Identification of Tidal-Traffic Patterns in Metro-Area Mobile Networks via Matrix Factorization Based Model

Sebastian Troia*, Gao Sheng†, Rodolfo Alvizu *, Guido Alberto Maier*, Achille Pattavina *
* Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy
† Beijing of Posts and Telecommunications, Beijing 100876, China

***Abstract*—Due to the highly predictable daily movements of citizens in urban areas, mobile traffic shows repetitive patterns with spatio-temporal variations. This phenomenon is known as Tidal Effect analogy to the rise and fall of the sea levels. Recognizing and defining traffic load patterns at the base station thus plays a vital role in traffic engineering, network design and load balancing since it represents an important solution for the Internet Service Providers (ISPs) that face network congestion problems or over-provisioning of the link capacity. Previous works have dealt with the classification and identification of patterns through the use of techniques, which inspect the flow of data of a particular application. But they assume prior knowledge on the stream of data packets, making the trend identification much inefficient. Recent methods based on machine learning techniques build their classification models based on sample data collected at certain points of the network with high accuracy. Therefore, in this paper, we address the problem by applying matrix factorization based models on real-world datasets, identifying typical patterns from data streams, which frequently occur in the network, without investigating the type of flows. For that, we propose a Collective Non-negative Matrix Factorization based model combining multi-source data, such as point of interests attributes, traffic data and base station information, identifying the basic patterns of those areas of the city that present the same type of attributes. The experimental results show the effectiveness of our proposed approach compared with the baselines.**

***Index Terms*—Non-Negative Matrix Factorization (NMF), Collective NMF (C-NMF), Machine Learning, Point of Interest (POI), Call Detail Record (CDR).**

I. Introduction

Due to the highly predictable daily movements of citizens in urban areas [1], mobile traffic shows repetitive patterns with spatio-temporal variations. This phenomena is known as Tidal Effect, compared to the rise and fall of the sea levels. Tidal traffic may create regular patterns given by the human mobility making the presence of people in certain regions periodic in time. The repetitive recurrence and sometimes deterministic trends in traffic, such as peaks during busy hours and valleys during inert hours, have attracted the attention of researchers since they carry a large fraction of information and play a vital role in traffic engineering, network design, load balancing and pricing. In reality, methods for internet traffic *classification* and *identification* may be very complex. Commonly, IP traffic classification techniques have been based around direct inspection of each packet's contents in order to determine typical flows of some applications [2]. This kind of classification assumes that most applications use well known TCP or UDP port numbers. Unfortunately, the effectiveness of such packet inspection techniques is diminishing because they rely on two related assumptions: 1) third parties unaffiliated are able to inspect each IP packet's payload; 2) the classifier knows the syntax of each application's packet payloads. Newer approaches classify internet traffic by recognising statistical patterns in externally observable attributes [3] (such as typical packet lengths and inter-packet arrival times). They aim to cluster IP traffic flows into groups that have similar traffic patterns, or classify one or more applications. A basic assumption of these methods is that traffic (at the network layer) has unique statistical properties for certain classes of applications in order to partition the incoming flows.

In recent years, a lot of methods for the identification of patterns are emerging, and they are different from classification approaches. A number of researchers are looking particularly closely at the application of Machine Learning techniques (a subset of the wider Artificial Intelligence discipline) in order to *identify* trends that are frequent in the network. Methods such as K-means, Spectral Clustering, Principal Component Analysis [4] and Gaussian mixture model have been exploited to identify and extract the *basic traffic pattern* and capture the underlying traffic trend. A range of applications, as anomaly detection and load balancing, rely on basic pattern estimation, and it represents an important solution for those Internet Service Providers (ISPs) that face every day network congestion problems or over-provisioning of the link capacity.

In this paper, we propose a novel model for basic pattern identification based on matrix factorization methods. In the field of pattern recognition, the Non-negative Matrix Factorization (NMF) is one of the most used method [5][6] thanks to the ability to detect basic flows inside large matrices. Firstly we will apply the classical NMF method to a real-world dataset that collects the data traffic of mobile users at the base station level in the city of Milan. Afterwords, inspired by [7], we propose an integration of multi-domain data on the basis of NMF and denote it as Collective NMF (C-NMF) based model.

In this work, both data traffic and Point Of Interests (POIs) data will be taken into account when detecting basic patterns. The data traffic data-set comes from a challenge that Telecom Italia launched in 2014, called BIG DATA CHALLENGE [8] and it provides information about telecommunication

TABLE I
POINT OF INTERESTS

1	ADMINISTRATION
2	RELIGIOUS BUILDINGS
3	COMPANIES
4	CULTURE
5	RESIDENCES FOR SOCIAL ACTIVITIES
6	GOVERNMENT
7	TRANSPORT INFRASTRUCTURE
8	TECHNOLOGY INFRASTRUCTURE
9	ISTRUCTION
10	HEALTH
11	SOCIAL SERVICE
12	SECURITY
13	SPORT
14	TURISM
15	UNIVERSITY AND RESEARCH

activities alongs two months. The POIs data-set contains geographic locations of places of different categories present in the territory [9] (such as universities, industries, parks, clubs, etc.). When operating the C-NMF, multi-domain data will make some difference to the final result respect to the classical baseline models.

The paper is organized as follows. Section 2 identifies the proposed model for the basic pattern identification. Section 3 reports experiments and results. In the fourth part, we will report some related work. Section 5 presents the concluding remarks.

II. PROPOSED MODEL

In this section we will explore the datasets we used for the preliminary work (subsection II-A) and the proposed model that allow us to obtain typical Internet traffic patterns (subsection II-C).

A. Datasets

The original data of our research is composed by four datasets. The first one refers to the traffic of voice/sms/data of Milan city, measured during November and December 2013 [8]. It is the result of a computation over the Call Detail Records (CDRs) generated by the Telecom Italia cellular network. CDRs log the user activity for billing purposes and network management every ten minutes, creating 144 records for each day. The data-set contains the following information: Square cell ID (from 1 to 10000), Time interval, Country code, Received SMS, Sent SMS, Received Calls, Sent Calls, Internet. The second data-set contains geographic locations of different POIs categories present in the territory, indicating: Address, Latitude, Longitude and Category (see table I).

The third data-set collects information about the base stations of Telecom Italia deployed in Milan [10], such as: Base Station ID (from 1 to 1728), Latitude and Longitude. It was created by a collaborative project with the aim to realize a free worldwide database of Cell IDs and their corresponding location area identity [10]. Figure 1 shows the segmentation of Milan city obtained by the third data-set as Voronoi diagram [11]. As we can see from fig. 1, base stations are concentrated

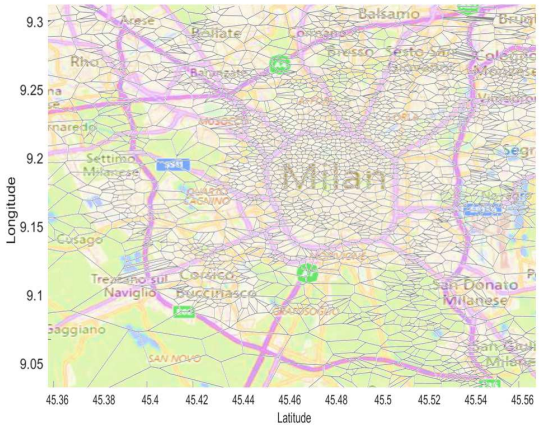

Fig. 1. Voronoi diagram on Milan city map

in the city center and along the main roads surrounding Milan. The last dataset is called DUSAF (Destinazione d'Uso dei Suoli Agricoli e forestali) map. It describes the general land usage of the entire Lombardy to monitor the changes that take place in all the region. We got the last version of 2012 [9] and used as ground truth information. We evaluated the quality of patterns by verifying their congruence with the underlying land usage. Before the research, some data pre-processing has been conducted in order to filter the CDRs data-set (considering only the information about Internet traffic), sampling hour by hour and aggregating the traffic from the 10000 squared cells to the 1728 base stations (fig.1).

B. Problem Definition

Definition 1. (Region) A region in a cell of the Voronoi map (see fig. 1) indicates the territory covered by the single Base Station.

Definition 2. (Pattern) A pattern is a time series segment whose elements repeat in a predictable manner.

Definition 3. (CDR matrix) The CDR matrix describes the number of CDR generated by the users, present in the area covered by the base station, at each hour of the day. In the CDR matrix V, where $V \in \mathbb{R}^{[N \times T]}$, each row is in the form $v_i = [v_{i,1}, v_{i,2}, v_{i,3}, ..., v_{i,N}]$ where $N = 1728$ represents the number of base stations, $T = 1488(hours)$ stands for the time interval and $v_{i,j}$ means the number of CDR generated by the base station i_{th} during the time interval j_{th}.

Definition 4. (POI matrix) POI means the point of interest in a city. They are specific places that someone may find useful or interesting. The POI matrix P, where $P \in \mathbb{R}^{[N \times M]}$ contains the POI vectors for each base station. Each row is in the form $p_i = [p_{i,1}, p_{i,2}, p_{i,3}, ..., p_{i,M}]$ where $M = 15$ represents the quantity of different type of POI (for example: schools, companies, industries), and $p_{i,j}$ is in proportion to the number of POI of kind j and inversely proportional to the distance from the POI to the base station.

Problem Definition. The main goal is to find a reliable method able to partition the traffic matrix in order to reveal similarities among base stations.

C. Collective Non-negative Matrix Factorization based Model (C-NMF)

Non-Negative Matrix Factorization (NMF) is a set of algorithms in multivariate analysis and linear algebra where a matrix V is factorized in a product of two positive sub-matrices W and H, where H represents the matrix of the basic flows and W the coefficient matrix.

Inspired by the work in [7] for discovering functional regions via matrix factorization based models, we integrated the POIs data-set to the NMF model to find more accurate basic patterns.

1) Problem Statement: The goal is to minimize the following objective function of C-NMF:

$$O(W, H_u, H_s) = \beta||V - WH_s||^2 + \alpha||P - WH_u||^2 + \lambda\left(||W||^2 + ||H_u||^2 + ||H_s||^2\right)$$

where $V \in \mathbb{R}^{[N\times T]}$ is the CDR matrix, $P \in \mathbb{R}^{[N\times M]}$ is the POI data-set, $W \in \mathbb{R}^{[N\times K]}$ represent the coefficients matrix, $H_s \in \mathbb{R}^{[K\times T]}$ with $H_u \in \mathbb{R}^{[K\times M]}$ are the basis matrix and are all non negative. The parameters α, β, λ scale the importance of the three blocks involved in the minimization. According to the objective function, each row of H_s can be denoted as basic traffic pattern, which indicates the typical behaviour of the base station. Each row of matrix W is the mixing coefficient for each base station. This means that, conducting a clustering on W, it will reveal similarities among regions (in terms of patterns) that would not have been visible from the initial CDR matrix. The last item of this formula play the role to minimize the structural risk.

2) Optimization: In order to learn the model, we applied the multiplicative updating algorithm. When optimizing the objective function, the algorithm updates the factor matrices at each iteration, until the function reaches the local/optimal minima. We derived the updating rules by applying the Kerush-Kuhn-Tucker (KKT) conditions [12].

$$W_{(i,j)} \leftarrow W_{(i,j)} \frac{\left(\alpha V H_s^T + \beta P H_u^T\right)_{(i,j)}}{\left(\alpha W\left(H_s H_s^T\right) + \beta W\left(H_u H_u^T\right) + \lambda W\right)_{(i,j)}} \quad (1)$$

$$H_{u(i,j)} \leftarrow H_{u(i,j)} \frac{\left(\alpha W^T P\right)_{(i,j)}}{\left(\alpha\left(WW^T\right) H_u + \lambda H_u\right)_{(i,j)}} \quad (2)$$

$$H_{s(i,j)} \leftarrow H_{s(i,j)} \frac{\left(\beta W^T V\right)_{(i,j)}}{\left(\beta\left(WW^T\right) H_s + \lambda H_s\right)_{(i,j)}} \quad (3)$$

By using these updating rules the algorithm will converge in limited time.

III. Experiments

A. Experiment setup

The experiment setup follows two procedures. First we applied two clustering algorithms, Kmeans and Spectral Clustering [13] [14], and later the NMF and C-NMF models. These two procedures follow differents steps:

TABLE II
Clustering indexes: *Davies-Bouldin (D&B)*, *Calinski Harabasz (CH)* and *Dunn*.

	D&B	**CH**	**DUNN**
KMEANS	1.22(12)	8894(8)	0.028(8)
SPECTRAL CLUSTERING	2.85(8)	6.92(15)	$1.93 \cdot 10^{-7}$(29)
NMF	0.91(38)	798(4)	0.031(38)
C-NMF	1.29(30)	1663(5)	0.069(52)

Procedure 1

1) Apply the Kmeans and Spectral Clustering to the matrix V (see section II-B), with number of clusters between 5 and 100.
2) Evaluate the clustering results deriving three clustering indexes: *Davies-Bouldin (D&B)*, *Calinski Harabasz (CH)* and *Dunn*, explained in section III-B.
3) Choose the correct number of clusters based on the clustering indexes.
4) Evaluate the clusters through the DUSAF map as ground truth information.

Procedure 2

1) Apply the NMF and C-NMF models to the matrix V, with different value of decomposition factor (k), deriving the sub-matrix W.
2) Cluster the matrix W with the Kmeans method, with number of clusters between 5 and 100.
3) Evaluate the clustering results deriving three clustering indexes: *Davies-Bouldin (D&B)*, *Calinski Harabasz (CH)* and *Dunn*, explained in section III-B.
4) Choose the correct number of clusters based on the clustering indexes.
5) Evaluate the clusters through the DUSAF map as ground truth information.

B. Clustering indexes

Calinski Harabasz (CH): called sometimes variance ratio criterion (VRC), evaluates the cluster validity based on the average between and within cluster sum of squares. Well-defined clusters have a large between-cluster variance and a small within-cluster variance.

Davies-Bouldin (D&B): it is based on a ratio of within-cluster and between cluster distances. For each cluster C, the similarities between C and all the other clusters are computed, and the highest value is assigned to C as its cluster similarity. The index is obtained by averaging all clusters similarities. So, we are looking for the smallest index.

Dunn: it is an internal evaluation scheme, where the result is based on the clustered data itself. The aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated. We are looking for the maximum value.

C. Experiment results

We evaluated the performance of the four methods on two fronts. First we calculated the accuracy of the clustering by

three parameters, see Table II (where the number next to the index is the number of clusters), and later, thanks to the DUSAF map, we looked directly at the areas covered by some clusters through the map of the city (fig.2). We are looking the smallest value of D&B and the highest one for CH and DUNN. In particular, we have seen that Kmeans and Spectral Clustering methods do not converge for any of the three parameters in a global minimum/maximum, but just in local ones. The number of clusters is around 10 and the DUSAF map does not give any information regarding the association between the patterns and land usage, the number of clusters should be more than 10.

The situation changes for NMF and C-NMF methods, because we found global minimum/maximum, making the evaluation more reliable. Even if C-NMF seems to be more accurate (according to CH and DUNN parameter), both NMF and C-NMF converge almost to the same number of clusters but obviously grouped differently. In this case we can see that the number of clusters is increased compared to the cases of classic clustering methods on the original matrix. This means that there are small clusters but distinguishing. In this article we show some of them obtained by the last method. Looking at the DUSAF map we have noticed that the last method gives a grouping of base stations that not only look alike in terms of patterns, but also from the point of view of land usage. Indeed more than 60% of base stations that compose each of the 5 clusters we analysed, converge with the land usage showed in the DUSAF map. In particular, we can see that the suburban patterns are fairly regular both for weekdays that for the weekends, and without any particular behavior (see fig. 3).

In fact, they are typical of large residential facilities and farms in the countryside. As for the town the situation changes. The algorithm is able to extract some typical pattern of certain well-defined areas that present same kind of POIs. Figure 4 shows the mean typical pattern of parks and touristic places, for weekdays and weekends. We can see that the traffic peaks recorded at noon and in the late afternoon are compatible with the habits of the people who frequent the parks. Figure 5 represents the typical patterns observed in the main roads and industries. The peak of the morning perfectly matches the typical traffic jam of large cities. Figure 6 and 7 are very interesting. The first one shows the patterns found around areas with commercial and restaurants facilities. In fact, the shape of the curve during the weekdays and weekends matches the mobility patterns of the people. Instead, the second figure shows a cluster where the city's night-life is very lively. In fact, this is typical of clubs and places where students and young people are used to spend their evenings. In the end, the last figure 8 reveals the patterns found nearby schools and universities.

These results confirm the presence of regular and periodic traffic due to the tidal effect experienced by the network. In fact, it depends not only on user habits but also by the land usage and POIs scattered in the city area.

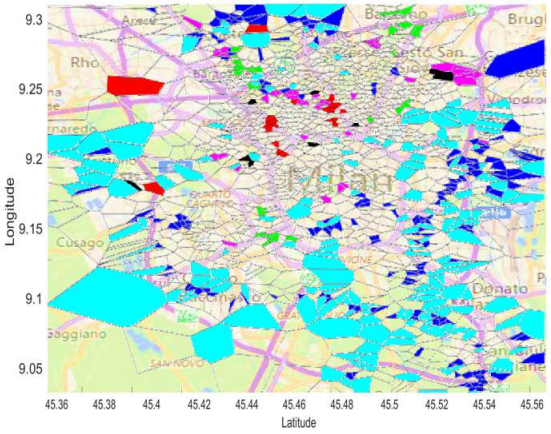

Fig. 2. Milan city map with some example clusters.

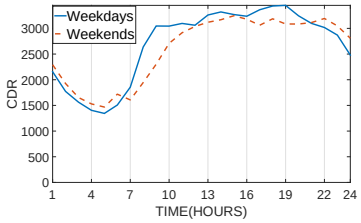

Fig. 3. Typical of the suburbs (LIGHT BLUE regions in fig. 2).

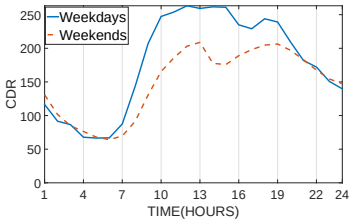

Fig. 4. Typical of Parks, Gardens and Touristic places (PINK regions in fig. 2).

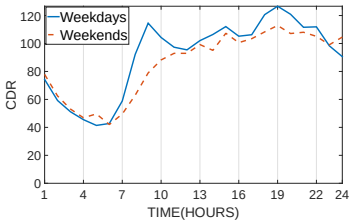

Fig. 5. Typical of Main roads and industrial places (BLUE regions in fig. 2).

IV. Related Work

In this section we will understand the reasons that led us to carry out this work. [1] showed that there is a high predictability (from 80% to 93%) of the user mobility thanks to the regularity of human behaviour. It makes the mobile traffic highly variable and therefore an issue for those Internet Service Providers that engaged in a static resource planning.

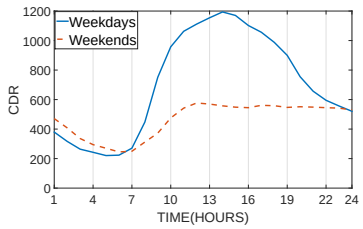

Fig. 6. Typical of Commercial, Restaurants and Companies area (RED regions in fig. 2).

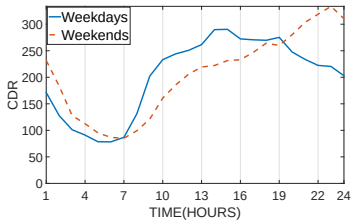

Fig. 7. Typical of Clubs and Theatres (BLACK regions in fig. 2).

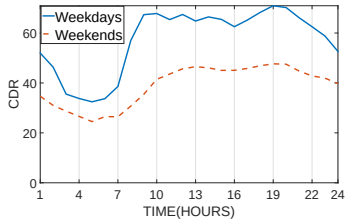

Fig. 8. Typical of Schools and Universities (GREEN regions in fig. 2).

It turns out that revealing the basic patterns of the traffic data at the base station level, can represent an important solution for those providers that face network congestion problems or over-provisioning of the link capacity. Then, we investigated previous work on traffic data classification and identification in [2],[3] and [4]. They deal with different methods, from statistical to principal component analysis approaches, obtaining results with high accuracy. Then, we addressed the problem by exploiting matrix factorization methods as showed in [5], [6] and [7], where these methods are widely studied. Authors in [15] investigate the heterogeneous patterns emerging in the mobile communication activity recorded within metropolitan regions. A. Furno *et al*, introduced an original technique to identify classes of mobile traffic signatures that are distinctive of different urban fabrics, showing that it creates mobile demand profiles that better agree with land use ground-truth data.

V. Conclusion

In this article, we addressed the problem to find typical traffic data pattern in the network by exploiting matrix factorization methods. The main goal is to characterize the internet traffic in order to better describe the tidal effect that occurs in the metro network with the aim to optimize the resources allocation and avoid network congestion. To achieve this goal we have seen how NMF can be improved by adding more information on the data, for example POIs, and therefore it is a great tool that allows us to better analyze the network traffic. Thanks to the spread of social networks we have many sources from which to draw, in order to tie user behavior with the internet traffic management and then improve internet services. In addition, the identification of traffic patterns could be improved by aggregating other information coming from the principal social networks, such as Facebook or Twitter, and understand why the network experiences unexpected peaks of traffic data during the day. The collection of these information could be used for many purposes, one of them is the prediction of the traffic load on-demand avoiding bottlenecks and providing dynamic allocation of network resources.

Acknowledgment

This work was supported by the EU FP7 IRSES Marie Curie MobileCloud project that has goal to foster research collaborations between Chinese and European university in Mobile Cloud Computing. The data-set of POIs was provided by the municipality of Milan and we thank them for their cooperation and commitment.

References

[1] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert Lszl Barabsi. Limits of predictability in human mobility. Science, 327(5968):10181021, 2010.
[2] Nguyen, Thuy TT, and Grenville Armitage. A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials 10.4 (2008): 56-76.
[3] Jaiswal, Rupesh Chandrakant, and Shashikant D. Lokhande. Machine learning based internet traffic recognition with statistical approach. 2013 Annual IEEE India Conference (INDICON). IEEE, 2013.
[4] Bandara, Vidarshana W., and Anura P. Jayasumana. Extracting baseline patterns in Internet traffic using Robust Principal Components. Local Computer Networks (LCN), 2011 IEEE 36th Conference on. IEEE, 2011.
[5] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556562, 2001.
[6] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. Journal of machine learning research, 5(Nov):14571469, 2004. 105
[7] Wang Shan, Xu Yajing, Gao Sheng. Revealing Functional Regions via Joint Matrix Factorization Based Model. Proceedings of NIDC2016
[8] https://dandelion.eu/datamine/open-big-data/
[9] http://www.territorio.regione.lombardia.it/
[10] http://opencellid.org/
[11] http://it.mathworks.com/help/matlab/math/voronoi-diagrams.html
[12] Hyunsoo Kim and Haesun Park. Non-Negative Matrix Factorization based on alternating Non-Negativity constrained least squares and active set method.
[13] Fan RK Chung. Spectral graph theory, volume 92. American Mathematical Soc., 1997.
[14] Bojan Mohar. Some applications of laplace eigenvalues of graphs. In Graph symmetry, pages 225275. Springer, 1997.
[15] A. Furno; M. Fiore; R. Stanica; C. Ziemlicki; Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas," in IEEE Transactions on Mobile Computing , vol.PP, no.99, pp.1-1 doi: 10.1109/TMC.2016.2637901