Modelling and Characterization of Customer Behavior in Cellular Networks

Thomas Couronné
Orange Labs
France Telecom R&D,
Paris, France
Email:Thomas.Couronne@orange-ftgroup.com

Valery Kirzner
Institute of Evolution
University of Haifa
Haifa, Israel
Email:valery@research.haifa.ac.il

Katerina Korenblat
Software Engineering Department
Ort Braude College
Karmiel, Israel
Email:katerina@braude.ac.il

Elena V. Ravve
Software Engineering Department
Ort Braude College
Karmiel, Israel
Email:cselena@braude.ac.il

Zeev Volkovich
Software Engineering Department
Ort Braude College
Karmiel, Israel
Email:vlvolkov@braude.ac.il

Abstract—**In this paper, we extend a model of the fundamental user profiles, developed in our previous works. We explore customer behavior in cellular networks. The study is based on investigation of activities of millions of customers of Orange, France. We propose a way of decomposition of the observed distributions according to certain external criteria. We analyze distribution of customers, having the same number of calls during a fixed period. A segmentation of the population is provided by an approximation of the considered distribution by means of a mixture of several more "basic" distributions presenting the "granularity" of the user's activity. In order to examine the meaning of the found approximation, a clustering of the customers is provided using their daily activity, and a new clustering procedure is constructed. The optimal number of clusters turned out to be three. The approximation is reduced in the optimal partition to a single-exponential one in one of the clusters and to two double-exponential in others. This fact confirms that the proposed partition corresponds to reliable consequential social groups.**

Keywords*–*Customer behavior pattern; Market segmentation; Probability distribution; Mixture distribution model; Machine learning; Unsupervised classification; Clustering.

I. Introduction

This paper presents an extended and improved version of [1], where we introduced the general framework of modelling behavior patterns in cellular networks.

Customer behavior is a way people, groups and companies purchase, operate with and organize goods, services, ideas and knowledge in order to suit to their needs and wants [2], [3]. Multidisciplinary studies of the customer behavior strive to comprehend the decision-making processes of customers and serve as a basis for market segmentation. Through market segmentation, large mixed markets are partitioned into smaller sufficiently homogeneous sectors having similar needs, wants, or demand characteristics.

In the cellular networks context, the mentioned products and services can be expressed in spending of the networks resources such as the number of calls, SMS messages and bandwidth. Market segmentation in this area is able to characterize behavior usage or preferences for each sector of customers. In other words, typifying of the customers‘ profiles is aimed at using this pattern in order to suitably adapt specific products and services to the clients in each market segment.

A segmentation of the population is provided by an approximation of the considered distribution by means of a mixture of several more "basic" distributions, which represent the "granularity" of the user's activity. Such mixture distribution models are conventional in machine learning due to their fruitful applications in unsupervised classification (clustering). In this framework, the underling probability distribution is decomposed into a mixture of several simple ones, which correspond to subgroups (clusters) with high inner homogeneity. In our application, hypothetically, each one of these clusters corresponds to a social group of users, having its own dynamics of calls depending upon the individual group social parameters.

The common applications of the known Expectation Maximization algorithm [4], which estimates parameters of the mixture models (for instance, in the clustering), suggest the Gaussian Mixture Model of the data. This well-understood technique is much admired because it satisfies a monotonic convergence property and can be easily implemented. Nevertheless, there are several known drawbacks. If there are multiple maxima, the algorithm may discover a local maximum, which is not a global one. In addition, the obtained solution strongly depends on the initial values [5]. Moreover, many studies are recently devoted to analysis of non-Gaussian processes, which are often related to the power law distributions.

While in clustering as a rule the Gaussian Mixture Model of the data is assumed, we treat the user activity in a cellular network as a mechanism generating *non-Gaussian* distributions.

In physics, hyperbolic dependencies are often observed (e.g., theories during phase transitions that clarify the corresponding mechanism). On the other hand, there are a number of general formal models (for example, the law of Yule [6]), where such a distribution appears. In these models, hyperbolic behavior is often observed as asymptotic or applicable to certain parts of the distribution.

Our research develops *a novel model of the fundamental user behavior patterns* (user profiles) in cellular networks. We adopt the standard simple regression methodology of [7] to our purposes. We show that empirical densities of the studied underlying distributions are monotone decreasing and do not exhibit multi-modality. These properties characterize mixtures of *the exponential distribution* [8], [9]. In this sense, we extend the study of [10], where it was shown that a parallel user activity in recording in an email address book leads to an appropriate exponential distribution of the clients.

In order to explore the meaning of the found approximation, a clustering of the customers is provided, based on their daily activity, and a *new clustering procedure* is constructed in the spirit of the bi-clustering methodology of [11], [12].

We base our study on analysis of the underlying distribution of customers, who have the same number of calls during a fixed period, say a day. In this research, an exponential distribution mixture model is applied. It is shown that a three-exponential distribution fits well the needed target.

The estimated optimal number of clusters turned out to be three. A straightforward clustering of the original data is hardly expected to deliver a robust and meaningful partition. Such a situation is a common place in the current practice. Moreover, in many applications the aim is to reveal not merely potential clusters, but also a quite small number of variables, which adequately settle that partition. For instance, the sparse K-means (SK-means) proposed in [13], at once discovers the clusters and the key clustering variables.

A *new procedure* in the spirit of such a bi-clustering methodology, where features and items are simultaneously clustered, is applied in this paper. Firstly, 24 hours inside a day (the features) are clustered consistent with the corresponding users' activity. In the next step, the users are divided in groups according to their occurrences in the previous partition of hours. As a result, a sufficiently robust clustering of users is obtained together with a description of the clusters in terms of call activity.

The observed dissimilarity between hours (from the point of view of the users' behavior) can be naturally characterized by a distance between the corresponding distributions. In this paper, we employ Kolmogorov-Smirnov two sample test statistic [14], [15], which is actually the maximal distance between two empirical normalized cumulative distribution functions.

Then we use the Partitioning Around Medoids clustering algorithm [16], in order to cluster the data. This algorithm operates with a distance matrix, but not with the items themselves. This is feasible for small data sets (such as the considered one, composed of 24 hours) and a small number of clusters three in our case. One of the input parameters of the algorithm is the number of clusters. We estimate the optimal number of their hours clusters, using the Silhouette coefficient of [17]. Here ideas of both cohesion and separation are combined: for individual points as well as for partitions. The number of clusters was checked in the interval of $[2-10]$, and the optimal one was found to be 3 for all the considered data sets.

When we obtain classification of users' activity across the hour clusters, we built a vector, composed of the fractions of calls falling within each hour cluster. At this stage, we produce user clustering employing this new data representation. Note that, due to the large amount of data, we deal here with a high complexity clustering task. It means that the traditional clustering algorithms cannot be directly applied to this situation. In order to resolve this problem we apply a resampling clustering procedure, according to which the whole data set is partitioned based on clustering of its samples.

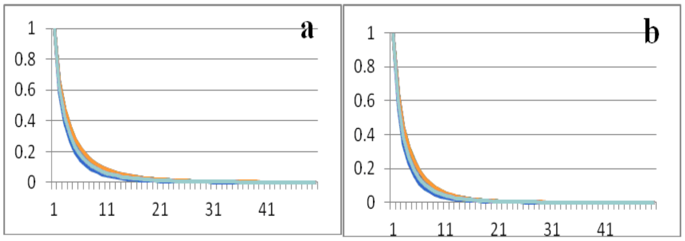


Figure 1. DSN curves for $InCalls$ (a) and $OutCalls$ (b), obtained on each day of the whole period of observation (13 days).

Finally, we obtain the optimal partition with a single-exponential call distribution in one of the clusters and two double-exponential call distributions in others. This fact confirms that the proposed partition corresponds to reliable consequential social groups. We emphasize the fact that the similarity measure, applied in the clustering process, is formed without any reference to the previously discussed mixture model.

The results, presented in the paper, are obtained by means of a study of the daily activity of a real group of users during the period from March 31, 2009 through April 11, 2009. For each considered day, several million users in this group are active (making one or more calls). The time of each input or output call is known. Note that the sets of active users on different days vary.

The paper is structured as follows. Section II is devoted to a distribution model of user activity and its decomposition. Section III describes the customer clustering procedure and its evaluations. The section presents model-based evaluation of the proposed customer classification. Section IV summarizes the paper and provides outlook.

II. Distribution model of user activity

In this section, we consider a mixture model approximation of the underlying distribution of users having the same number of calls during a day. We denote by DSN the *Day Same Number* distribution. We distinguish two types of user activity: input calls, denoted by $InCalls$, and output calls, denoted by $OutCalls$.

All users (about five millions) are divided into groups according to their number of calls per day, so that the i-th group contains all customers having exactly i calls per a particular day. The size of the i-th group is denoted by N_i. Obviously, the contents and sizes of the groups are not the same for different days. In addition, the number of groups with $i > 100$ is very small in the dataset, and these groups most likely contain "non-standard" users such that sales agents, call centers and so on. We omit such groups together with users, who do not call at all in a given day, since this lack of activity could be explained by factors that are not directly related to the user activity on the network.

The DSN curves, normalized to 1, of $InCalls$ (a) and $OutCalls$ (b) as well as the corresponding numbers of calls ranging from 1 to 50, are shown in Fig. 1. We note that the curves are of almost the same monotonically decreasing form for all 13 days of the observation.

TABLE I. p-values for different numbers of components

Number of components	1	2	**3**	4
p-value	0	$8.6e-06$	**0.025**	0.282

As it was mentioned in Section I, a mixture distribution model with exponential components seems to be an appropriate approximation of DSN. In our context, it is natural to assume that the underlying population is actually a mix of several different sub-populations. Mixture distribution models appear in many applications as an inherent and straightforward tool to pattern population heterogeneity. The assumption about *exponentially distributed* components of the mixture is commonly invoked in the study of lifetime or more universal duration data. Here you have a simple k-finite exponential mixture model, having a density function of the following form

$$f(x) = \sum_{j=1}^{k} A_j exp(-t_j \cdot x), \tag{1}$$

where A_j and t_j, $j = 1,...k$ are nonnegative numbers, and $\sum_{j=1}^{k} A_j = 1$. For a given number of components k, the Expectation–Maximization (EM) algorithm [4] is a traditional method for maximum likelihood estimation of finite mixtures.

However, we apply another approach in the spirit of the linear regression methodology without any prior assumption about k - the number of components. For this purpose, we initially form explanatory variable $X = (1, 2, ..., 100)$ and response Y. For each value $x \in X$, $Y = ln\,(f(x))$, where $f(x)$ is the normalized frequencies of DSN in a day.

Using the standard simple regression methodology of [7], a linear regression model is identified as $Y = a + b \cdot X$ and the first estimation of the density $f(x)$ in (1) is constructed: $f^{(1)}(x) = A_1 \exp(-t_1 \cdot x)$, for $A_1 = \exp(a)$ and $t_1 = -b$. In the next step, the new response is built: $Y = ln\left(f(x) - f^{(1)}(x) + C\right)$, where C is a sufficiently big positive number insuring that $f(x) - f^{(1)}(x) + C > 0$ for all x and j. In each step, p-value coefficient of significance:

$$F = \frac{R^2(X,Y)}{1 - R^2(X,Y)}(100 - 1) \tag{2}$$

is calculated. Here $R(X,Y)$ is the Pearson correlation coefficient between X and Y [18]. The described procedure is repeated until the actual p-value is less than the traditional level of significance 0.05. In our particular application, for all cases of daily activity, the procedure has been stopped after three components were extracted.

The parameters of (1), calculated for each of the 13 days of the observation, are presented in Tables II and III. They demonstrate high stability of the exponent indexes t_1, t_2, t_3, which are practically independent of time but are somewhat different on the weekends, i.e., Saturdays 4 and 11 of April 2009 and Sunday 5 of April 2009. Amplitudes A_1, A_2, A_3 differ to a greater degree (in percentage terms). Thus, the absolute number of active users varies from day to day to a greater extent than the distribution pattern, which actually corresponds to a set of exponent indexes. The p-values, calculated for the first of the considered days, are presented in Table I.

TABLE II. Values of the approximation function parameters on different days for $InCalls$. (The designations of the amplitudes (A) and indexes (t) correspond to (1))

Dates	A_1	t_1	A_2	t_2	A_3	t_3
03_30	80001	0.12	399893	0.32	420568	1.02
03_31	110555	0.12	441268	0.33	380258	1.15
04_01	94021	0.11	421456	0.31	401002	1.01
04_02	99683	0.11	419564	0.3	411258	1.05
04_03	96660	0.11	409176	0.29	405050	0.98
04_04	90424	0.12	406385	0.34	420161	1.07
04_05	59971	0.12	399873	0.36	530064	1.08
04_06	91189	0.11	425022	0.31	450957	1.04
04_07	83467	0.11	415012	0.3	431301	0.96
04_08	93358	0.11	430842	0.31	422297	1
04_09	102169	0.11	426124	0.31	416794	1.07
04_10	97814	0.11	402832	0.3	408717	1
04_11	65206	0.11	353998	0.33	439797	1.01

TABLE III. Values of the approximation function parameters on different days for $OutCalls$. (The designations of the amplitudes (A) and indexes (t) correspond to (1))

Dates	A_1	t_1	A_2	t_2	A_3	t_3
03_30	100684	0.16	561222	0.37	527907	1.25
03_31	119660	0.16	560344	0.36	514682	1.32
04_01	116329	0.16	564578	0.35	498085	1.32
04_02	118910	0.16	546314	0.35	494688	1.27
04_03	130193	0.16	538984	0.34	497177	1.3
04_04	95354	0.16	524779	0.39	548041	1.34
04_05	87109	0.17	522660	0.46	617407	1.41
04_06	110086	0.16	562064	0.36	548389	1.34
04_07	102030	0.15	560233	0.35	497784	1.22
04_08	90481	0.15	568191	0.34	510487	1.21
04_09	115820	0.16	543334	0.34	505349	1.26
04_10	121782	0.16	518418	0.34	500068	1.24
04_11	80915	0.15	445910	0.39	538691	1.22

In the case of $InCalls$ (Table II), the ratio of the exponent indexes is: $3{\cdot}t_1 \approx t_2, 3{\cdot}t_2 \approx t_3$. In the case of $OutCalls$ (Table III), this ratio is somewhat different: $2 \cdot t_1 \approx t_2, 3.5 \cdot t_2 \approx t_3$. The decay value x_0 of each component in (1) is chosen in order to normalize the component value at this point to 1.

The components are not equivalent in the sense of their decay value (see Table IV). In fact, the exponent with index $t_3 = 1.0$ and amplitude $A_3 = 500,000$ (these values are typical for one of the three exponents, which describe the daily activity) already decays at $x_0 = 13$. For the second typical pair of the values: $t_2 = 0.33$ and $A_2 = 400,000$, the decay occurs at $x_0 = 39$. Moreover, the exponent with $t_1 = 0.12$ and $A_1 = 90,000$ has the longest effect on DSN: $x_0 = 95$. Accordingly, two of the three components, which describe the user activity, disappear in the middle of the considered interval of calls. Only the third exponent continues, and its values may be considered as the ”asymptotic behavior” of the distribution.

The relatively complex nature of the obtained empirical distribution model of user activity may indicate the heterogeneity of the entire set of the users. This set is conceivably composed of a few groups such that the total user activity in a group is described by a certain simpler distribution.

TABLE IV. Decay value for each component of DSN on different days. Columns 1, 2, 3 show decay values, x_0, of the corresponding components.

Date	$InCalls$			$OutCalls$		
	1	2	3	1	2	3
03_30	94	40	12	71	35	10
03_31	96	39	11	73	36	9
04_01	104	41	12	72	37	9
04_02	104	43	12	73	37	10
04_03	104	44	13	73	38	10
04_04	95	37	12	71	33	9
04_05	91	35	12	66	28	9
04_06	103	41	12	72	36	9
04_07	103	43	13	76	37	10
04_08	104	41	12	76	38	10
04_09	104	41	12	72	38	10
04_10	104	43	12	73	38	10
04_11	100	38	12	75	33	10

Obviously, the social status, gender and age of the users affect their activity on telephone networks. However, such types of personal data are not available for us. Therefore, in the following section, we divide the users into groups, based merely on the features of their individual activity during a given day. It is assumed that these features are related to some of the social characteristics of the users. A justification for this assumption may be found, for example in [19].

III. User classification

We assume that the obtained three-component exponential mixture model reflects the inner customers' behavior patterns, exposed by the observed data. In order to identify these patterns, all the users are divided into groups according to a comparable daily performance. In this way, analysis of the overall cluster behavior can characterize the corresponding pattern.

We apply a procedure in the spirit of the bi-clustering methodology of [13]. First of all, we cluster 24 hours inside a day (the features) according to the corresponding users' activity. Then, the users are divided in groups according to their occurrences in the hour's partition. As a result, a sufficiently robust clustering of users is obtained together with a description of the clusters in terms of call activity.

A. Clustering of hours

First of all, we try to outline a similarity between hours in a day. For this purpose, we consider each hour as a distribution of users across the actual numbers of calls within this hour. It means that we examine how many people did not call at all in this hour, how many people called just one time, two times and so on.

The observed dissimilarity between hours can be naturally characterized by a distance between the corresponding distributions. Generally speaking, any asymptotically distribution-free statistic is suitable for this purpose. In fact, the distribution of an asymptotically distribution-free statistic does not depend on the underlying distribution of the populations for samples of sufficiently large size. Here, we employ the well-known Kolmogorov-Smirnov (KS) two sample test statistic [14], [15].

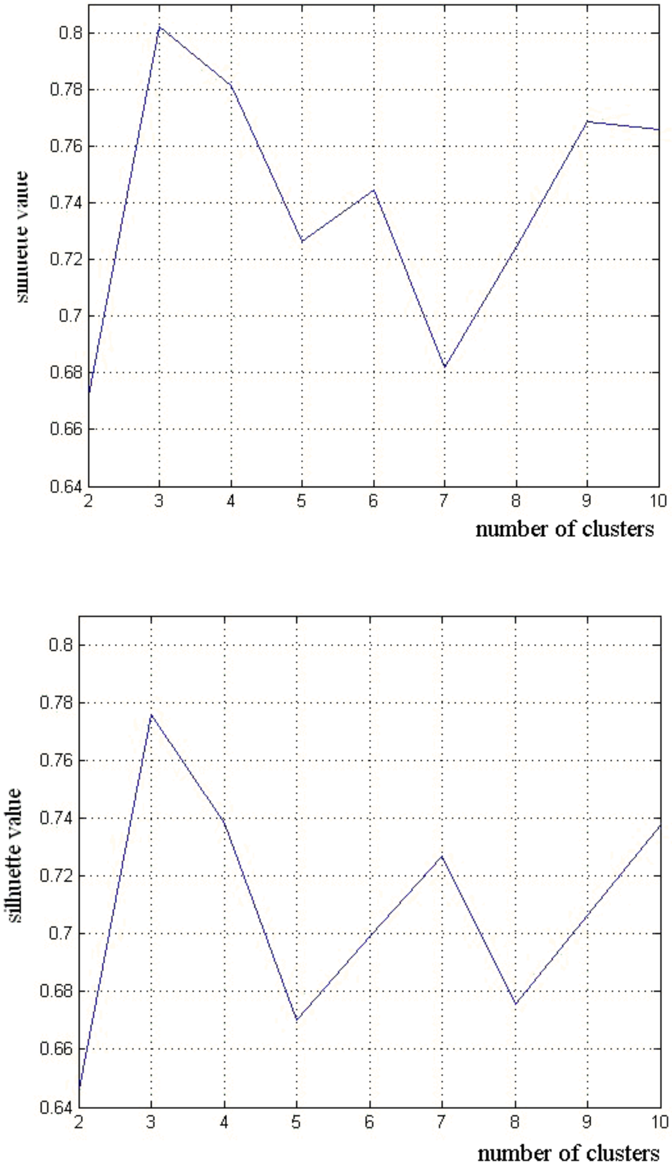


Figure 2. Silhouette plots for April 5 and April 10

Calculating the KS-distance for each pair of hours, we get a 24×24 distance matrix. The Partitioning Around Medoids (PAM) clustering algorithm [16] is applied now to cluster the data. In order to divide a data set into k clusters using PAM, firstly, k objects from the data are chosen as initial cluster centers (medoids) with the intention to attain the minimal total scattering around them (to reduce the loss function value). Then, the procedure iteratively replaces each one of these center points by non-center ones with the same purpose. If no one of further changes can improve the value of the loss function then the procedure ends.

In addition to the clustered data, PAM requires as an input parameter the number of clusters k. Hence, the first step of our procedure is devoted to estimation of the optimal number of the hour's clusters. For this purpose, we use the

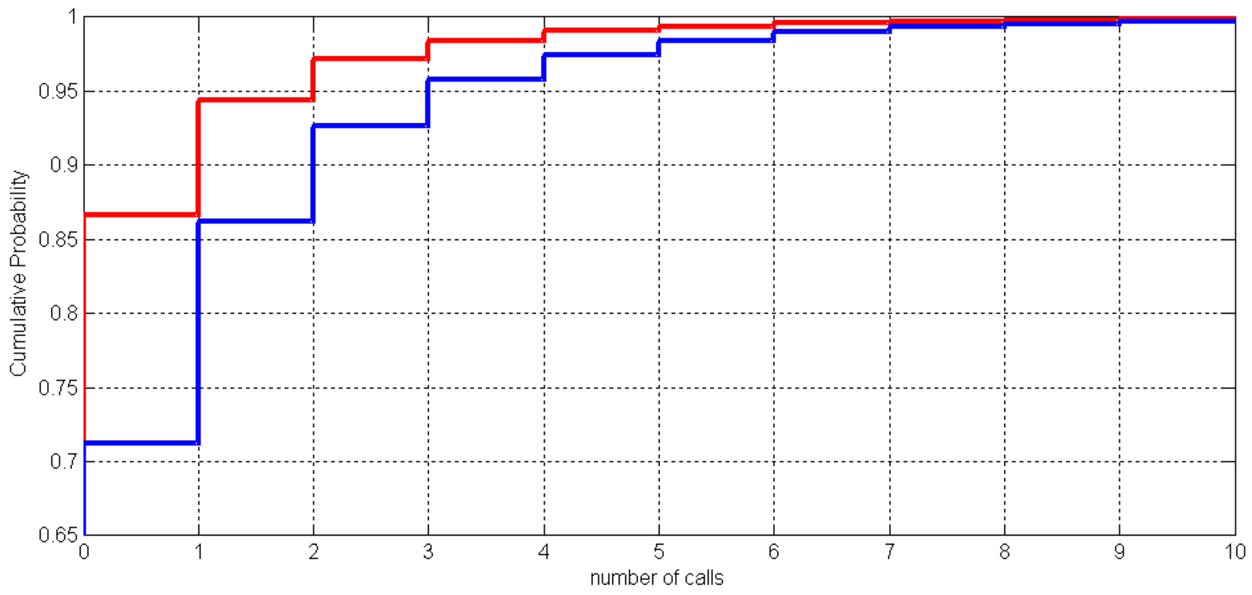


Figure 3. Empirical cumulative distribution functions for hours 9 (upper) and 19 (lower).

Silhouette coefficient of [17]. For each point, the Silhouette index takes values in $[-1, 1]$ interval, if that the Silhouette mean value, calculated across the whole data, is close to 1 specifies ”well clustered” data, and value -1 characterizes a very ”poor” clustering solution. Therefore, the Silhouette mean value, found for several different numbers of clusters, can indicate the most appropriate number of clusters by its maximal value. The number of clusters was checked in the interval of $[2-10]$, and the optimal one was found to be 3 for all the considered data sets (i.e., for all considered days). An example of Silhouette plots (for 5 of April and 10 of April) is shown in Fig. 2. Fig. 3 presents examples of two different normalized cumulative distribution curves, calculated for hours 9 and 19.

The partition of 24 hours into 3 hour clusters is presented in Fig. 4 for three different dates. It can be concluded that although the partitions slightly depend on the particular data set (date), the overall structure of the clusters is preserved. Namely, there is a silent ’night’ cluster (red), an active ’day’ cluster (blue), and a ’morning/evening’ cluster (green). Table V shows the same distribution of 24 hours with another color convention: ’night’ cluster (dark gray), ’morning/evening’ cluster (light gray) and ’day’ cluster (white) for all the considered dates. The procedure was successfully applied to our data sets, which contains information on the activity of about 5 million users during the period of observation: 13 days. We recall that only active users, those having at least one, are considered in our procedure. Based on the results of this clustering of hours, we can obtain information from the original data regarding the user activity during those hours, which correspond to the clusters.

B. *Clustering of users*

We obtained classification of users’ activity across the hour clusters. Apparently, a user can move from cluster to cluster, for example, in case when the corresponding SIM card is transferred to another person like a family member. However, as it was mentioned, the clustering structure is very similar for different working days, e.g., the most of the users do not change their behavior in a cellular network.

Now, for each user we built a vector, composed of the fractions of calls falling within each hour cluster. We produce user clustering employing this new data representation. Due to the large amount of data, we are dealing here with a high complexity clustering task. We apply a resampling clustering procedure. User behavior patterns are obtained from analysis of the users falling within a certain cluster. In this section, we describe the proposed classification procedure and its results.

1) Clusterization procedure: The main aim of the users clustering procedure is to divide the clients into groups, using information about their activity in each one of the hour clusters, obtained in the previous stage. We present each user as three dimensional vector (r_1, r_2, r_3), where r_i is the ratio of a user’s activity during a cluster of hours number i. More precisely, r_i is a fraction of a user’s calls during the cluster i in the total number of calls during a day.

The proposed resampling clustering procedure is based on the well-known K-means algorithm [20], and implementing de-facto the idea, proposed in [21]. The K-means algorithm has two input parameters: the number of clusters k and the data set to be clustered X. It strives to find a partition $\pi(X) = \{\pi_1(X), \dots, \pi_k(X)\}$ minimizing the following loss function

$$\rho_{\{c_1,\dots c_k\}}(\pi(X)) = \frac{1}{N}\sum_{j=1}^{k}\sum_{x\in\pi_j(X)}\|x-c_j\|^2\,, \qquad (3)$$

where $c_j,\ j = 1, ..., k$ is the mean position (the cluster centroid) of the objects belonging to cluster $\pi_j(X)$, and N is the size of X.

TABLE V. 24 hours partition into dark gray (night), light gray (morning/evening) and white (day) clusters for different dates.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
30.03	2	2	2	2	2	2	2	2	1	1	1	3	3	3	3	3	3	3	3	3	3	1	1	2
31.03	2	2	2	2	2	2	2	2	1	1	1	3	3	3	1	3	3	3	3	3	3	1	1	2
01.04	2	2	2	2	2	2	2	2	1	1	3	3	3	3	3	3	3	3	3	3	3	1	1	1
02.04	2	2	2	2	2	2	2	2	1	1	1	1	3	3	1	1	3	3	3	3	3	1	1	2
03.04	2	2	2	2	2	2	2	2	1	1	1	3	3	3	3	3	3	3	3	3	3	1	1	2
04.04	2	2	2	2	2	2	2	2	2	1	3	3	3	3	3	3	3	3	3	3	3	1	1	1
05.04	1	2	2	2	2	2	2	2	1	1	3	3	3	3	3	3	3	3	3	3	3	3	1	1
06.04	2	2	2	2	2	2	2	2	1	1	1	3	3	3	3	3	3	3	3	3	3	1	1	2
07.04	2	2	2	2	2	2	2	2	1	1	1	3	3	3	1	3	3	3	3	3	3	1	1	2
08.04	2	2	2	2	2	2	2	2	1	1	3	3	3	3	3	3	3	3	3	3	3	1	1	1
09.04	2	2	2	2	2	2	2	2	1	1	1	3	3	3	3	3	3	3	3	3	3	1	1	2
10.04	2	2	2	2	2	2	2	2	1	1	3	3	3	3	3	3	3	3	3	3	3	1	1	1
11.04	2	2	2	2	2	2	2	2	1	1	1	3	3	3	3	3	3	3	3	3	3	1	1	2

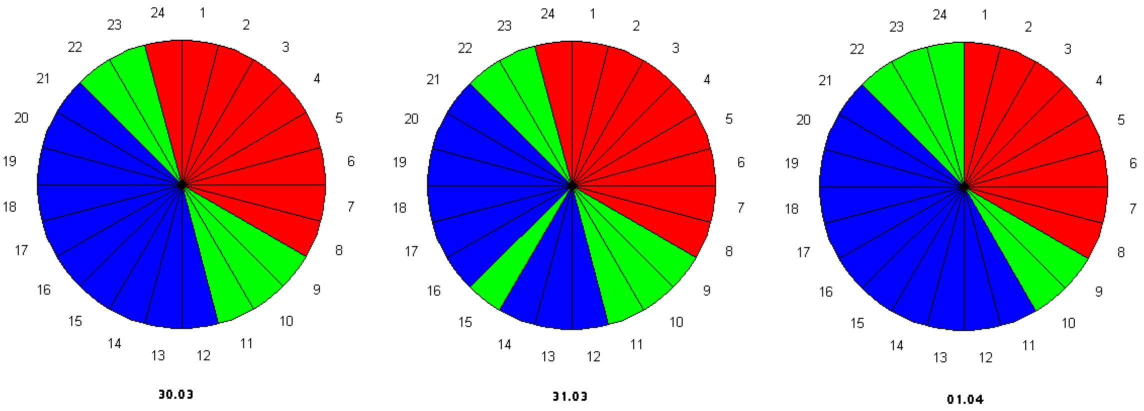


Figure 4. 24-hour partition for Mars 30, Mars 31, April 01: 'night' cluster (red), 'day' cluster (blue), and 'morning/evening' cluster (green).

Initially, the centroid set can be predefined or chosen randomly. Using the current centroid set, the K-means algorithm assigns each point to the nearest centroid aiming to form the current clusters and then recalculates centroids as the clusters means. The process is reiterated until the centroids are stabilized. In the general case, as a result of this procedure, the objective function (3) reaches its local minimum. Note that in the K-means algorithm, a partition is unambiguously defined by the centroid set and vise versa. Moreover, in the general case, the loss function (3) can be used for assessing the quality of arbitrary partition $\widehat{\pi}(X)$ with respect to given centroid set $\{c_1, \dots c_k\}$. The resampling procedure allows partitioning a large data set based on partitioning its parts as presented in Algorithm 1.

2) *Choosing the number of users clusters:* In order to evaluate the optimal number of clusters, usually, one compares stability of the obtained partition for different numbers of clusters. To this aim, we repeat the users' clustering procedure ten times on the same data set and evaluate the Rand index value between all obtained partitions. The Rand index [22] represents the measure of similarity between two partitions. It is calculated by counting the pairs of samples, which are assigned to the same or to different clusters in these partitions. The closeness of the Rand index value to 1 indicates similarity of the considered partitions.

For the same purpose, the Adjusted Rand index of [23], which is the corrected-for-chance version of the Rand index, can be used as well. However, in our consideration, it is more suitable to use the regular one because it still reflects well the closeness of the partitions. The mean values of the obtained Rand indexes naturally characterize stability of the partition by the maximal value. So, the "true" number of clusters corresponds to the most stable partition.

C. *Experimental study*

In this section, we are mostly concentrated on the data set for April 1, which is taken as a typical example of the original data sets. The results (obtained for other data sets) are very similar including all the parameters, considered below.

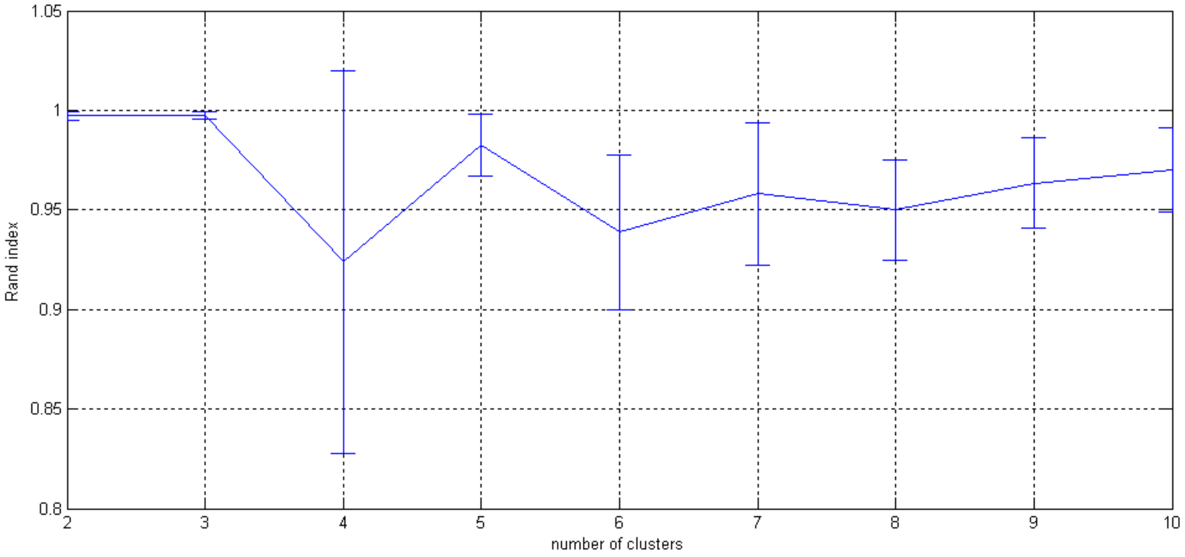


Figure 5. Rand index plot for the data set of April 1.

```
Algorithm 1: Partitioning
Input:
    $X$ dataset to be clustered;
    $k$ the number of clusters;
    $N$ the number of samples;
    $m$ the sample size;
    $\varepsilon$ the threshold value.
Resampling procedure: Randomly draw $N$ samples $S_i$ of size $m$ from $X$ without replacement.
 1: for all $S_i$ do
 2:     In the first iteration, set of centroids $C$ is chosen randomly.
 3:     Cluster $S_i$ by $K$-means algorithm, starting from the given centroid set $C$.
 4:     Cluster $X$ by assignment to the nearest centroid, using the centroids, obtained in the previous step.
 5:     Calculate the object function value of the partition $\pi(X)$ from the previous step according to (3).
 6: end for
 7: Choose from the set $\{S_1, \ldots, S_N\}$ a sample $S_0$ with the minimal object function value.
 8: if the first iteration is being processed or if the absolute difference between two minimal object function values, which are calculated for two sequential iterations, is greater than $\varepsilon$ then
 9:     replace $C$ with the set of centroids of $\pi(S_0)$, and return to step 1
10: else
11:     stop
12: end if
end
```

1) Estimation of the "true" number of clusters: In order to estimate the optimal number of clusters in the users' clusterization procedure described in Section III-B1, we repeat the clustering stability evaluation procedure described in Section III-B2 for each of the possible number of clusters in interval $[2, 10]$. The results for all dates are very similar. Fig. 5 demonstrates an example of Rand-index curve for April 1. It is easy to see that the maximal stability attitudes appear for $N = 2$ and $N = 3$.

Recall that the main purpose of the user clustering is to recognize behavior patterns, which represent the general structure of the users' population. Let us consider two possible estimators for the "true" number of clusters from this point of view. We describe a behavior pattern via an average level of the users' activity within each of the three hour clusters, defined in Section III-A. So, we take a three-dimensional representation of users, introduced in Section III-B1, and calculate the mean and standard deviation of each coordinate in each user cluster.

The user activity patterns for April 1 are shown in Fig. 6 by means of the error bar plot of values in each hour cluster. Recall that for the given date we obtained a 'night' cluster (red) with hours 1-8; a 'day' cluster (blue) with hours 11-21; and a 'morning/evening' cluster (green) containing hours 9-10 and 22-24 (see Fig. 4). For example, pattern A (the left panel in the picture) is characterized by the prevalence of the day activity since the average activity value is 0.84 for the 'day' hour cluster, in comparison with the values of 0.09 and 0.06 for the other hour clusters. Similarly, the behavior pattern B (the middle panel) describes users with significant activity in all hour clusters, while pattern C (the right panel) is characterized by high activity in the morning-evening hours.

The obtained result shows that we have a "clear" partition into 2 clusters and that one of them is well divided into 2 more sub-clusters. In fact, the two-clusters partitions contain the cluster corresponding to Pattern B and the united cluster for Patterns A and C. For our purposes, therefore,

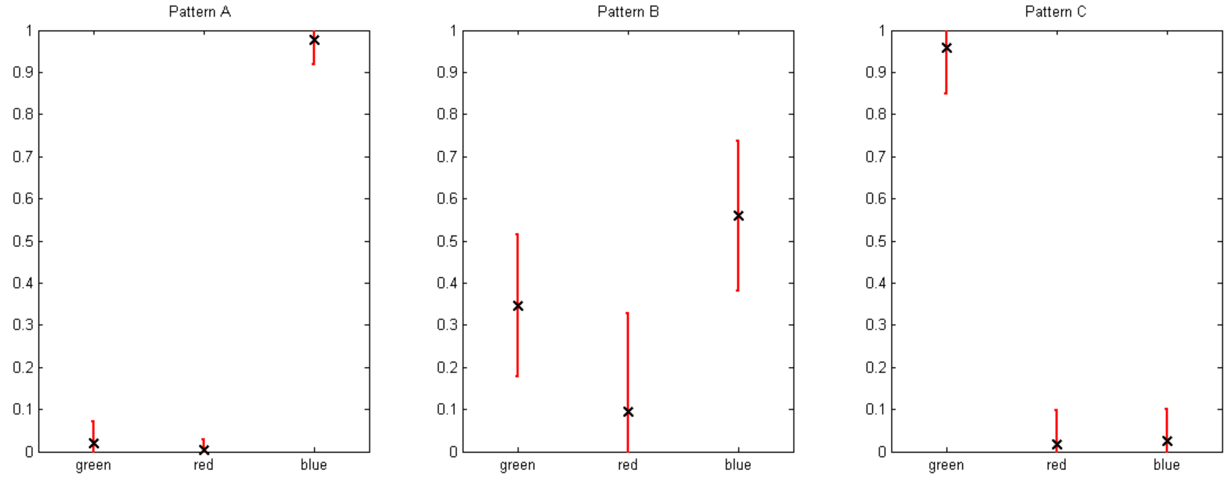


Figure 6. Profiles of 3 customer clusters (green, red, blue) for work day (April 1; $InCalls$).

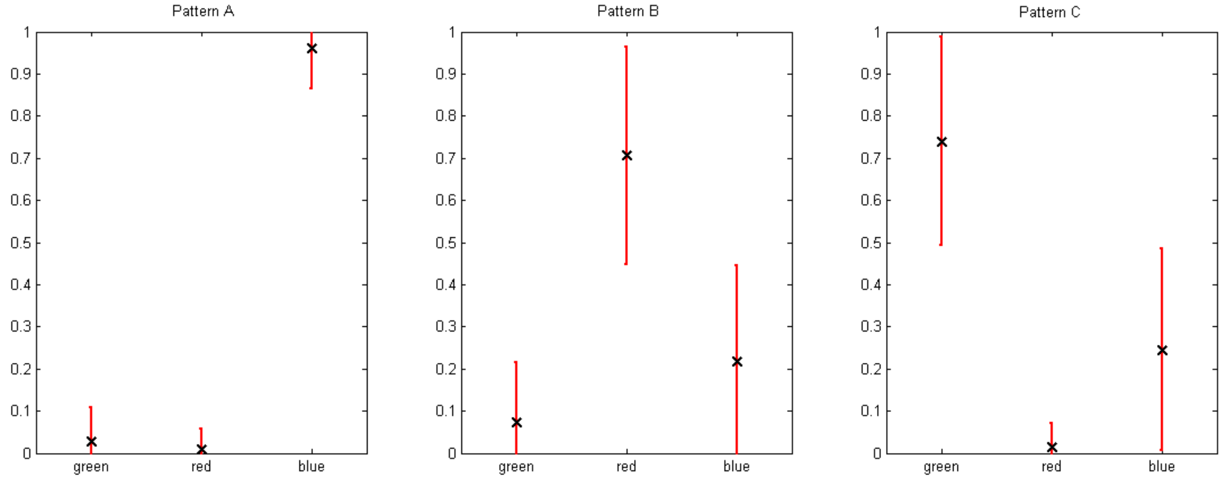


Figure 7. Profiles of 3 customer clusters (green, red, blue) for off day (April 5; $InCalls$).

it is natural to choose 3 as the ”true” number of clusters. Note, that it is common situation in cluster analysis: the ”ill-pose” number of clusters determination task can have several solutions depending on the model resolution. In the most cases, the population is partitioned into three clusters, with about 70, 20, and 10 user percentages in these clusters. This fact demonstrates the distribution within the population, connected to the call activity. This fact may be related to the nature of the people working time.

2) Procedure convergence: Now, we demonstrate that the resampling clustering procedure (Algorithm 1) converges very fast. In fact, Table VI shows the minimal objective function values for the first five iterations of the resampling procedure, executed on 100 samples for $k = 3$ (for others k, the situation is similar). The results show that, even in the second iteration, the minimal average of the distances does not change significantly as compared to the first iteration. In the subsequent iterations, this value remains constant to within 0.0001.

3) Profile stability: Further, we use behavior patterns for comparison of the results of our procedure on different datasets. The profiles for each considered date are shown in Tables VII and VIII. It is easy to see that they are stable both for work days and off days. However, the difference between work and off days is significant (see Fig. 6 and Fig. 7 for comparison). Although qualitative descriptions of profiles are very similar in both cases (pattern A with prevalent ”day” activity; pattern B with significant activity throughout 24 hours and pattern C with prevalent ”morning-evening” activity), in off days higher ”night” activity is detected.

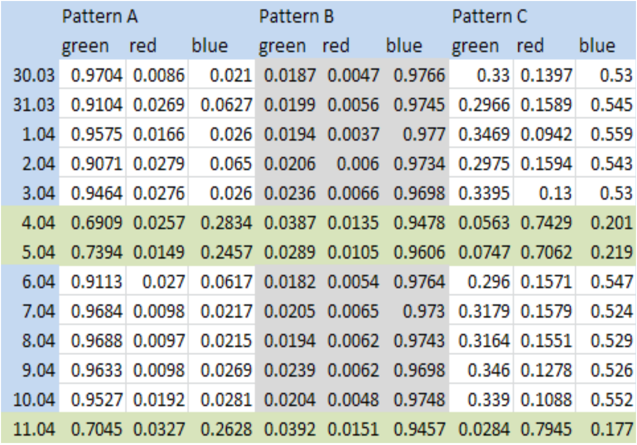
D. Call activity, associated within patterns

Now, we consider the call activity of the users, who correspond to each one of the three found clusters. The total activity of all the users within a day has a density with two

TABLE VI. Minimum of average distances to the nearest centroid for the first 5 iterations of the resampling procedure of Algorithm 1.

Iteration	1	2	3	4	5
Minimum	0.014487	0.013302	0.013295	0.013309	0.0132901

TABLE VII. Mean values for different Patterns in 3 clusters partition. (For Pattern X mean values for each hour cluster (red - 'night', green - 'morning/evening' and blue - 'day')).



	Pattern A			Pattern B			Pattern C		
	green	red	blue	green	red	blue	green	red	blue
30.03	0.9704	0.0086	0.021	0.0187	0.0047	0.9766	0.33	0.1397	0.53
31.03	0.9104	0.0269	0.0627	0.0199	0.0056	0.9745	0.2966	0.1589	0.545
1.04	0.9575	0.0166	0.026	0.0194	0.0037	0.977	0.3469	0.0942	0.559
2.04	0.9071	0.0279	0.065	0.0206	0.006	0.9734	0.2975	0.1594	0.543
3.04	0.9464	0.0276	0.026	0.0236	0.0066	0.9698	0.3395	0.13	0.53
4.04	0.6909	0.0257	0.2834	0.0387	0.0135	0.9478	0.0563	0.7429	0.201
5.04	0.7394	0.0149	0.2457	0.0289	0.0105	0.9606	0.0747	0.7062	0.219
6.04	0.9113	0.027	0.0617	0.0182	0.0054	0.9764	0.296	0.1571	0.547
7.04	0.9684	0.0098	0.0217	0.0205	0.0065	0.973	0.3179	0.1579	0.524
8.04	0.9688	0.0097	0.0215	0.0194	0.0062	0.9743	0.3164	0.1551	0.529
9.04	0.9633	0.0098	0.0269	0.0239	0.0062	0.9698	0.346	0.1278	0.526
10.04	0.9527	0.0192	0.0281	0.0204	0.0048	0.9748	0.339	0.1088	0.552
11.04	0.7045	0.0327	0.2628	0.0392	0.0151	0.9457	0.0284	0.7945	0.177

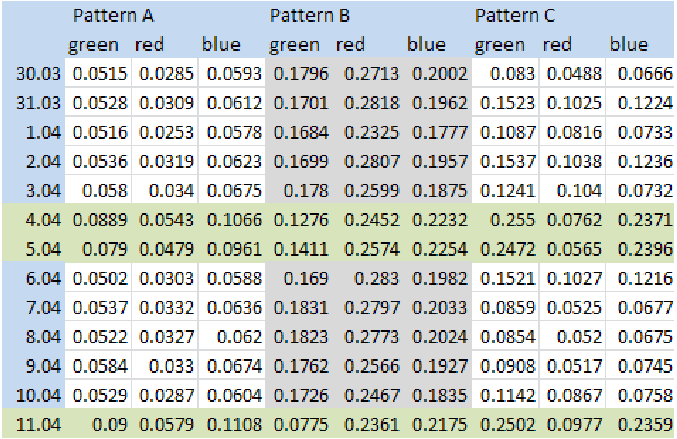
peaks. One of them is placed in the workday middle, and the second one, the higher peak, is located in the period after 7 p.m such that a local activity minimum is observed immediately after.

The shape of the corresponding density in the first cluster (A) is actually the same. However, the user's activity almost does not vary in the second cluster (B), i.e., the density curve has several insignificant peaks, and the activity decreases at 10 p.m. The total activity of the users belonging to cluster three (C) has two peaks, which are located in the morning and in the evening of a day.

The corresponding curves are shown in Fig. 8 and Fig. 9, where columns A, C present $InCalls$, and columns B, D present $OutCalls$. Here, the blue curves corresponds to the total activity densities of all the users; the red, green and brown ones give the total activity densities for clusters 1, 2 and 3, respectively. Note that both activity types have the same distribution shapes. Furthermore, the distribution of calls during a day for all three clusters is almost independent on the activity type (see Fig. 10a).

1) Features of the cluster model parameters: The model that we use reveals major differences between the DSN of the entire set of users and the DSNs for the individual clusters. In fact, for $InCalls$, the DSN for Cluster 1 is almost always best fitted by a single exponent. On the other hand, in more than half of the observed cases, the DSN for Cluster 2 is fitted by two exponents. Moreover, during the weekend period, the curve is fitted by three exponents. The DSN for Cluster 3 is usually fitted by two exponents, while the three-exponent fit sometimes arises without regard for the day of the week (Table IX). For $OutCalls$ (Table X), the above irregularities are more pronounced for Clusters 1 and 2, since all the best fits for Cluster 3 are two-exponential.

TABLE VIII. Standard deviation for different Patterns in 3 clusters partition. (For Pattern X std values for each hour cluster (red - 'night', green - 'morning/evening' and blue - 'day')).



	Pattern A			Pattern B			Pattern C		
	green	red	blue	green	red	blue	green	red	blue
30.03	0.0515	0.0285	0.0593	0.1796	0.2713	0.2002	0.083	0.0488	0.0666
31.03	0.0528	0.0309	0.0612	0.1701	0.2818	0.1962	0.1523	0.1025	0.1224
1.04	0.0516	0.0253	0.0578	0.1684	0.2325	0.1777	0.1087	0.0816	0.0733
2.04	0.0536	0.0319	0.0623	0.1699	0.2807	0.1957	0.1537	0.1038	0.1236
3.04	0.058	0.034	0.0675	0.178	0.2599	0.1875	0.1241	0.104	0.0732
4.04	0.0889	0.0543	0.1066	0.1276	0.2452	0.2232	0.255	0.0762	0.2371
5.04	0.079	0.0479	0.0961	0.1411	0.2574	0.2254	0.2472	0.0565	0.2396
6.04	0.0502	0.0303	0.0588	0.169	0.283	0.1982	0.1521	0.1027	0.1216
7.04	0.0537	0.0332	0.0636	0.1831	0.2797	0.2033	0.0859	0.0525	0.0677
8.04	0.0522	0.0327	0.062	0.1823	0.2773	0.2024	0.0854	0.052	0.0675
9.04	0.0584	0.033	0.0674	0.1762	0.2566	0.1927	0.0908	0.0517	0.0745
10.04	0.0529	0.0287	0.0604	0.1726	0.2467	0.1835	0.1142	0.0867	0.0758
11.04	0.09	0.0579	0.1108	0.0775	0.2361	0.2175	0.2502	0.0977	0.2359

Tables IX and X demonstrate comparison of the DSNs performance in Clusters 1-3 with the DSN, which is found within the total set of users. Each one of the curves exhibits its own cluster behavior characterizing the group. Nevertheless, joining any two of these clusters results in a three-component DSN. At the same time, we split the data randomly. This random partition into three clusters (with the same number of users as in the calculation of Clusters 1, 2, 3 as mentioned above) yields the same three exponent indexes, $t1 = 0.11$, $t2 = 0.31$ and $t3 = 1.01$ for all three clusters, which coincide with those calculated for the total set of users on the same day (see Table XI).

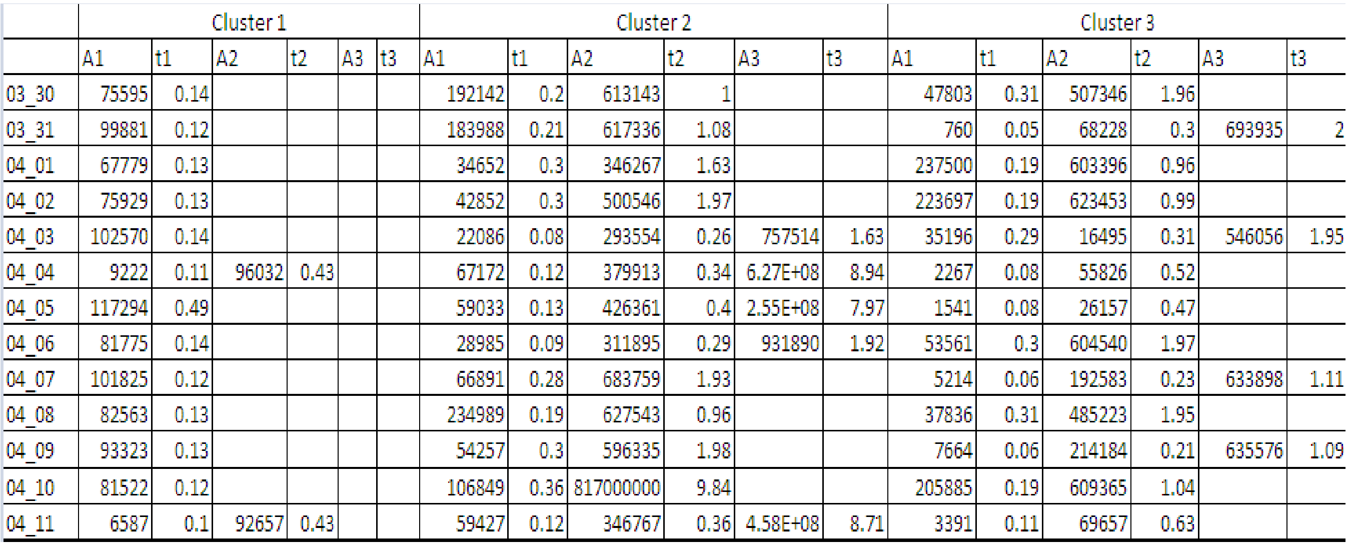
Thus, simplification of the cluster model shows that the partition into Clusters 1-3 actually reflects different activity characteristics for different groups of users. There are some differences on the weekends. However, in general, the parameters of a particular DSN are the same for each day. Note also that the DSNs of Clusters 2 and 3 are not in the least close to the second or third component (exponent) of the total set DSN. Indeed, in our model, the DSN of Cluster 2 consists mainly of two exponents, with one exponent disappearing at the decay value of 30, while the other (as a rule) is not decaying up to the value of 70. The DSN of Cluster 3 also has long-lasting components (up to 100 and more - see Table XII).

IV. Conclusion and Further Studies

Most of the recent studies, which consider the analysis of non-Gaussian processes, are related to hyperbolic distribution. The mere existence of such a distribution does not depend on the particular model, but rather is the result of the process being non-Gaussian in nature. Indeed, the Gnedenko-Doeblin limit theorem imposes restrictions on the form of a non-Gaussian distribution. Namely, its asymptotic behavior coincides with the Zipf distribution to within a slowly varying function.

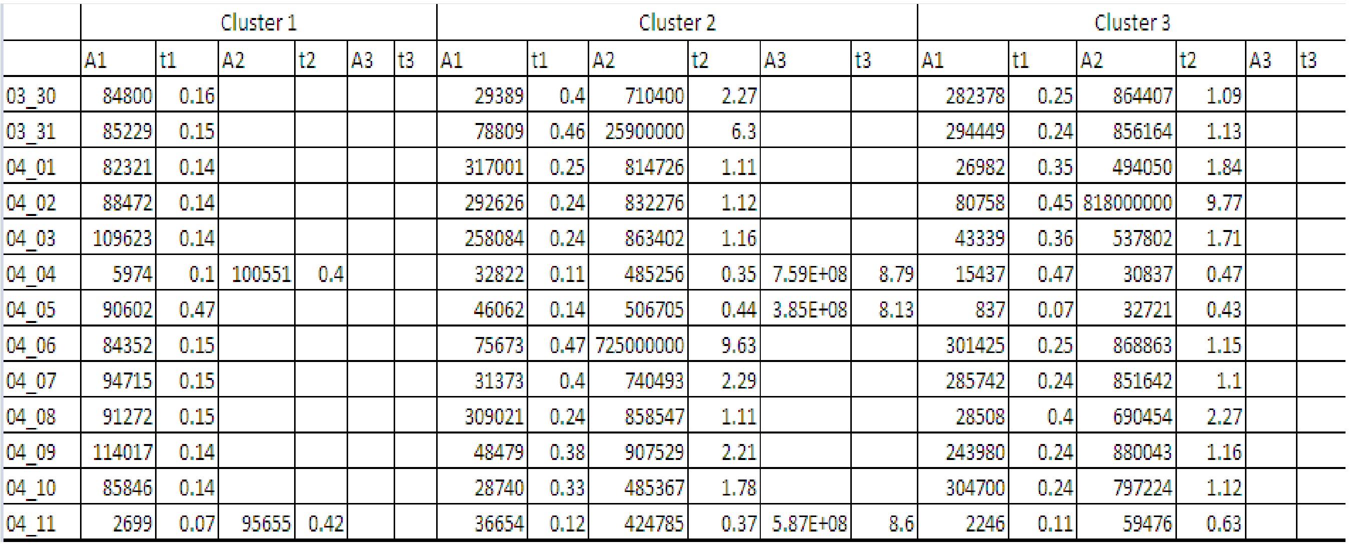
For example, the hyperbolic distribution was first observed in some fields of human endeavor, e.g., Pareto distribution of people according to their income and Zipf's law for the frequency of words in a text, [24]. It later turned out that the

TABLE IX. Parameters of the approximating curves for Clusters 1-3 during the days of observation ($InCalls$).



	Cluster 1						Cluster 2						Cluster 3					
	A1	t1	A2	t2	A3	t3	A1	t1	A2	t2	A3	t3	A1	t1	A2	t2	A3	t3
03_30	75595	0.14					192142	0.2	613143	1			47803	0.31	507346	1.96		
03_31	99881	0.12					183988	0.21	617336	1.08			760	0.05	68228	0.3	693935	2
04_01	67779	0.13					34652	0.3	346267	1.63			237500	0.19	603396	0.96		
04_02	75929	0.13					42852	0.3	500546	1.97			223697	0.19	623453	0.99		
04_03	102570	0.14					22086	0.08	293554	0.26	757514	1.63	35196	0.29	16495	0.31	546056	1.95
04_04	9222	0.11	96032	0.43			67172	0.12	379913	0.34	6.27E+08	8.94	2267	0.08	55826	0.52		
04_05	117294	0.49					59033	0.13	426361	0.4	2.55E+08	7.97	1541	0.08	26157	0.47		
04_06	81775	0.14					28985	0.09	311895	0.29	931890	1.92	53561	0.3	604540	1.97		
04_07	101825	0.12					66891	0.28	683759	1.93			5214	0.06	192583	0.23	633898	1.11
04_08	82563	0.13					234989	0.19	627543	0.96			37836	0.31	485223	1.95		
04_09	93323	0.13					54257	0.3	596335	1.98			7664	0.06	214184	0.21	635576	1.09
04_10	81522	0.12					106849	0.36	817000000	9.84			205885	0.19	609365	1.04		
04_11	6587	0.1	92657	0.43			59427	0.12	346767	0.36	4.58E+08	8.71	3391	0.11	69657	0.63		

TABLE X. Parameters of the approximating curves for Clusters 1-3 during the days of observation ($OutCalls$)



	Cluster 1						Cluster 2						Cluster 3					
	A1	t1	A2	t2	A3	t3	A1	t1	A2	t2	A3	t3	A1	t1	A2	t2	A3	t3
03_30	84800	0.16					29389	0.4	710400	2.27			282378	0.25	864407	1.09		
03_31	85229	0.15					78809	0.46	25900000	6.3			294449	0.24	856164	1.13		
04_01	82321	0.14					317001	0.25	814726	1.11			26982	0.35	494050	1.84		
04_02	88472	0.14					292626	0.24	832276	1.12			80758	0.45	818000000	9.77		
04_03	109623	0.14					258084	0.24	863402	1.16			43339	0.36	537802	1.71		
04_04	5974	0.1	100551	0.4			32822	0.11	485256	0.35	7.59E+08	8.79	15437	0.47	30837	0.47		
04_05	90602	0.47					46062	0.14	506705	0.44	3.85E+08	8.13	837	0.07	32721	0.43		
04_06	84352	0.15					75673	0.47	725000000	9.63			301425	0.25	868863	1.15		
04_07	94715	0.15					31373	0.4	740493	2.29			285742	0.24	851642	1.1		
04_08	91272	0.15					309021	0.24	858547	1.11			28508	0.4	690454	2.27		
04_09	114017	0.14					48479	0.38	907529	2.21			243980	0.24	880043	1.16		
04_10	85846	0.14					28740	0.33	485367	1.78			304700	0.24	797224	1.12		
04_11	2699	0.07	95655	0.42			36654	0.12	424785	0.37	5.87E+08	8.6	2246	0.11	59476	0.63		

same laws could be detected in other areas of human endeavor (e.g., the distribution of cities according to population) as well as in natural phenomena (e.g., time distribution of disasters). Internet activity and, in particular, user activity on social networks, appears to be an appropriate area for such analysis. Numerous studies suggest different models of social networks and try to link particular network characteristics with some measure of user activity. These characteristics often obey the hyperbolic law in one form or another.

From a practical point of view, the difference between non-Gaussian distributions (sometimes referred to as heavy-tailed) and the Gaussian distribution is quite important. The frequencies of extreme deviations in the two distributions are very different. The moments of non-Gaussian distributions increase with sample size, but do not tend to be limited as in the Gaussian case.

Although the social activity distribution of a population takes a specific and constant form, it can be assumed that the observed distribution is in some sense an averaged one. Obviously, it is composed of various types of distributions, generated by different social layers. We have in mind not only the groups, arising from the simplest types of differences, such as age and gender, but also the more complex features of the population under consideration. The purpose of this paper is to analyze the phenom as well as to decompose the observed distributions, according to certain external criteria.

It can be assumed that the hyperbolic law or the combination of distribution laws for various social groups, depend on the nature of the user joint activity. In some cases, each user's actions are in some sense sequential, so that their average behavior can be considered in the framework of a single law.

In the cases where users' actions occur in parallel, each user group, which is uniform with respect to some criterion, can generate its own law of activity distribution. Typically,

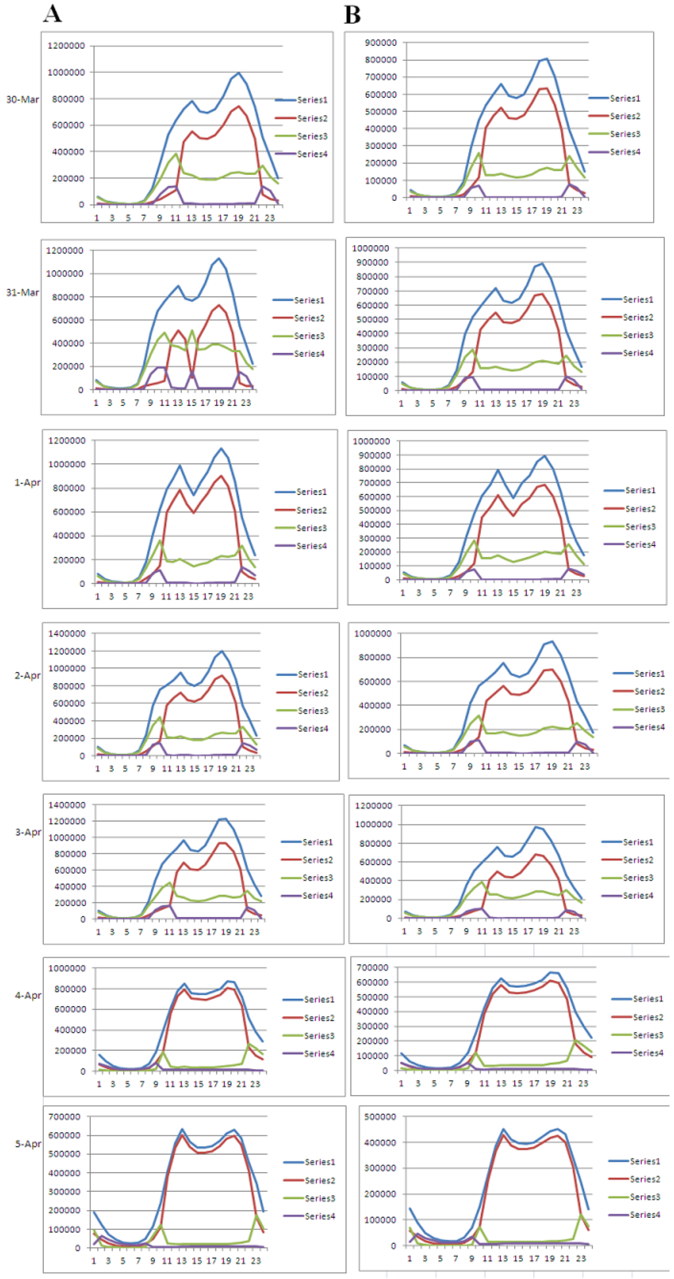


Figure 8. Distribution of $InCalls$ and $OutCalls$ during the day in the three clusters of users (A) and (B).

users actions in such a group are aggregated in order to pattern the group behavior. Researchers, who study the parameters of such networks, often find fractal properties and hyperbolic distributions. An example of parallel user activity is the number of records in an email address book. In a population of 16,881 users of a large university computer system, the cumulative distribution is not a powerful one, [10].

Since telephone calls are also more likely to be a parallel user's activity in the sense described above, we expected to find that the observed distribution of calls is the sum of several distribution functions, corresponding to different social groups

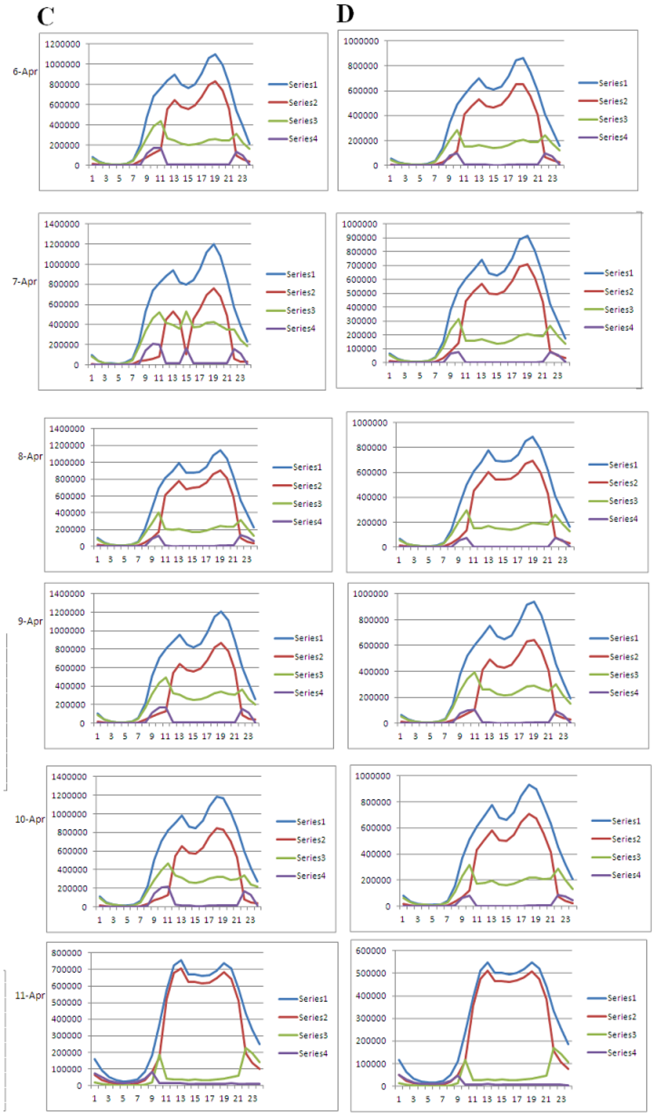


Figure 9. Distribution of $InCalls$ and $OutCalls$ during the day in the three clusters of users (C) and (D).

TABLE XI. Decay value of each DSN component in the activity/cluster model for all the three clusters considered ($InCalls$)

	cluster 1			cluster 2			cluster 3		
	1	2	3	1	2	3	1	2	3
03_30	80	0	0	60	13	0	34	6	0
03_31	95	0	0	57	12	0	132	37	6
04_01	85	0	0	34	7	0	65	13	0
04_02	86	0	0	35	6	0	64	13	0
04_03	82	0	0	125	48	8	36	31	6
04_04	82	26	0	92	37	2	96	21	0
04_05	23	0	0	84	32	2	91	21	0
04_06	80	0	0	114	43	7	36	6	0
04_07	96	0	0	39	6	0	142	52	12
04_08	87	0	0	65	13	0	34	6	0
04_09	88	0	0	36	6	0	149	58	12
04_10	94	0	0	32	2	0	64	12	0
04_11	87	26	0	91	35	2	73	17	0

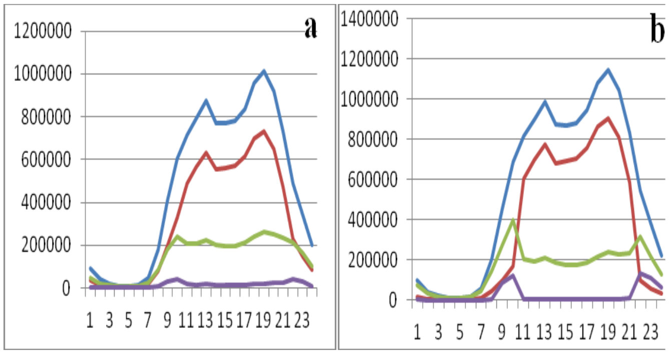


Figure 10. (a) Distribution of $InCalls$ for the clusters obtained for $OutCalls$. (b) Distribution of $InCalls$ for the clusters obtained for $InCalls$. Date: April 8. Notations of the curves are the same as in Fig. 8.

TABLE XII. Decay value of each DSN component in the activity/cluster model for all the three clusters considered ($OutCalls$)

	cluster 1			cluster 2			cluster 3		
	1	2	3	1	2	3	1	2	3
03_30	70	0	0	25	5	0	50	12	0
03_31	75	0	0	24	2	0	52	12	0
04_01	80	0	0	50	12	0	29	7	0
04_02	81	0	0	52	12	0	25	2	0
04_03	82	0	0	51	11	0	29	7	0
04_04	86	28	0	94	37	2	20	21	0
04_05	24	0	0	76	29	2	96	24	0
04_06	75	0	0	23	2	0	50	11	0
04_07	76	0	0	25	5	0	52	12	0
04_08	76	0	0	52	12	0	25	5	0
04_09	83	0	0	28	6	0	51	11	0
04_10	81	0	0	31	7	0	52	12	0
04_11	112	27	0	87	35	2	70	17	0

of users. The limited number of these groups is an important prerequisite for such differentiation because averaging over the groups is absent in this case. In [25], we introduced the notion of user strategy and showed that the number of different strategies is small. Therefore, we expected to obtain a small number of groups with equivalent user activity. Having no real-life socio-relevant parameters, we assumed that the peculiarities of a user's activity during a day may correlate with the user's social status.

We split the population into three clusters and showed that these clusters have simpler distribution functions than those for the total population. Yet, it is quite possible that a more detailed partition exists with even simpler distributions for each group.

References

[1] T. Couronne, V. Kirzner, K. Korenblat, E. Ravve, and Z. Volkovich, "Modelling behavior patterns in cellular networks," in Proceedings of ICCGI-2016, The Eleventh International Multi-Conference on Computing in the Global Information Technology, November 13-17, 2016, 2017, pp. 64–71.

[2] I. Simonson, Z. Carmon, R. Dhar, A. Drolet, and S. Nowlis, "Consumer research: in search of identity," Annual Review of Psychology, vol. 52, 2001, pp. 249–275.

[3] P. Kotler and K. Keller, Marketing Management, ser. MARKETING MANAGEMENT. Pearson Prentice Hall, 2006.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, no. 1, 1977, pp. 1–38.

[5] G. McLachlan and D. Peel, Finite mixture models, ser. Wiley series in probability and statistics. New York: J. Wiley & Sons, 2000.

[6] J. Willis and G. Yule, "Some statistics of evolution and geographical distribution in plants and animals, and their significance," Nature, vol. 109, 1922, pp. 177–179.

[7] J. Kenney and E. Keeping, "Linear regression and correlation," in Mathematics of Statistics: Part 1, 3rd ed. NJ: Princeton, Van Nostrand, 1962, ch. 15, pp. 252–285.

[8] N. Jewell, "Mixtures of exponential distributions," The Annals of Statistics, vol. 10, no. 2, 1982, pp. 479–848.

[9] J. Heckman, R. Robb, and J. Walker, "Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments," Journal of the American Statistical Association, vol. 85, no. 410, 1990, pp. 582–589.

[10] M. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," Physical Review E, vol. 66, Sep 2002, p. 035101.

[11] J. Hartigan, "Direct clustering of a data matrix," J. Amer. Statist. Assoc, vol. 67, 1972, pp. 123–132.

[12] Y. Cheng and G. Church, "Biclustering of expression data," in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, 2000, pp. 93–103.

[13] D. Witten and R. Tibshirani, "A framework for feature selection in clustering," Journal of the American Statistical Association, vol. 105, no. 490, 2010, pp. 713–726.

[14] R. Lopes, P. Hobson, and I. Reid, "The two-dimensional Kolmogorov-Smirnov test," in Proceeding of the XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Nikhef, Amsterdam, the Netherlands, April 23-27, 2007. Proceedings of Science, 2007.

[15] ——, "Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test," Journal of Physics: Conference Series, vol. 119, no. 4, 2008, p. 042019.

[16] L. Kaufman and P. Rousseeuw, Finding groups in data: an introduction to cluster analysis, ser. Wiley series in probability and mathematical statistics. New York: Wiley, 1990, a Wiley-Interscience publication.

[17] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, 1987, pp. 53 – 65.

[18] F. Galton, "Typical laws of heredity," Nature, vol. 12, 1877, pp. 492–495, 512–533.

[19] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in Human Behavior Understanding: First International Workshop, HBU 2010, Istanbul, Turkey, August 22, 2010. Proceedings, A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 14–25.

[20] D. MacKay, Information Theory, Inference & Learning Algorithms. New York, NY, USA: Cambridge University Press, 2002.

[21] J. Kogan, C. Nicholas, and M. Teboulle, Grouping Multidimensional Data: Recent Advances in Clustering, 1st ed. Springer Publishing Company, Incorporated, 2010.

[22] N. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in Proceedings of the 26th Annual International Conference on Machine Learning, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1073–1080.

[23] S. Wagner and D. Wagner, "Comparing Clusterings – An Overview," Universität Karlsruhe (TH), Tech. Rep. 2006-04, 2007.

[24] R. Baayen, Word Frequency Distributions, ser. Text, Speech and Language Technology. Springer Netherlands, 2001, no. 1.

[25] T. Couronné, V. Kirzhner, K. Korenblat, and Z. Volkovich, "Some features of the users activities in the mobile telephone network," Journal of Pattern Recognition Research, vol. 1, 2013, pp. 59–65.