Behavioural Network Traffic Analytics for Securing 5G Networks

Stavros Papadopoulos, Anastasios Drosou, Ilias Kalamaras, and Dimitrios Tzovaras
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
Emails: spap@iti.gr, drosou@iti.gr, kalamar@iti.gr, dimitrios.tzovaras@iti.gr

***Abstract*—The analysis of the network traffic in 5G networks is of high significance to the network security administrator, since it could allow for the identification of different behavioural groups and the distinction of anomalous from normal activity. The problem is the multi-dimensional nature of the data, e.g. SMS, call, Internet, services etc. that makes it difficult to analyse. This is even more challenging in 5G networks, compared to previous generation networks, since one more dimension is added to the traffic, representing different network slices. In this respect, activity that is normal in one slice can be anomalous in another. This paper presents a graph-based method for network mining and visualization of user activities in a mobile network. The raw multi-dimensional network traffic data are used for the construction of multiple multi-dimensional graph-based features that capture specific behavioural aspects for each user. Within each feature, graph matching techniques are applied in order to identify groups of users with similar behaviour. The dissimilarity results for each feature are combined using a multi-objective visualization method. The outcome is a data visualization in which users with similar behaviour are depicted as points close to each other. The network analyst is able to select the desired trade-off among the multiple features, and visually detect groups of users with similar behaviours, as well as possible anomalous clusters or outliers. Experimental evaluation of the proposed approach in several application scenarios verify its efficiency.**

I. Introduction

As the number of user devices in the mobile networks increases, the problem of network mining for security in cellular networks is becoming more and more challenging. The vast number of mobile devices that communicate every day results in a huge amount of information stored. Additional challenges are also introduced by the multi-dimensional nature of data generated in 5G networks, due to their network slicing mechanism, since each slice represents different services with different activity requirements. The detection of anomalous mobile devices has traditionally been handled by analysing the time-series of communications events of individual users (e.g. [1]). However, such techniques have lower accuracy when compared to more sophisticated approaches (e.g. graph-based approaches [2]), while they also need the manual definition of specific parameters to function well. Additionally, previous work has shown that the users form natural groups in the way they act in a social network [3], and thus clustering can be efficiently applied not only to identify such groups, but also to detect anomalies in an efficient manner. Such clustering techniques provide a higher level of abstraction that also addresses the issue of data minimization, by allowing the analyst to focus only on specific clusters and their attributes, e.g. size, relative relationship etc, instead of raw data. The use of data minimization makes the use of clustering techniques appropriate for application with Big Data.

The diversity of the malware types and behaviours renders the problem of anomaly detection as a very challenging one. Behavioural-based approaches (e.g. [4][5]) are promising in this respect, since they extract features that describe different aspects of the behaviour of malicious and normal actors, allowing for their efficient discrimination [6]. In this respect, the term “behaviour” represents the range of actions taken by actors in conjunction with themselves and their environment. In the context of mobile networks, the actors are the mobile phone users, and the actions are the communication types among them and between them and the network components. The behavioural-based analysis is even more challenging in 5G networks, since depending on the slice, the same activities can be either anomalous or normal, and thus, the analysis must be slice-aware.

In this paper, novel behavioural features are defined, which are extracted from Call Detail Record (CDR) data. These features are based on multidimensional graph representations of the communication activity of each mobile device in a 5G network. The different dimensions correspond to data coming from different network traffic types, services, and slices. Afterwards, graph matching techniques are applied in order to identify similarities between the users with respect to each graph-based feature. These similarities are given as input to the multi-objective visualization approach of [7]. This multi-objective approach is used for the identification and visualization of the various optimal trade-offs of the features, leading to the identification of various aspects of user behaviour and the subsequent identification of malicious behavior by the human operator.

II. Related Work

Information visualization systems provide insight into large amounts of complex data and are especially useful for network security applications [8]. Mobile network security related works have employed simple plots and graphs in order to

DOI: 10.1109/ICCW.2018.8403674

visually encode the network-related information, and help discriminate between different events. For example, *Eagle et al.* [9] proposed the use of entropy measures to characterize the CDR activity of each mobile device and utilized 2-dimensional plots in order to visually represent the changes of entropy over time. *Ye et al.* [10] utilized graphs in order to visualize CDR data, whereby vertices represent users and the edges call communication events. Similarly, *Shen et al.* [11] proposed Mobivis, a graph-based method for the analysis of CDR data. In Mobivis the vertices of the graph are more generic, since they can represent ontologies (e.g. users, locations, etc.), while the edges encode time dependent relationships between the ontologies. This more generic approach that allows for the vertices to represent arbitrary ontologies is also utilized in our work. The difference, however, is that the proposed ontology graph is created by the attributes of the CDRs, and also that it is not utilized for direct visualization, but for feature extraction as a preprocessing before the visualization. Most of the recent techniques for visualizing large amounts of data are specialized in a specific field, while supporting only a few data types and not being scalable to more than some thousands of elements. In a broader context, such methods fall short, and thus arises a need for new representations.

Since several attributes are available for each communication event, such as origin, destination, duration, time etc., they can be combined so as to lead to better insights about the data. In the literature, the combination of multiple sources of information for clustering, classification, visualization etc. has mainly been handled by multi-modal fusion methods [12]. The first most simple multi-modal fusion approach is to simultaneously combine the characteristics of all the available modalities, e.g. through weighted sums. For example, *Lin et al.* [13] also employed Multiple Kernel Learning [14] for graph-based fusion. A second multi-modal fusion approach is to utilize information of one modality in order to assist the learning of another modality in an iterative manner, such as [15]. A different approach is proposed in works such as [7][5][16], where the visualization of the modality fusion is formulated as an multi-objective optimization problem. This allows to simultaneously optimize all objectives by identifying a set of Pareto-optimal solutions, instead of only one solution. Despite the fact that such techniques are efficient for multi-modal fusion, their application in mobile network security is limited. In this respect, this paper follows this multi-objective approach, using graph-based features as modalities and a graph matching technique for measuring similarities.

While the use of graph representations for anomaly detection is very common in multiple fields, their use in mobile communication networks is limited (survey on [3]). In addition, most of the graph-based analytics techniques utilize feature extraction from the graphs, and not the graphs as features directly. The direct use of graphs has only been used in the area of object recognition, combined with graph matching techniques [17][18]. Under these considerations, this paper proposes and evaluates the use of graphs directly as features, and applies graph matching techniques for the analytical task of detecting groups of users with similar behaviours.

III. Method Description

Inspired by the multi-objective approaches that focus on the clustering of entities in an environment (e.g. [7] where entities are images, or [5] where entities are mobile devices), the proposed method uses an entity-based analysis scheme in order to analyze any type of record data. Each entity is defined as a collection of attributes or features. Examples of entities include: mobile devices, database records, user profiles, articles, time periods, network slices etc. After the specification of the entities, and motivated by the efficiency of graphs for feature extraction [3] and object recognition [18], multiple multi-dimensional graph based features are extracted for each entity, so as to capture their behavioural characteristics. The distance between the different entities for each feature is computed as the dissimilarity between the respective graphs. Finally, the multi-objective visualization methods presented in [7] is used for feature fusion and visualization.

Compared to previous graph-based methods, the proposed approach does not utilize a predefined set of features extracted from the graphs, but uses the structure of the graphs directly as behavioural descriptors. This fact renders the proposed approach more generic, since it is not focused on capturing only specific predefined characteristics of the communication activities.

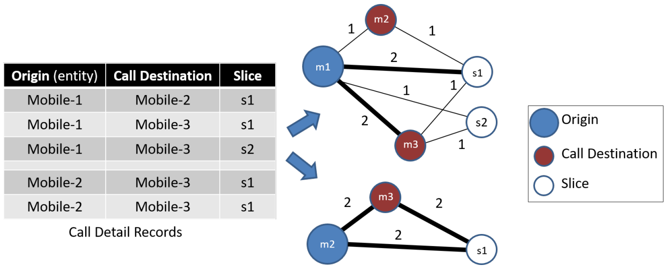
A. Definition of the multi-dimensional graph-based features

For simplicity and without loss of generality, the input dataset is considered to be comprised of a set of multi-dimensional attributes $A = \{a_1, ..., a_{|A|}\}$, where $a_l = \{v_1, .., v_{|a_l|}\}$ is a set of possible values, and a number of records $R = \{r_1, ..., r_{|R|}\}$, where each record is a set of attribute values $r_j = \{v_1, ..., v_{|r_j|}\}$ for $v_l \in a_l$. All the attributes are considered to be discrete, and the continuous attributes are transformed into discrete using binning.

The dataset entities are defined based on the values of a specific attribute, which is selected depending on the analysis task at hand. More specifically, the set of entities a_{ent} is defined as the set of different values of a specific attribute $a_l \in A$, where $a_{ent} \equiv a_l = \{v_1, .., v_{|a_l|}\}$. For example, in a dataset of CDRs from a mobile network, and the task of the identification of anomalous mobile devices, the entities are the mobile devices, as defined from the set of different mobile devices found in the "source of the call" attribute of the communication records.

The set of entities a_{ent} is used to separate the set of records R into $|a_{ent}|$ disjoint sets R_k, such that $R = \bigcup_{k \in [1, |a_{ent}|]} R_k$ and $R_i \cap R_j = \varnothing$ for $i \neq j$. Each subset of records R_k is constructed from the records that contain the specific entity $v_k \in a_{ent}$: $R_k = \{r_j | \forall v_k \in r_j, v_k \in a_{ent}\}$

As noted earlier, the behavioural characteristics of each entity are captured using graph-based features. Each graph-based feature of an entity $v_k \in a_{ent}$ is an undirected weighted graph $G_k^i(V_k^i, E_k^i, f_k^i)$, where V_k^i is the set of vertices, $E_k^i \subseteq V_k^i \times V_k^i$ is the set of edges, $f_k^i : E_k^i \to \mathbb{R}^+$ is a function that maps the



Origin (entity)	Call Destination	Slice
Mobile-1	Mobile-2	s1
Mobile-1	Mobile-3	s1
Mobile-1	Mobile-3	s2
Mobile-2	Mobile-3	s1
Mobile-2	Mobile-3	s1

Call Detail Records

Fig. 1. CDRs representing the origin and the destination of the communication calls and the network slice used for the call. The entity attribute a_{ent} is set to be the origin of the call, and the set of attributes of the definition of the graph-based feature F^i is comprised of the destination of the call and the slice. The weights of the edges of each vertex-pair are equal to the number of their co-occurrences in the CDRs.

edges to their respective (positive) weights, $i \in [1, n]$ is the index of the i_{th} feature out of a total of n features, and k is the index of the k_{th} entity $v_k \in a_{ent}$.

For the creation of the graph feature G_k^i, firstly a set of dataset attributes is selected $F^i \subseteq A$. Afterwards, the set of vertices is defined as:

$$V_k^i = \bigcup_{a_l \in F^i} a_l \tag{1}$$

while the set of edges is defined as:

$$\begin{aligned} E_k^i = \{(v_p, v_j) | \forall\ v_p \in a_l,\ v_j \in a_k,\ and \\ a_l, a_k \in F^i, and\ l \neq k,\ and\ v_p, v_j \in r_q, \\ where\ r_q \in R_k\} \end{aligned} \tag{2}$$

The weight of each edge is defined as the number of records that contain the corresponding two vertices used for the creation of the edge. More formally:

$$f_k^i(e_q^i) = |R_k^{q,i}| \tag{3}$$

where $R_k^{q,i} \subseteq R_k$ and:

$$R_k^{q,i} = \{r_t | \forall v_i, v_j \in e_q^i,\ and\ v_i, v_j \in r_t\} \tag{4}$$

Figure 1 shows an example of graph-based feature creation for each entity of the dataset. In this case, the dataset is a set of CDRs representing the origin, the destination of the communication calls, and the network slice used for the call. The entity attribute a_{ent} is set to be the origin of the call, and the set of attributes of the graph-based feature F^i are comprised of the destination of the call and the slice. The weight of the edges corresponds to the co-occurrences of the corresponding vertex-pair in the CDRs.

In the example shown in Figure 1, and for simplicity of presentation, there are only one feature-graph defined per entity, using a two attributes, i.e. the destination and slice attributed. As shown earlier in equations (1) and (2), the proposed approach provides the possibility to create multiple graph-based features, and each feature can be created using more than one attributes. The number of features utilized corresponds to the number of modalities that must be combined. In this respect, Figure 1 results in a uni-modal approach (i.e. a single feature). On the other hand, in the case that k features were utilized, the result would be a k-modal approach.

The proposed multidimensional graph encodes the communication activities and the slice in which each communication event has occurred. The graph is also generic and can incorporate activities from multiple slices. Additionally, the generic approach of the graph definition allows for the creation of slice specific features, i.e. creation of a set of graphs that encode call/SMS activities for each slice separately. This would allow for the explicit definition of slice activities encoded in the structure of the graph.

B. Multi-objective Behavioural Visualization

After the calculation of the graph-based features for each entity, this paper employs graph matching techniques [19] [20] to compute the pairwise distances between the entity graphs for each feature. More specifically, the distance between two entities v_k and v_l with respect to feature F_i is defined as follows:

$$D(G_i^k, G_i^l) = D_{eig} + D_{adj} \tag{5}$$

where D_{eig} is the eigenvalue graph matching method [21] which takes into account the structure of the graph, and D_{adj} is the absolute difference between the weighted adjacency matrices of G_i^k and G_i^l which takes into account the content of the graph. Given M_i^k as the weighted adjacency matrix of G_i^k: $D_{adj}(G_i^k, G_i^l) = |M_i^k - M_i^l|$.

Figure 2 illustrates examples of graphs created using their CDRs, similarly as in Figure 1. For simplicity, only the destination attribute of the CDRs is considered. Four cases are considered, two users with low activity ((a) and (b)), and two users with high activity ((c) and (d)). Blue colour represents the source and red the destination of the communication events. The width of the edges represents the number of communication events between the corresponding users. The low activity users (a) and (b) communicate with two and three destinations respectively with similar distributions, i.e. high communications towards one destinations. Thus, the graph matching procedure will output small distance between these two graphs. Furthermore, graphs (c) and (d) are similar with each other, since both users communicate with a large number of users (six and seven respectively) with similar distributions, i.e. large number of communications towards many destinations. On the other hand the graphs of the low and high activity users have many structural dissimilarities (different number of destinations and different edge weight distributions), and thus, the graph matching procedure will output larger distance between them than when using users of the same class, i.e. either low or high activity class.

After the calculation of the pairwise entity distances for each feature, the method proposed in [7][5] is used for behavioural clustering. Specifically, the computed distances

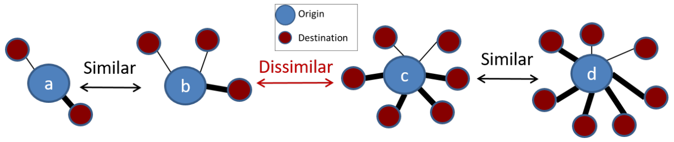


Fig. 2. Examples of graphs representing low activity, (a) and (b), and high activity, (c) and (d) , mobile users. Blue colour represents the source and red the destination of the communication events. The width of the edges represents the number of communication events between the corresponding users.

are used to construct minimum spanning trees H_i for each feature F_i, where the vertices are the entities and the edges have weights equal to the corresponding entity distances. The multiple graphs are used as the input to the multi-objective problem [7][5], and the solution is a set of Pareto-optimal solutions, namely the *Pareto front*, representing multiple trade-offs among the various behavioural characteristics. By selecting different solutions of the front, the operator can put more focus on one feature or another, or equally to all available features.

IV. Experimental Evaluation

This section presents the application of the proposed multi-objective visualization approach for network mining on multiple simulated datasets in cellular mobile networks. In this respect, the raw data consist of CDRs. The CDR data contain billing information about the calls and SMSs performed by the mobile users, including the time of the communication, its duration, the IDs of the communication origin and the recipient, the slice used for each event etc. In this context two slices are considered in the 5G network, one for call and one for SMS traffic.

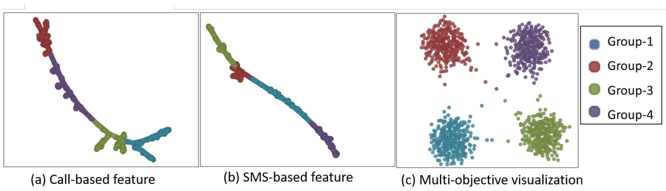
It should be noted that it is very difficult to apply analytical approaches for the evaluation of anomaly detection algorithms, due to the high dimensionality and complexity of the communication activities. This is the reason why most of the previous approaches utilize either real data or simulations for evaluation.

In order to quantitatively evaluate the visualizations in this section, the Dunn Index has been used, as also done in [7] and [5]. The Dunn Index is a clustering evaluation index that uses the available ground truth class labels of the data in order to quantitatively evaluate how well separated the data classes are in the visualization. It has been chosen since it considers the spatial separation of clusters, instead of just logical separation, thus making it appropriate for evaluating visual clusters.

It should be noted that in this section, weights are used to denote the various solutions of the Pareto set. In the case that the weight of a feature is larger than an another, then the corresponding solution favours this feature.

A. Application 1: Identification of user behavioral groups

This section presents the application of the proposed approach on the task of detecting different user behavioural groups from the CDRs in a mobile cellular network. The dataset was simulated using GEDIS studio [22]. Specifically, the dataset is comprised of 1,000 mobile devices, performing calls and SMSs for the duration of one day. Four different behavioural groups were simulated, as shown in Table I.



(a) Call-based feature (b) SMS-based feature (c) Multi-objective visualization

Fig. 3. Application of the proposed approach for the identification of different user behavioural groups in a cellular mobile network. Each point corresponds to one user. Different colors correspond to different behavioural groups. (a)-(b) MST representations of a feature created using the destination and time (with 24 quantization levels) attributes for the Call/SMS communications respectively. (c) Multi-objective visualization using the two features in (a) and (b) with equal weighting, i.e. 0.5 and 0.5. The Dunn Index value is 3.91.

TABLE I
Different groups of behaviours that were simulated for application 1.

Group ID	Short description
Group-1	250 users with normal SMS, and normal Call behaviour
Group-2	250 users with high SMS, and normal Call behaviour
Group-3	250 users with normal SMS, and high Call behaviour
Group-4	250 users with high SMS, and high Call behaviour

The normal SMS and normal Call behaviours have specific distributions and correspond to 50 communication events within the day. The high SMS and high Call behaviours have different distributions than the normal behaviours, and correspond to 100 communication events within the day. The CDRs are comprised of the following fields:

- Origin: Identifier of the origin of the communication event.
- Destination: Identifier of the destination of the communication event.
- Time: Timestamp of the communication event.
- Communication type: Call (in the first slice) or SMS (in the second slice).

For the analysis, the Origin attribute is used for the creation of the entities. Additionally, two graph-based features are created: 1) Using the destination and time (with 24 quantization levels) attributes for the SMS communications, and 2) Using the destination and time (with 24 quantization levels) attributes for the Call communications.

Figure 3 shows the application of the proposed approach for the identification of different user behavioural groups in the aforementioned dataset. Each point represents an origin of the communication events, while colour is utilized to illustrate one of the four different behavioural groups. Figures 3(a) and (b) show the single-feature representations (according to [5]) of a feature created using the destination and time (with 24 quantization levels) attributes for the Call/SMS communications respectively. The different classes are well separated in each

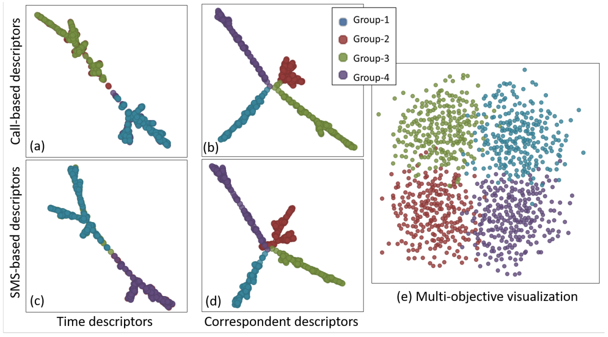


Fig. 4. Result of the approach proposed in [5] using four different features. Each point correspond to one user. Different colours correspond to different behavioural groups. (a),(c) Time Histogram Descriptors (THD) with respect to the Call and SMS activities respectively. (b), (d) Recipient Histogram Descriptors (RHD) for the Call and SMS activities respectively. (e) Multi-objective visualization using the four features in (a), (b), (c), and (d) with equal weighting. The Dunn Index is 1.82.

feature. Figure 3(c) shows the multi-objective visualization using the two features with equal importance, i.e. weights 0.5 and 0.5 respectively. The different clusters that are formed are well separated and easily identified, while they also correspond to the different behavioural groups. The Dunn Index of this figure is equal to 3.91.

Figure 4 illustrates the result of the approach proposed in [5] using four different features. Each point correspond to one user. Different colours correspond to different behavioural groups. Figures 4 (a) and (c) show the Time Histogram Descriptors (THD) with respect to the Call and SMS activities respectively. According to [5], these features are histograms of the frequency of the communication events within a day, with bin sizes equal to one hour. This is similar to the graph-based features proposed in this paper if we select as entities the source of the communication events, and as a feature only the time of the events, with quantization equal to one hour. Figures 4 (b) and (d) show the Recipient Histogram Descriptors (RHD) with respect to the Call and SMS activities respectively. These features are histograms in which each bin corresponds to a destination of the communication event for the specific origin user. The size of the bin is equal to the number of communication events towards the specific destination. This is similar to the approach proposed in this paper by creating a graph using the source of the communication events as entities, and as a feature only the destination of the communication events. The distance metric used by [5] for the histogram-based features is L1 norm. The RHD features, in (b) and (d), are able to efficiently identify the different behavioural groups, since they have different number of destinations. On the other hand, the THD features, in (a) and (c) are not able to completely separate the different behavioural groups since the groups 1 and 3 have normal SMS behaviour and groups 1 and 2 have normal Call behaviour.

Figure 4 (e) shows the multi-objective visualization using the four features in (a), (b), (c), and (d) with equal importances i.e. weights equal to 0.25, 0.25, 0.25, and 0.25 respectively. The different behavioural groups are not separated well. This happened due to the inclusion of the RHD features, in (b) and (d), which are not able to completely separate the different behavioural groups. The Dunn Index is equal to 1.82.

B. Application 2: SMS flood attack

In this section, the efficiency of the proposed approach is demonstrated on the task of identifying anomalous users which are involved in an SMS flood attack against the core network. The dataset was simulated using Gedis studio [22]. The dataset simulates a period of 7 days, the first 6 are the normal days, while the last day is the day of the attack. There are in total three different behaviours that where included in the dataset, as shown in Table II. Compared to our previous work [2], where a similar dataset is utilized, the proposed approach is able to identify the anomalies without feature engineering, which would make it application specific, and thus, is more generic. The proposed approach is able to not only identify the anomalous devices, but also different (not anomalous) behavioural groups.

TABLE II
DIFFERENT GROUPS OF BEHAVIORS THAT WERE SIMULATED FOR APPLICATION 2.

Group ID	Short description
Group-1	500 users with normal SMS, and normal Call behaviour
Group-2	500 users with high SMS, and normal Call behaviour
Group-3	100 users (anomalous users active in only the last day of the simulation) with anomalous SMS behaviour, and normal Call behaviour

The normal SMS and normal Call behaviours have specific distributions and correspond to 50 communication events within the day. The high SMS behaviour has different distributions than the normal behaviours, and correspond to 100 communication events within the day. The anomalous SMS behaviour have 10x times more SMS than normal, in order to perform the SMS flood attack according to the setting of [23]. The CDRs are comprised of the following fields (similarly to the previous section): Origin, Destination, Time, and Communication type.

For the analysis, the Origin attribute is used for the creation of the entities. Additionally, two graph based features are created: 1) Using the destination and time (with 24 quantization levels) attributes for the SMS communications, and 2) Using the destination and time (with 24 quantization levels)index attributes for the Call communications.

Figure 5 shows the results of the proposed approach and the approach proposed on [5] on the SMS flood attack dataset using all the features with equal weighting (the same features as in Figures 3 and 4). Each point in the visualization represents a different origin of the communication events, while colours are used to represents the three different behavioural groups. Specifically, red colour represents the low SMS users, green the high SMS users, and purple the anomalous users (which are active only in the last day). Figures 5 (a) and

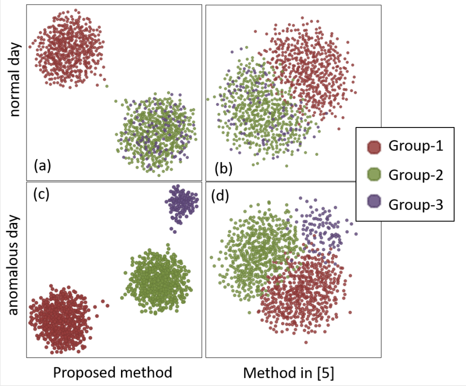


Fig. 5. The results of the proposed approach and the approach proposed on [5] on the SMS flood attack dataset using all the features with equal weighting (the same features as in Figures 3 and 4). Colors represent the ground-truth user groups: red for low SMS users, green for high SMS users and purple for anomalous users. (a),(b) Normal day using the proposed approach and the approach of [5]. The Dunn Indexes are 3.78 and 1.61 respectively. (c),(d) Anomalous day using the proposed approach and the approach of [5]. The Dunn Indexes are 3.2 and 1.69 respectively.

(b) illustrate a normal day using the proposed approach and the approach of [5]. The proposed approach is able to more efficiently discriminate between the two different normal SMS behaviours. The Dunn Indexes are 3.78 and 1.61 respectively. Figures 5 (c) and (d) show the anomalous day using the proposed approach and the approach of [5]. The proposed approach is able to efficiently separate the anomalous cluster from the two normal ones. The reason for this is that the graph-based features and the graph matching techniques are able to more efficiently characterize the user activities of the users than the simple histogram features. The Dunn Indexes are 3.2 and 1.69 respectively.

V. Conclusion

This paper presented a novel graph-based feature extraction process for the visualization of mobile network data and the identification of clusters with distinct behaviours in 5G networks. The proposed graph-based features are able to efficiently encode behaviours related to different communication patterns, such as the destination, the time of the communications events, or different network slice activities. Multiple graph-based features are extracted for the mobile users, encoding different aspects of their behaviour. These different graph-based features are exploited in the visualization, by adopting the multi-objective visualization approach of [7]. This approach is able to identify Pareto-optimal visualizations, which correspond to different trade-offs between the available features. Selecting a solution in the middle of the Pareto front results in visualization that combine the characteristics of all the available features, which can uncover useful data relationships. Experimental results on multiple scenarios provide evidence with respect to the efficiency in visualizing the behavioural similarities of users and in separating different behavioural patterns.

Acknowledgment

This work has been partially supported by the European Commission through project 5G-MoNArch funded by the European Union Horizon 2020 programme under Grant Agreement n^{o}761445. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

References

[1] I. Murynets and R. P. Jover, "Anomaly detection in cellular Machine-to-Machine communications," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 2138–2143, IEEE, 2013.

[2] S. Papadopoulos, A. Drosou, and D. Tzovaras, "A Novel Graph-based Descriptor for the Detection of Billing-related Anomalies in Cellular Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2655 – 2668, 2016.

[3] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, pp. 1–63, 2014.

[4] L. Cao, "In-depth behavior understanding and use: the behavior informatics approach," *Information Sciences*, vol. 180, no. 17, pp. 3067–3085, 2010.

[5] I. Kalamaras, A. Drosou, and D. Tzovaras, "A multi-objective clustering approach for the detection of abnormal behaviors in mobile networks," in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pp. 1491–1496, IEEE, 2015.

[6] M. Iliofotou, "Exploring graph-based network traffic monitoring," in *INFOCOM Workshops 2009, IEEE*, pp. 1–2, IEEE, 2009.

[7] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-Objective Optimization for Multimodal Visualization," *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1460–1472, 2014.

[8] H. Shiravi, A. Shiravi, and A. A. Ghorbani, "A survey of visualization systems for network security," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 8, pp. 1313–1329, 2012.

[9] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[10] Q. Ye, T. Zhu, D. Hu, B. Wu, N. Du, and B. Wang, "Cell phone mini challenge award: Social network accuracy–exploring temporal communication in mobile call graphs," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, IEEE, 2008.

[11] Z. Shen and K.-L. Ma, "Mobivis: A visualization system for exploring mobile data," in *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*, pp. 175–182, IEEE, 2008.

[12] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[13] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Multiple kernel learning for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 6, pp. 1147–1160, 2011.

[14] M. Gönen and E. Alpayd, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[15] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*, pp. 86–93, ACM, 2000.

[16] I. Kalamaras, S. Papadopoulos, A. Drosou, and D. Tzovaras, "MoVA: A Visual Analytics Tool Providing Insight in the Big Mobile Network Data," in *Artificial Intelligence Applications and Innovations*, pp. 383–396, Springer, 2015.

[17] T. Tung and F. Schmitt, "The Augmented Multiresolution Reeb Graph Approach for Content-based Retrieval of 3d Shapes.," *International Journal of Shape Modeling*, vol. 11, no. 1, pp. 91–120, 2005.

[18] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Combining topological and geometrical features for global and partial 3-D shape retrieval," *Multimedia, IEEE Transactions on*, vol. 10, no. 5, pp. 819–831, 2008.

[19] L. Livi and A. Rizzi, "The graph matching problem," *Pattern Analysis and Applications*, vol. 16, no. 3, pp. 253–283, 2013.

[20] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis and applications*, vol. 13, no. 1, pp. 113–129, 2010.

[21] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, “Algorithms for graph similarity and subgraph matching,” tech. rep., Technical Report of Carnegie Mellon University, 2011.

[22] GenieLog, “GEDIS Studio online,” 2014.

[23] E. K. Kim, P. McDaniel, and T. La Porta, “A detection mechanism for SMS flooding attacks in cellular networks,” in *Security and Privacy in Communication Networks*, pp. 76–93, Springer, 2013.