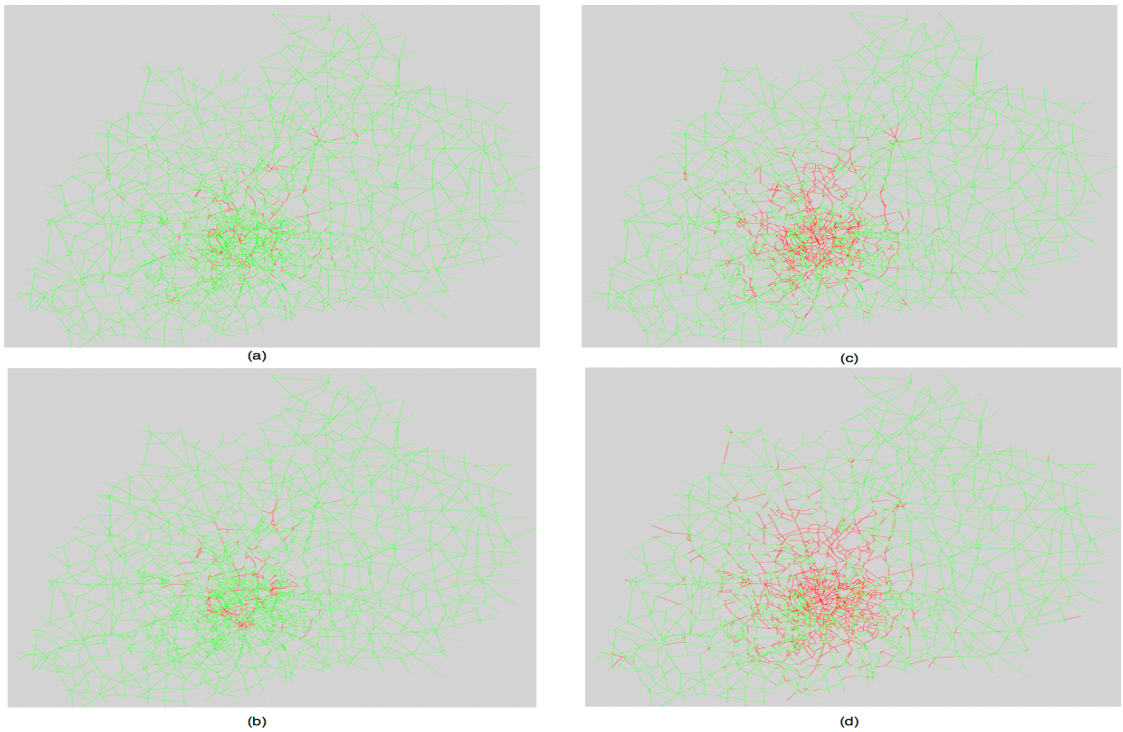

Figure.6 Exemplars of sub-clusters. (a) and (b) are exemplars of sub-clusters labeled by pink and purple legends respectively. (c) and (d) are exemplars of sub-clusters labeled by black and light green legends respectively.

In a typical traffic scenario simulation, the whole traffic network is free-flow at the beginning. Subsequently, congestion emerges and becomes heavier and heavier until reaching the peak of traffic jam. Finally, network-level traffic states recover gradually. Therefore, network-level traffic states evolve in circular trajectories in the PCA space, as shown in Figure.7. Along the trajectories, transitions from the free-flow state to clusters corresponding to different network-level congestion patterns imply totally different temporal traffic dynamics in the network. To make it intuitive, we select two kinds of trajectories. Both start from the free-flowing state but achieve peaks of traffic congestions in the clusters labeled by pink and black legends respectively. Each

trajectory is composed by observations of network-level traffic states at 48 time steps in our benchmark database. For each time step, we take the average traffic index value of all 13627 links in the network as a crude measure of global traffic state at the current time. The lower it is, the heavier congestion occurs in the network. Within each type of trajectories, we calculate average of the mean traffic index at each time step of each trajectory, which results in a 48-D sequence of average values as a general temporal dynamic pattern of global traffic states in the corresponding type of trajectories. In Figure.8, we compare both dynamic patterns. Trajectories with its peak of congestion located in the cluster labeled by black legends contain traffic jams with longer durations and heavier congestion at their peak points. These two different temporal evolution processe represent different settings of requirements and supplies of traffic resources in the network. Crude as it is, the analysis gives us a hint that we could analyse evolution of network-level traffic states based on the clustering results.

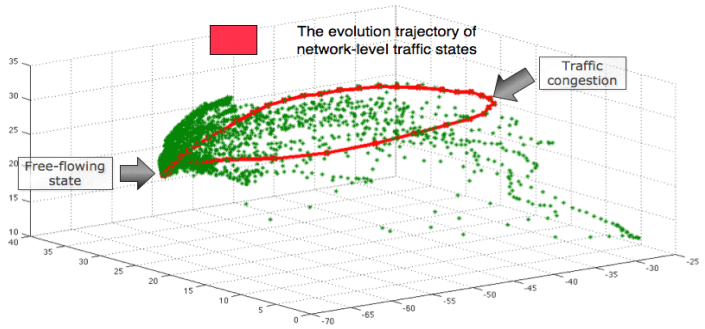

Figure.7 The evolution trajectory of network-level traffic states

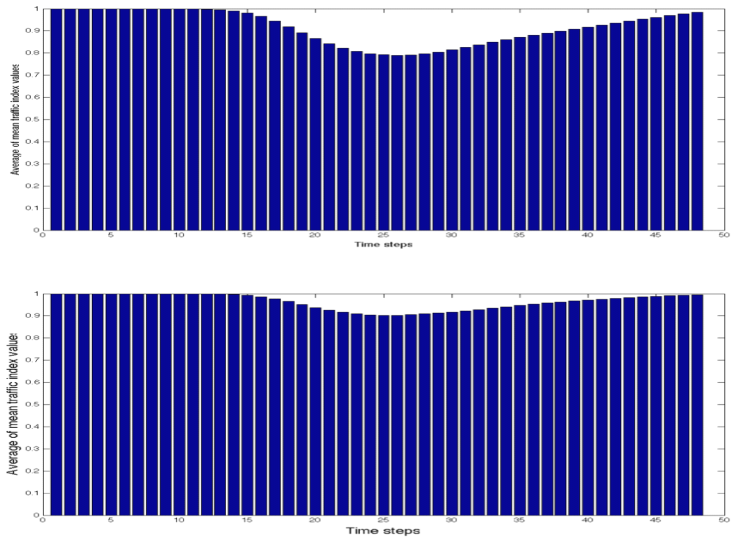

Figure.8 Average temporal evolution patterns of the two sub-clusters

CONCLUSION

In this paper, we propose a geometrical weighted similarity measure to evaluate both link-level differences of traffic states and spatial constraints between traffic states of

neighboring incoming and out-coming links, which is important in describing spatial configurations of network-level states. Using it, we adopt a topology preservative non-negative matrix factoirzation method to obtain a low dimensional model of the high-dimensional network-level traffic states, which keeps the similarity information between original network-level traffic patterns. Thanks to the low-dimensional representation, we can describe or predict the network-level traffic states much more easily. In our work, we focus on clustering of the network-level states. Experimental results not only indicate characteristics of spatial configuration patterns with respect to traffic states in the whole network, but also denote promising applications of the low dimensional representation and clustering results in describing temporal dynamic patterns of the network-level traffic states.

ACKNOWLEDGEMENT

This work was supported by the grant ANR-08-SYSC-017 from the French National Research Agency. The author specially thanks Cyril Furtlehner and Jean-Marc Lasgouttes for providing advice and the benchmark database used in this paper.

REFERENCE

[1] R.Herring, A.Hofleitner, S.Amin, T,Nasr, A.Khalek, P.Abbeel and A.Bayen, "Using mobile phones to forecast arterial traffic through statistical learning",in *the 89th Transportation Research Board Annual Meeting*, January 10-14 2010.

[2] B.Ghosh, B.Basu, and M.O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis", *IEEE Trans. Intell. Transport. Sys.*, vol. 10, no. 2, pp. 246–254, 2009.

[3] H. Kanoh et al., "Short-term traffic prediction using fuzzy c-means and cellular automata in a wide-area road network," in *Proceedings of the 8th International, ser. Conf. Intell. Transport. Sys*. Vienna, Austria, 2005.

[4] D.Cai, X.F.He, X.Y.Wu, and J.W.Han,"Non-negative Matrix Factorization on Manifold", in *Proceedings of International Conference on Data Mining,* Italy, 2008.

[5] D.Cai, X.Fei He, X.H.Wang, H.J.Bao and J.W.Han, "Locality Preserving Nonnegative Matrix Factorization", In *Proceedings of International Joint Conference on Artificial Intelligence*, Pasadena, CA, 2009.

[6] F.R.K.Chung, "Spectral Graph Theory", in *Proceedings of AMS Regional Conference Series in Mathematics*, vol.92,1997.

[7] W.Xu, X.Liu and Y.H.Gong, "Document Clustering Based on Non-negative Matrix Factorization", in *Proceedings of ACM SIGIR 2003,* Canada, 2003.

[8] F. Marchal, "Contribution to dynamic transportation models," *Ph.D.dissertation*, University of Cergy-Pontoise, 2001.

[9] C.Furtlehner, Y.F.Han, J.M.Lasgouttes, V.Martin, F.Marchal and F.Moutarde, "Spatial and Temporal Analysis of Traffic States on Large Scale Networks", In *Proceedings of Intelligent Transportation Systems Conference*,Portugal, 2010.

[10] "TRAVESTI Project Wiki", http://travesti.gforge.inria.fr/