IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 February 2024, accepted 14 April 2024, date of publication 18 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391061

RESEARCH ARTICLE

City Hotspot Identification Using Smart Cyber-Physical Social System

FARHAN AMIN[1], (Member, IEEE), LIGANG HE[2], (Member, IEEE), AND GYU SANG CHOI[1], (Member, IEEE)

[1]School of Computer Science and Engineering, Yeungnam University, Gyeongsan 38541, South Korea
[2]Department of Computer Science, University of Warwick, CV4 7AL Coventry, U.K.

Corresponding author: Gyu Sang Choi (castchoi@ynu.ac.kr)

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF 2021R1A6A1A03039493).

ABSTRACT Recently, the concept of smart cities has become popular and got researchers' attention because it helps to improve citizens' lives by providing valuable services, for instance, smart transportation, smart homes, telecommunication, infrastructure, etc. Hotspot analysis is a classic problem concerned with spatial analysis. Telecommunication operators and companies always care to identify the Hotspots in the city. The hotspots are the places with very high communication strength relative to others. It is evident from the current literature that cyber physics social systems (CPSS) are useful in the identification of hotspots in a smart city. However, big data storage, analysis, processing, accuracy, and robustness are the key concerns. Thus herein, we propose a smart cyber-physical-social system for the analysis of hotspots using telecom data. Herein, our proposed CPS model is comprised of three layers and each layer has different functionality. In our proposed model, initially, raw Call Detail Data (CDR) data is collected at the data collection layer. Then smart CPSS passed it to the next layer. In the Data processing layer, CPSS performs pre-processing, data storage, and analysis. Then, it constructs a graph and performs a social network analysis (SNA). Herein, different from traditional centrality measures, we suggest Eigenvector and k-shell as social network similarity and Jaccard, cosine, as social behavioral measures. Herein, the process of city hotspot identification is performed, followed by SNA, which is conducted by quantifying the importance of each hotspot based on metrics. Finally, our proposed smart CPSS model accurately identifies Top-Ten hotspots. In this study, we use five-day data and compare the changes in the hotspot patterns. We validate our findings of hotspots with the original dataset and confirm the robustness and accuracy using autocorrelation and cross-correlation functions.

INDEX TERMS Cyber-physical social systems (CPSS), cyber-physical systems (CPS), smart city, big data.

I. INTRODUCTION

The concept of a smart city is popular nowadays and is used to improve people's lives. In general, information and communications technology (ICT) plays an important role in the development of a smart city [1]. A smart city aims to exchange information using smart devices and provide various services to the citizens [1]. The development of a smart city is a critical task because it requires intelligent choice and the planning infrastructure [1]. One of the challenging tasks of a smart city is to build an ICT structure [1], [2]. The second problem of smart cities is the lack of telecommunication infrastructure. In the development and provision of intelligent services in smart cities, graphs play an important role [3].

Graph theory is used in the modeling of highly connected systems for example; social networks, computer systems, biological and complex systems [3]. The graph theory models combine the modeling of system components and device-level logic. Big data is a recent emerging technology and is widely used in smart cities and telecommunication. In the

The associate editor coordinating the review of this manuscript and approving it for publication was Ting Yang.

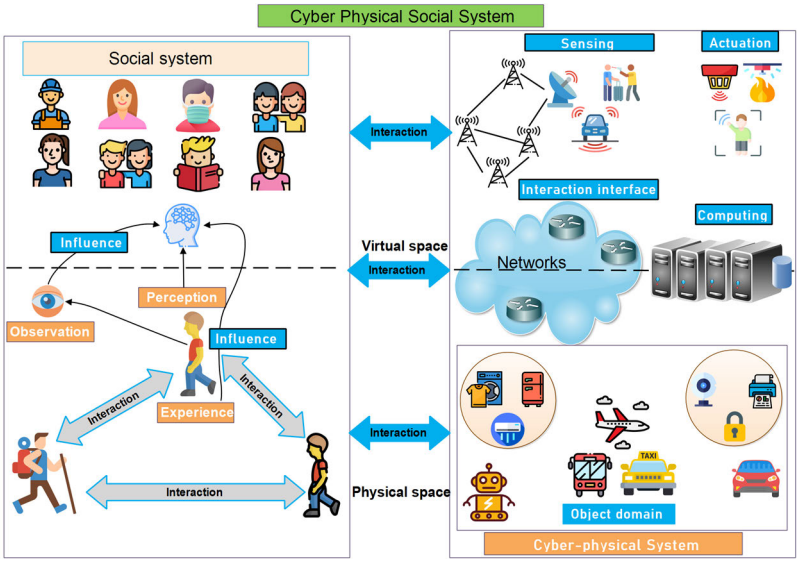


FIGURE 1. The structure of CPSS.

telecom industry, the Telecom Call Detail Records (CDR) [4] are considered a gold mine for data scientists due to their huge usage and high potential [5]. The challenges associated with data are it should be clean, free from errors, with no data duplication, and fewer missing values. In addition, it should be available in real-time [5]. The use of big data to mine customer behavior is called customer analytics [6] When a person calls another using a mobile phone, the CDR event gets generated [7]. A person doesn't have to use a mobile phone or GPS. In General, when a person calls another person, SMS, or even accesses the internet, the CDR is generated in the database. In many cases, telecom operators store the data in the system database. In most cases, the telecom operators have a separate department for this purpose [6]. CDRs are the main factor in customer analytics that need to be carefully investigated [8]. Cyber-physical systems (CPSs) are known as the next generation of intelligent systems and are composed of software and hardware that can control and monitor the physical environments using smart objects such as actuators and sensors [9]. These are used in the development of smart cities [10]. The smart objects in CPSS were connected to the real world using the Internet [11]. Recently, the concepts of CPSs have become a reality and become the core part of Industry 4.0 [12]. This phenomenon acts as a base for cyber-physical-social systems (CPSSs) [13].

CPSSs use big data and perform analysis to provide valuable services [9]. Nowadays, the global world is shifting towards the advancements and the integration of three aspects cyber, physical, and social [12]. A CPSS is the integration of a CPS and a cyber social system (CSS) [11]. CPSs are not limited to communication, multimedia, or entertainment [14].

Fig. 1 shows the basic structure of a CPSS [15]. In this figure, a reader can see that cyber, physical, and social spaces are connected. The first component named the social system mainly comprised of people or citizens. These people have relations and the relations are formed based on interaction, personal experiences, observations, and also perceptions. The second component is named as 'physical system'. This component comprised sensors and actuators. The sensing devices comprised sensors, actuators temperature sensors, etc. These objects are known as smart objects and were connected using communication technologies. The communication technologies are comprised (both wireless and wired) to process the data in the system and are shown in virtual space. CPSS is a recent and active research area and a lot of research has been carried out. The promising feature of CPSS is that it provides an interface between the objects so that they can send and receive the data and also carry out necessary actions. The promising integration feature of CPSS helps to improve the telecom and user services, especially in smart cities. Researchers have worked in this direction for example; in [16], the researchers proposed a use case study and performed network analysis using mobile network telecom data. Herein, the proposed data set is large and comprised of CDRs, including topological and country information. Similar work has been carried out by Visan et al. in [17]. In this research, the authors highlighted the communication service market problems faced by telecom operators [17]. Later, they presented various models and scenarios using telecom big data. Amin et al. [18] proposed CPSs for the analysis of hotspots in a smart city. The proposed system uses graph-based metrics for the identification of hotspots. However, the metrics are very basic and thus the accuracy is compromised and also the robustness is not discussed.

In General, the provision of valuable services to the users is the key part of a CPSS. The incorporation of modern communication and cutting-edge technologies focuses on providing high-quality services using low latencies.

Usually, the telecom big data is comprised of calls, SMS, and Internet data and big data are telecom transactions and pass through mobile devices. The challenges associated with big data are efficient data storage, analysis, and processing. These challenges became more difficult, especially with the incorporation of modern techniques named social network analysis (SNA) or machine learning [19]. These modern research methods require suitable big storage, analysis, and also modern distributed processing solutions.

From these aspects, it is necessary to propose and develop a big data model that can handle and effectively process, store, and analyze the big data. In addition, the proposed model should be smart so that it can provide fast calculation and processing time [20]. Therefore, based on these facts and ground truths, Herein we proposed a powerful big data platform with the ability to solve the above-stated challenges. Our proposed CPSS model is very smart because it can handle and process large-scale data using different layers.

Problem Statemen:

Hotspots or high-traffic communication areas [21] have a high activity and density compared to the other areas in a smart city [22]. Hotspot analysis is a classical problem concerned with spatial analysis [23]. Telecommunication operators and companies always care to identify the hotspots in a city to improve the quality of service.

Motivation:

The motivation of our research is to propose and develop a smart CPSS model that can efficiently process telecom data and perform data analytics. The proposed CPSS acts as a solution to the challenges associated with the extraction of large-scale data. The hotspots have a high density as compared to the other areas of a city. Thus, hotspot identification is useful to telecom operators and companies to focus only on specific areas in providing high-quality services.

The secondary motivation of this proposal is to provide a real-time big data model that will help telecom decision-makers. It is evident from the literature that, telecom operators and companies always take care of providing good services to the customers. As the influential hotspots in a network increase the service-providing features. Thus, it has importance in the telecom domain.

Key Contributions:

The key contributions are summarized below:

- Our proposed CPSS model is smart and comprised of three layers. Each layer has different functionality and hence, different functions have been performed by each layer. Our proposed model initially extracts the hotspots as high-traffic areas from a graph and later performs Social network analytics (SNA). Herein, we suggested social network similarity and social behavioral measures. These measures are used to quantify the importance of each node. Thus, our proposed CPSS model identifies Top-10 high influencers based on suggested metrics and it favors accurate analysis of telecom data.
- In previous studies, traditional centrality methods were used. Our proposed model is unique in all aspects because we have selected social and behavioral measures to detect the hotspots or high communication areas. Thus, it makes our proposed model more efficient. In addition, our proposed CPSS model is efficient because it provides accuracy and robustness which are not supported by the traditional methods.
- In this proposal, we confirmed that social network similarity and behavioral measures are useful in the identification of high communication areas. This will help the telecom operators to perform accurate analysis of large-scale telecom data.

Research Benefits:

- This research provides big data analysis using telecom data. Thus, it helps the telecom operators and the companies to identify the hotspots (high communication areas) in a smart city. It has a benefit for the telecom companies so that they can pay more attention to these areas in providing more good services in target areas.
- The proposed CPSS model is smart because it helps to identify the high communication areas in a smart city.
- In this proposal two research fields can be combined, i.e. Graph theory and communication.

The paper is organized as follows. In Section II, we presented related work. In Section III, we introduced our proposed smart CPSS model. In Section IV, we presented the details of a dataset. Section V, presents data analysis and a detailed discussion of experimental results. Finally, section VI, concludes the conclusion of our study.

II. RELATED WORK

In this section, we reviewed the state of the art in this area [20]. For instance; Onnela et al. presented an analysis of telecom data using CDRs [24]. For this, they have selected a large dataset consisting of millions of CDR data records. They have used customer call records and considered them as weighted graphs [25]. They performed analysis and suggested weighted distribution, weighted clustering, and degrees to identify the correlation between quantities. According to these authors, it will help the readers to understand the network structure. Similar work has been carried out by the Nanavati et al. They suggested degree distribution and neighborhood distribution [26] in their proposed model and analyzed a large data provided by Indian Telecom.

Nattapon et al. presented research on CDRs using a telecom dataset [27]. In this study, they propose a method to clean large data using ‘‘filters to filter’ to remove anomalies. Ahmad et al. [28] presented an advanced framework named the churn prediction SNA model. In this model, they combine big data and machine learning [28]. Herein, they suggested various network centrality measures to provide an equality analysis between each node pair. They perform an analysis and hence each node pair interacts with the others using links [20].

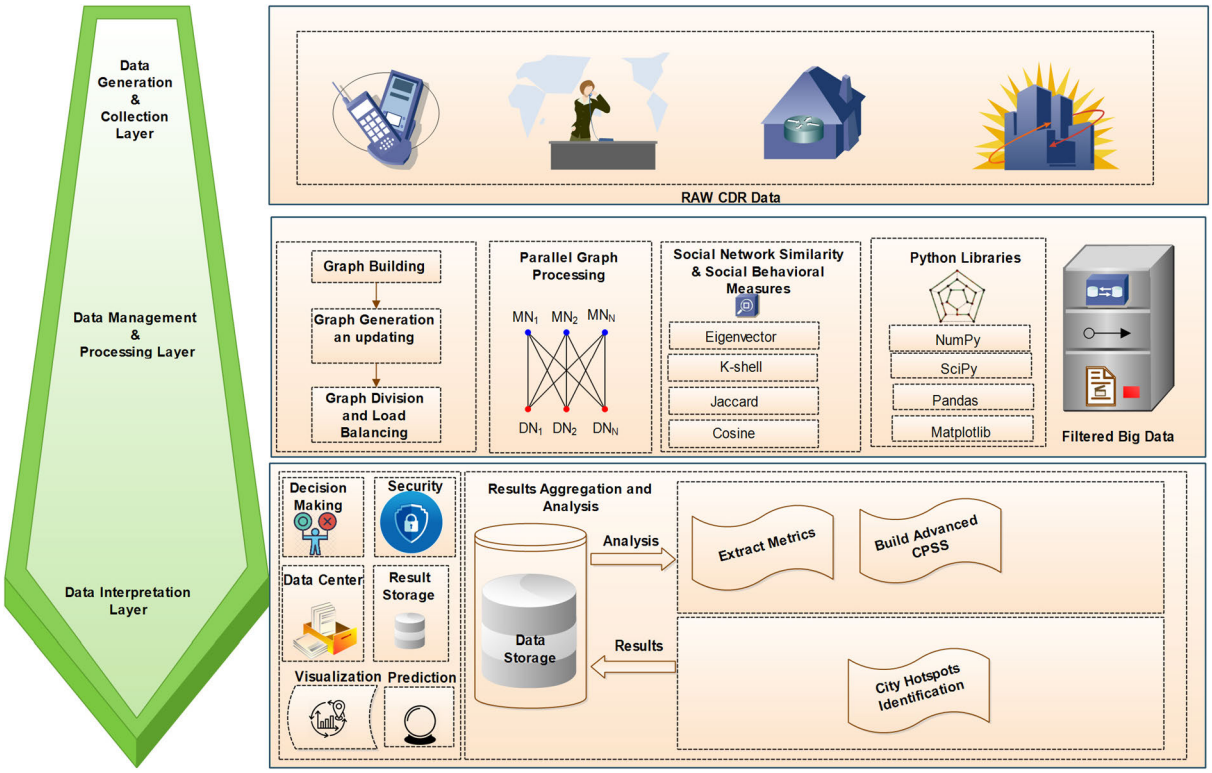


FIGURE 2. Proposed smart CPSS model.

It is evident from the literature that SNA and the centrality measures were used in churn prediction. Modarresi et al. [29] proposed a graph-based analysis model intending to increase the resilience of smart homes [29]. Herein, they suggest several topologies using smart home scenarios. Mededovic et al. [30] explored various centrality metrics and then concluded that they were used in the analysis of hotspots in a certain area [30]. Herein, they performed a detailed analysis using two weeks of telecom data to find the hotspots in the network and also measure the interaction. In this research, they used Eigenvector as a key measure to rank the hotspots. Seufert et al. [31] proposed a Wi-Fi hotspot model for the building of a smart city [31]. Herein, the Top-ten Wi-Fi hotspot locations were identified using a public Wi-Fi dataset. They concluded that the different Wi-Fi locations can be modeled using a uniform distribution. The angles and the gamma distribution can be maintained using minimum distance. This is a very simple Wi-Fi hotspot model and the locations are used to create the spatial distributions.

Peiyan et al. [32] presented an advanced data-forwarding method for opportunistic networks [32]. In this research, they explored various sizes of hotspots in the network. Herein, they propose a Hoten as a metric used for routing. This metric is used in human mobility. Another measure named entropy is employed. The function of entropy is to identify the public and personal Hotspots. Brdar et al. [33] presented a knowledge retrieval model using telecom data [33]. Herein, They suggested various centrality measures named closeness and degree centrality [34]. Finally, Amin et al. [18] proposed CPSs for the analysis of hotspots in a smart city. The proposed system uses graph-based metrics for the identification of hotspots. However, the accuracy and the robustness are not presented.

Briefly in the above review, we presented various researcher's work and they used traditional centrality measures for example Degree, closeness etc. In summary, these centrality measures are used for detecting the influencers in small or medium-scale networks. It is noticed that they are not suitable for large-scale networks. Similarly, a few measures for example; PageRank, etc. is incompatible with the telecom data. Because it is used to rank web pages over the Internet [20]. Therefore, to overcome these issues. Herein, we propose a smart CPSS model to measure the large traffic areas in smart cities. Our proposed model is unique in all aspects because we have selected social and network measures to detect the hotspots. It is noted that in previous studies, these measures were not used. Thus, it makes our proposed model more efficient. In addition, our proposed CPSS model is smarter because it provides accuracy and robustness which are not supported by the traditional methods. The details of our proposed conceptual model are discussed in section III.

III. INTRODUCTION TO THE PROPOSED SMART CPSS MODEL

Herein, we explain our proposed smart conceptual CPSS model. Our proposed model is a four-layer model, and each layer has different functionality. Fig. 2 shows our proposed smart conceptual framework. In this diagram, the functionality and the working mechanism of each layer are given.

TABLE 1. Dataset description.

Number	Dataset type	Issuer	Area	Rows	Column	size
1	Grid	Telecom Italia	Milan, Trentino	1048576	8	79.0 MB

TABLE 2. Genrated data.

Datetime	CellID	Countrycode	Smsin	Smsout	Callin	Callout	internet	Sms	calls
2013-11-01	1	0	0.3521	0.0000	0.0000	0.0273	0.0000	0.3521	0.0273
2013-11-01	1	33	0.0000	0.0000	0.0000	0.0000	0.0261	0.0000	0.000
2013-11-01	1	39	1.7322	1.1047	0.591	0.4020	57.7	2.8369	0.9939

A. DATA GENERATION AND COLLECTION LAYER

This layer is responsible for the data generation and the basic function is to generate and efficiently process the data. Initially, raw low-level CDR data is collected as shown in Fig. 2. The data is usually comprised of both CDR and customer data. At first CDR data includes call type (incoming/outgoing), calling number, called number, switch ID, and call duration. On the other hand, the customer data comprised the customer's name, age, address, sex, and customer ID. The user CDR consists of SMS, call and voice calls, etc. Herein first the data is collected and then transmitted to the next layer for further processing.

B. DATA MANAGEMENT AND PROCESSING LAYER

This layer receives the data from the upper layer. The basic function here is to normalize the data into a meaningful form. In this way, meaningful information is extracted from that data. It is observed that when we process large amounts of data it requires a variety of resources and power. It is evident from the literature that traditional data processing methods are not feasible to handle large-scale data. Thus, we should propose and develop a smart model that can preprocess large-scale data. Herein, our smart CPSS model can tackle this issue by following these steps.

Initially, the incoming large data stream is stored in a database named; filtered big data, and then handled by using the Python data analysis library (pandas). The data handling procedure is shown in Figure 1. The basic function of this layer is to store the data. Herein, the data storage system creates filtered big data. Herein, the initial processing steps have been performed, for instance; handling the redundant data and dealing with data with errors. Subsequently, the received data is converted into a graph database. Herein, the Pandas perform cleansing, transforming, and the manipulation of the data [35].

Network X [35], uses that database and generates a graph. The graph database can then extract the hotspots. Network X has a data structure for graphs for example; directed, in-directed, etc. Herein, our proposed social network similarity and social behavioral measures are applied. The data storage unit initially creates filtered and redundant data and the errors were removed. The smart CPSS processes an equal amount of data and generates the graphs as output at the same time. Finally, the error-free data is converted into graphs. The processed data is forwarded to the graph-building component.

1) GRAPH BUILDING COMPONENT

This component is a key part of our proposed model. Herein the process of graph construction has been performed. Milan city is represented as a Graph G. In G the set of nodes is considered a hotspot and the set of edges is considered a hotspot. A weighted network is represented as $G{=}(V, E)$. We have assigned two different types of weights that connect each edge. These weights depend upon the call duration between both sides of the edges. The graph-building component increases the efficiency of our proposed model by dividing the graph into numerous mutually exclusive subgraphs. After completing this task the independent subgraphs are sent to the processing server for further processing. Hence, the overall system load balance is maintained. Herein, the parallel graph processing component processes the graph, and also multiple servers are available to process each sub-graph as shown in Fig. 2. Herein, each parallel graph processing server is equipped with a specific graph algorithm. The specific algorithm runs based on the user's request. As a result, every server produces corresponding output to each graph algorithm for each subgraph of the main graph. When graph processing is required, the subgraph forwards all independent subgraphs to the processing server, here the load balancing task has been performed. After that data management and processing layer output in the form of segments. Each segment corresponds to the result on each graph. These chunks were aggregated and here the finalized dataset was sent for further processing to a designated server. Herein, the graph data is first stored on the local disk and then forwarded to the next layer.

C. DATA INTERPRETATION LAYE

The data interpretation layer comprised three units named cloud server, a storage device, and a data center as shown in Figure 2. After completing of data management and processing tasks, the attained results were ready for compilation. Herein, the parallel graph processing server forwards all results including, partial and complete results (performed in the data processing layer) to the data storage center (Data interpretation layer). Herein, initially, these graphs were stored in the result storage and the data center (As storage locations). The functionality of both storages is similar and therefore, it depends upon the user's choice. It should be noted that the cost of both storage devices will be different. Herein, the security component provides basic security to these storage locations. The data interpretation layer handles

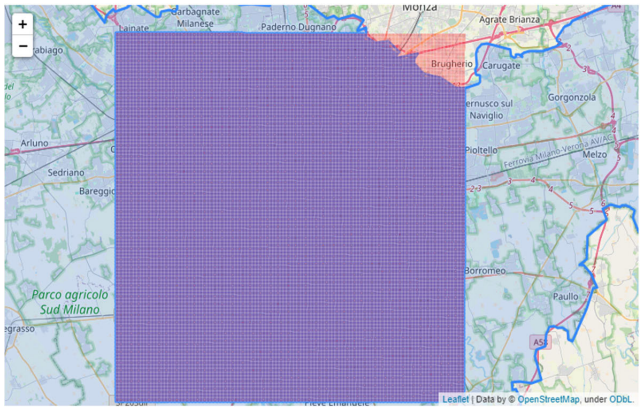


FIGURE 3. Area identification: milan cells.

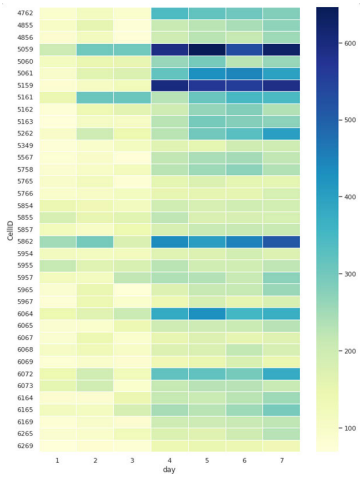


FIGURE 4. HeatMap telecom activity.

both database management and storage tasks efficiently. Finally, our proposed smart model stores, analyzes, and visualizes the achieved results to end users. In the next section, we discuss the details of the dataset and the achieved results.

IV. DATA PREPARATION

we have used the "Telecom Italia Big Data Challenge" dataset [22]. This data was provided by the telecom provider and collected in November 2013. Data between Trento and Milan cities were collected and a description of the dataset is given in Table 1. In this table, the dataset type, issuer, and size of the dataset are given. The key elements are given below:

- The ID field shows the identification numbers for Milan and Trento.
- The volume field shows the incoming/outgoing connections for SMS.
- The time field shows the time for an event.
- The incoming /outgoing connections field for any calls is shown as volume.
- The Internet traffic and country code fields are given in the dataset.

We have used "mobile phone activity" which contains, records, Internet data, and SMS as key elements. Herein, the description details are given below.

- *id*1 field shows the number of squares in Milan/Trento.
- *id*2 field shows the square of the Milan/ Trento grid.
- The Times field shows the time occurrence of events.

A. THE IDENTIFICATION OF HIGH COMMUNICATION AREAS

In this section, we explain the process of identifying high communication areas and later, we suggest metrics for these communication areas.

As mentioned earlier, the first step is to identify the high communication areas. Herein, we have defined a parameter named threshold. This is a key parameter and is used to identify high communication areas. This parameter extracts the minimum amount of communication traffic in a certain area. This is a dynamic parameter and can be changed according to the environment and circumstances.

Let, i is denoted as a high communication area.

$$I_i > \frac{1}{N} \cdot \sum_{j=1}^{N} I_i + \omega \tag{1}$$

I_i is the communication strength for the area i and ω is computed using the below equation.

$$\omega = \left(Trif - \frac{1}{N} \cdot \sum_{j=1}^{N} I_j \right) P \tag{2}$$

In Equation 2 $Trif$ is denoted as maximum amount of communication in all areas, and P is a cutoff or threshold parameter. By applying equation (1) *and* (2). The extracted specific areas of Milano province's grid are shown in Fig.3.

In this figure, a reader can see that Milano province's grid divides the whole province into various cells (approximately ten thousand). Herein, we ignore, those few cells that are outside of the province boundary, and the cells within the boundary are only considered. Fig.4 shows a heat map of telecom activity. The x-axis shows the day and the y-axis shows the CellID. We can see that the cell is increased from day 4 to 7. It is noted that it is relatively low from day 1 to 3. The darker ones (represented in different colors) shown in this figure are more active than the other cells (shown in yellow or orange color).

Fig. 5 shows the highly active cells on the map. Fig. 6 is the extended form of Fig.5 shows the overlay of those locations (Popular point locations) on the map. In this figure, the highlighted cells have maximum telecom activity and the points shown on the map are popular locations in the province. The identified locations shown on the map have high traffic as compared to the other areas in a city.

Fig.7 shows the visualization of mobile phone and internet activity in these areas. Herein, we can see that high communication activity has been observed in Dumo, Bocconi, and Navigli. A reader can see in the figure that the number of connections in the Duomo is higher than in both Bocconi and Navigli. This is because a lot of people were living in both

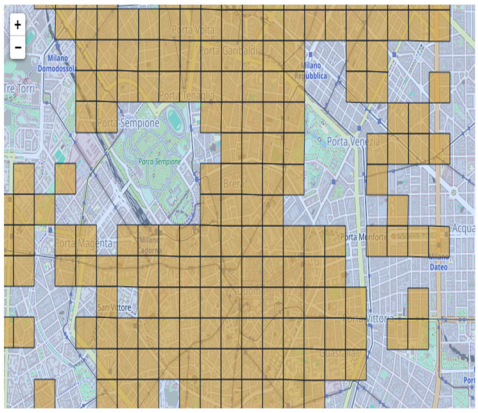


FIGURE 5. Highly active cells Identification.

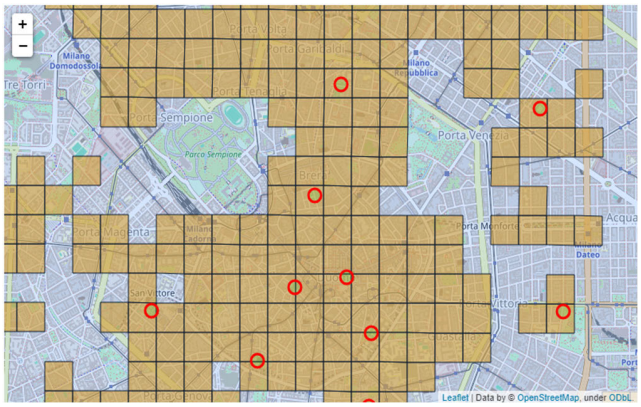


FIGURE 6. Popular point locations overlaid on active cells.

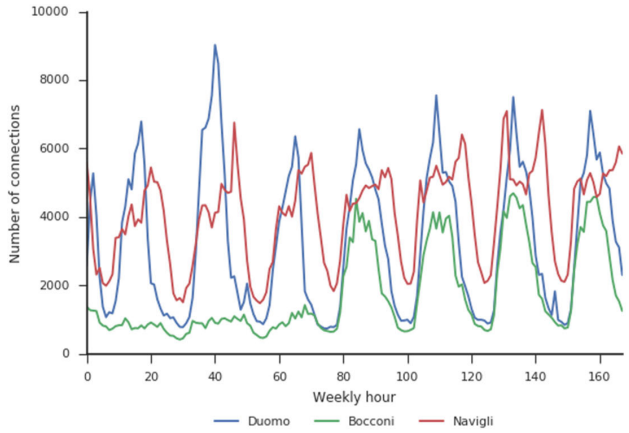


FIGURE 7. Mobile phone and internet activity in different areas.

areas. The calls were reduced during the weekend. Table 2 shows the generated output of data along with key elements.

B. USEFUL SOCIAL NETWORK SIMILARITY AND SOCIAL BEHAVIORAL MEASURE

After finding high-strength communication areas, the next step is to use social network similarity and social network behavioral measures to quantify the importance of each hotspot. The suggested measures are used to get useful insights into the networks and are also helpful in understanding the structural and behavioral properties of nodes. In addition, it shows the differentiating of nodes based on their importance in the network.

1) IDENTIFYING TOP-TEN HOTSPOTS USING SUGGESTED MEASURES

In the previous research discussed in [5], and [18] traditional centrality measures were used to identify the influentials in the network. However, these traditional measures were used for detecting the influencers in small or medium-scale networks. It is noticed that these measures are not suitable for large-scale networks. In addition to computation, they need complex calculations. Simialry, few measures for instance; PageRank is incompatible with the telecom data. Therefore, these are not enough, and herein, we employed more advanced measures to quantify the influentials [36], [37]. The complete description of social network similarity and social behavioral measures measures are given below.

2) EIGENVECTOR CENTRALITY

Eigenvectors are used to identify the most important nodes while considering the importance of neighbors in a network. It is an extended form of a degree of centrality [38]. The eigenvector in a node's influence is calculated based on the number of connections to others in the network [39], [40]. The eigenvector for a node v is defined as:

$$EV\ (v) = \frac{1}{\lambda} \sum_{j=1} Aij\ vj \tag{3}$$

where λ is a constant scaler value. The adjacency matrix is square and is represented by Aij [40].

$$A_{ij} = \begin{cases} 1 & If\ there\ is\ an\ edge\ between\ v_i\ and\ v_j \\ 0 & Otherwise \end{cases} \tag{4}$$

3) K-SHELL MEASUR

Wang et al. [41] proposed this measure based on the k-shell value. The k-shell is an iteration factor and is computed as;

$$\sigma_v = k_s.\left(1 + \frac{n}{m}\right) \tag{5}$$

Herein, k_s is the k-shell value for a node v and, m is the iteration number. The v is the removed node in the $n - th$ iteration of k degree. The proposed influence capability is computed based on:

$$IC_v = \sigma_v.D_v + \sum\nolimits_{v \in N(v)} \sigma_u.D_u \tag{6}$$

where IC_v presents the influence capability of a node v and σ_v is the k-shell iteration factor. This is a global measure. Here, the IC takes into consideration both global and local measures for the finding of the most influential nodes in the network.

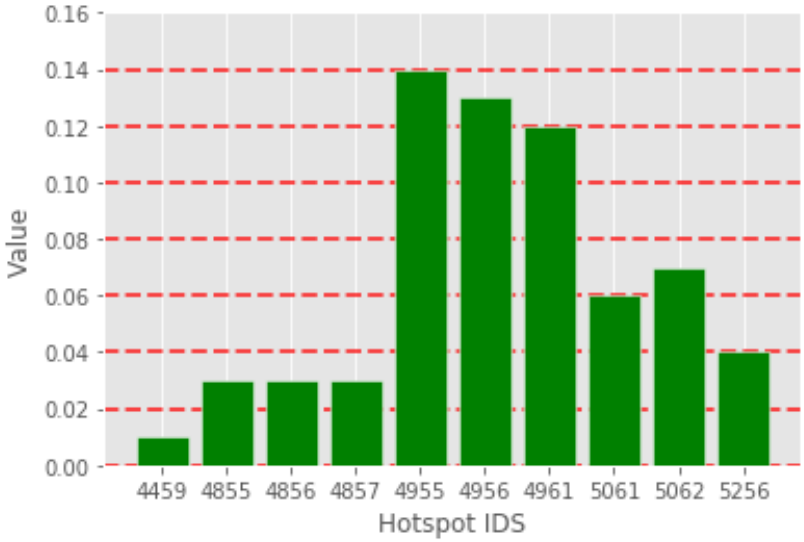


FIGURE 8. Top-ten hotspots identification using eigenvector as social network similarity measure.

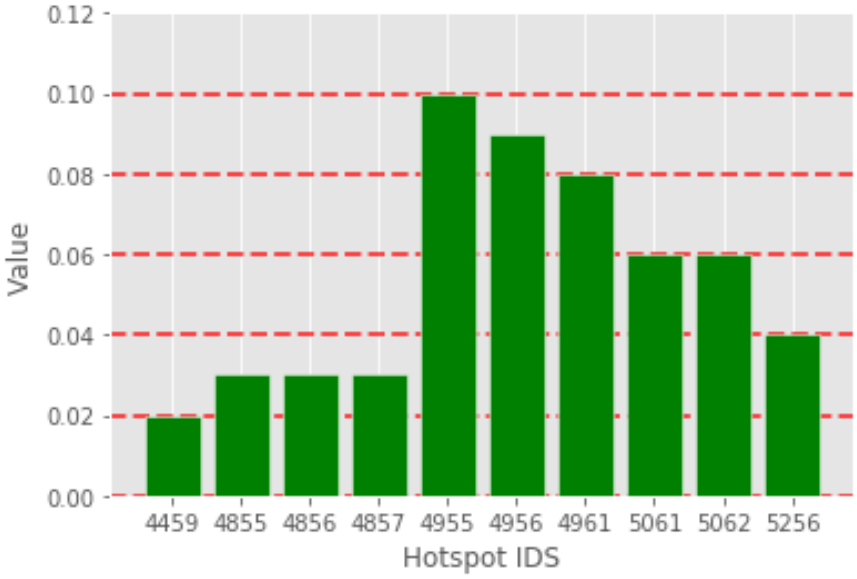


FIGURE 9. Top-ten hotspots identification using K-shell as social network similarity measure.

4) JACCARD MEASUR

This measure is used to normalize the number of shared neighbors between two nodes in the network. The size depends upon the union of two neighbors in the network.

$$Sim_{Jacc}\,(v,u) = \frac{N(v) \cap N(u)}{N(v) \cup N(u)} \tag{7}$$

Herein, $N(u)$ shows the neighbors of node v.

5) COSINE MEASURE

It is the cosine angle between the feature vectors of neighboring nodes in the network. The measure is computed based on the below equation.

$$Sim_{Cos}\,(v,u) = \frac{N(v) \cap N(u)}{\sqrt{|N\,(v)| * |N\,(u)\,|}} \tag{8}$$

Herein, the similarity score is computed based on the above equation. This achieved score is the average between two similarity measures. The similarity score is used to detect the pairs that have the probability of being similar.

V. DATA ANALYSIS AND EXPERIMENTAL RESULT

In this section, we performed detailed data analysis and also initiated a detailed discussion on the achieved experimental results. Herein, we use 'Network X', as is a popular

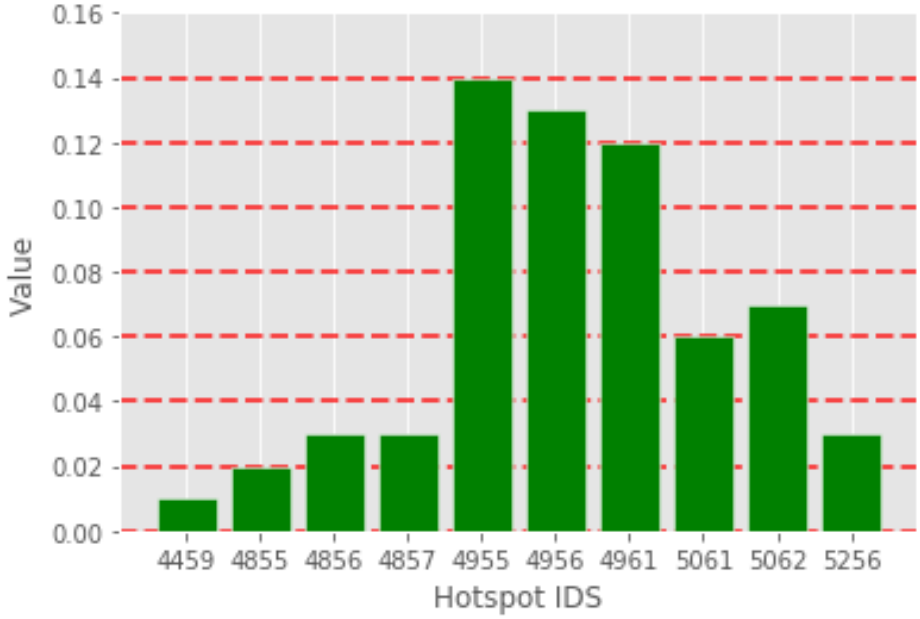


FIGURE 10. Top-ten hotspots identification using jaccard as social behavioral measure.

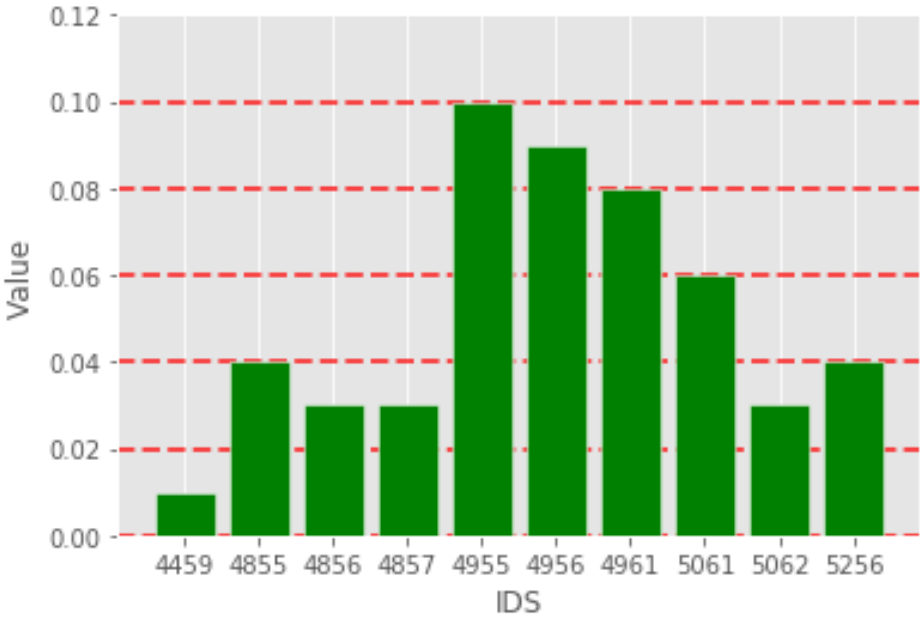


FIGURE 11. Top-ten hotspots identification using cosine as social behavioral measure.

network package used for the analysis of networks. Network X core package provides a complete data structure and is comprised of various algorithms used for directed and undirected graphs. It has the support of a powerful modern programming language named Python. Python provides various features including flexibility. In addition, It includes many useful and powerful Python libraries named Pandas, *NumPy*, *SciPy*, *and Matplotlib* [35].

To measure the importance of each hotspot. Herein, we first calculate every hotspot by using suggested metrics. Herein, we have used the ''Telecom Italia'' dataset mainly comprised of Milan City for the big data analysis.

We first identified the high communication areas in a smart city using Equations (1) and (2) and the probability p= 0.75. Then, our proposed smart CPS model applies the suggested social network similarity and social network behavioral measures. Finally, we identify Top-Ten hotspots using our proposed smart CPSS model.

Herein, $M(x)$ is the set of nodes except node x, and $d(..)$ is a function that calculates the distance between two connected

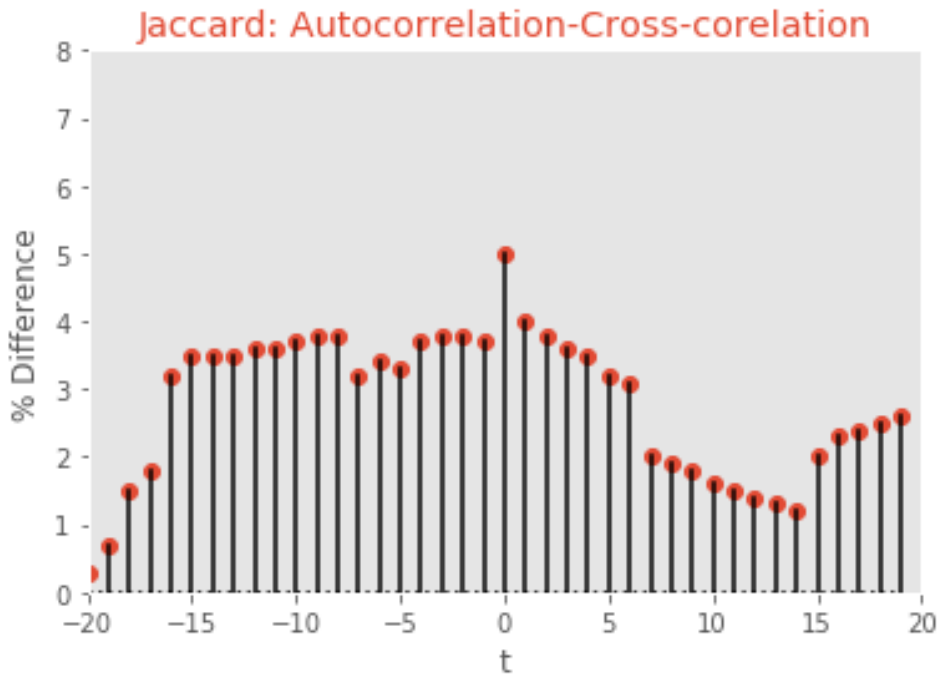


FIGURE 12. Auto- correlation | Jaccard.

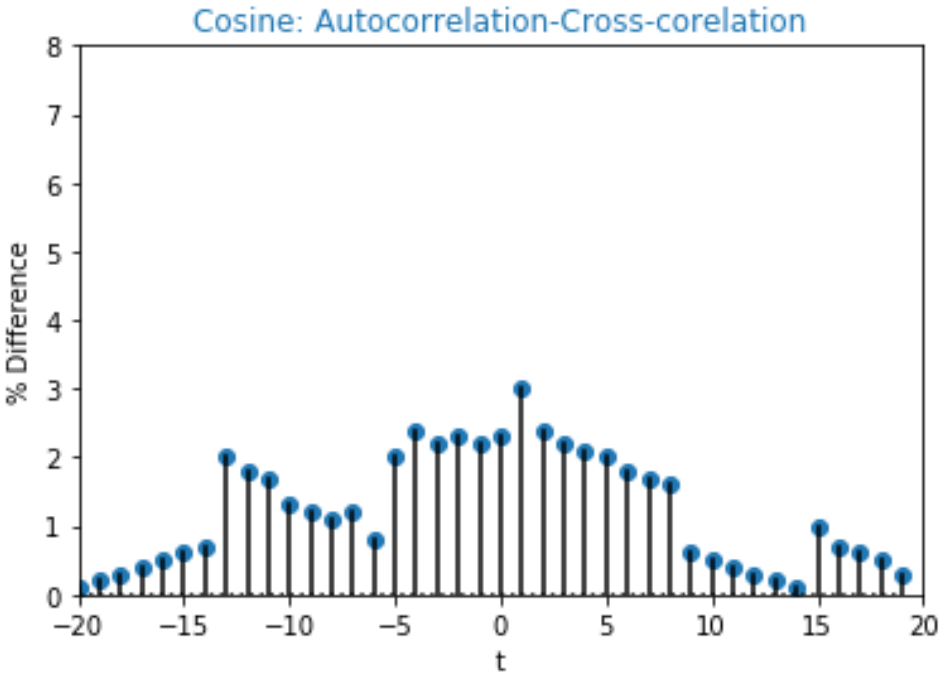


FIGURE 13. Auto- correlation | Cosine.

nodes in a graph. To measure the importance of high communication areas, we calculate each hotspot using suggested measures.

The first experimental result shows identified Top-Ten hotspots using Eignvactor. we use this measure as a social network similarity measure and the obtained result is shown in Fig. 8. In this figure, a reader can see that the x-axis present the hotspot IDs and the y-axis shows attained eigenvector values. As we mentioned earlier the Eigen vector favors high weights. A reader can see that (4955, 4956, 4961)IDs have the highest value. On the other hand (4459, and 5256) have the lowest values.

Now we calculate k-shell as a social network similarity measure using same dataset. In Figs.9, we can see that ID number (4955, 4956, 4961) has the highest value among others. We can see that the result is quite similar to

the eigenvector measure. The conclusion drawn from these results is that both measures always favor those hotspots having high weights. Another key finding from this result is that the adjacent IDs as hotspots were the neighbors to others. If we look again at Fig.6 and compare with these results, we can see that mostly active cells were discovered in the center of Milan city. Herein, the achieved result proves that most of the hotspot IDS are located in the center. Thus, this observation simplifies from analysis that the achieved results are efficient.

The summary drawn from these results is given below:

- ID numbers 4955, 4956, 4961 have the highest scores in both measures. This is due to the dynamic behavior of these measures. Because it always favors high-weight and also supports short links in the network. Eigenvector and Jaccard always favor traffic areas with high weights. If we combine these facts, it can be noticed that has the same score.

Fig. 10 and Fig. 11 show the achieved results from Jaccard and cosine social network behavioral measures respectively. In this figure, we can see that more than one identified Hotspot IDs were the same as shown in Fig. 8 and Fig.9. This means that the ranking of hotspots remains practically the same for all metrics.

A. ACCURACY AND ROBUSTNESS

To compare the accuracy of achieved results. Herein, we suggest a cross-co relation function. This function is used to detect if there is a correlation between the achieved results received from two-time series are the same or not. In our scenario, we have a discrete number of hotspots, therefore, we suggest a discrete version of the cross-correlation function. This function is defined as [42]:

$$(F * g)[n] = \sum_{m=-\infty}^{+\infty} f[m].g[m+n] \tag{9}$$

To measure the robustness of these results. Herein, we use the cross-correlation function. In addition, we have used corresponding auto-correlation. This is used to show the correlation of time series with a delayed copy of itself [42].

$$(F * f)[n] = \sum_{m=-\infty}^{+\infty} f[m].f[m+n] \tag{10}$$

Fig 12 and Fig 13 both show the relative difference between cross-correlation and auto-correlation. Fig. 12 shows the experimental result of the auto-correlation using the Jaccard measure. In this figure, we can see the difference between the autocorrelation and the cross-relation. The x-axis shows the shift τ. The y-axis shows the difference in %. On the other hand, Fig. 13 shows the result of autocorrelation and the cross-relation of the cosine measure. Let us examine these figures carefully, Herein, a reader can see that the percent difference is 5% for Jaccard and cosine is less than 3%. This is used to quantify the difference between the observed level of similarity and the perfect ones. The difference is less than 5% for Jaccard and cosine is less than 3%. These experimental results prove the consistency of our results. Our finding is that the ranking of both hotspots and the relative difference of metrics per hotspot does not vary significantly.

In summary, the overall evaluation of our proposed model shows that social network similarity and social behavioral measures gave us more efficient and accurate results in the finding of hotspots in a smart city.

VI. CONCLUSION

Herein, we proposed a smart CPSS model on a big data platform by using telecom data. The smart CPSS model is divided into different layers and each layer has different functionality. At first, the data collection layer receives raw telecom data. The next step is to pass through the data processing layer. The data processing layer performs different functions, for instance, processing, storage and analysis, etc. Then, it constructs a graph and performs a social network analysis (SNA). Herein, the high communication areas in a city were identified and secondly, Top-10 hotspots were discovered using social network similarity and social behavioral measures. It is evident from the results that our proposed big data analysis has shown that the ranking of hotspots remains practiced under these metrics. In addition, we found that the variance of results is significantly smaller for Milan. This research is helpful the traffic forecasting. In the future, will perform a detailed analysis of the complete dataset that comprises every week's data for Trento.

REFERENCES

[1] A. K. Jha and N. R. Sunitha, "Evaluation and optimization of smart cities using betweenness centrality," in *Proc. Int. Conf. Algorithms, Methodol., Models Appl. Emerg. Technol. (ICAMMAET)*, Feb. 2017, pp. 1–3.
[2] F. Amin, A. Ahmad, and G.-S. Choi, "To study and analyse human behaviours on social networks," in *Proc. 4th Annu. Int. Conf. Netw. Inf. Syst. Comput. (ICNISC)*, Apr. 2018, pp. 233–236.
[3] A. Modarresi and J. P. G. Sterbenz, "Towards a model and graph representation for smart homes in the IoT," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Sep. 2018, pp. 1–5.
[4] K. Sultan, H. Ali, and Z. Zhang, "Call detail records driven anomaly detection and traffic prediction in mobile cellular networks," *IEEE Access*, vol. 6, pp. 41728–41737, 2018.
[5] G. Maji, S. Mandal, and S. Sen, "Identification of city hotspots by analyzing telecom call detail records using complex network modeling," *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119298.
[6] A. A. Khade, "Performing customer behavior analysis using big data analytics," *Proc. Comput. Sci.*, vol. 79, pp. 986–992, Jan. 2016.
[7] L. E. Daniel and L. E. Daniel, "Cellular system evidence and call detail records," in *Digital Forensics for Legal Professionals*. Oxford, U.K.: Syngress, 2012, pp. 225–237.
[8] A. Ahmad, M. Babar, S. Din, S. Khalid, M. M. Ullah, A. Paul, A. G. Reddy, and N. Min-Allah, "Socio-cyber network: The potential of cyber-physical system to define human behaviors using big data analytics," *Future Gener. Comput. Syst.*, vol. 92, pp. 868–878, Mar. 2019.
[9] S. De, Y. Zhou, I. Larizgoitia Abad, and K. Moessner, "Cyber–physical–social frameworks for urban big data systems: A survey," *Appl. Sci.*, vol. 7, no. 10, p. 1017, Oct. 2017.
[10] E. Abba, A. M. Aibinu, and J. K. Alhassan, "Development of multiple mobile networks call detailed records and its forensic analysis," *Digit. Commun. Netw.*, vol. 5, no. 4, pp. 256–265, Nov. 2019.
[11] J. Zeng, L. T. Yang, M. Lin, H. Ning, and J. Ma, "A survey: Cyber-physical-social systems and their system-level design methodology," *Future Gener. Comput. Syst.*, vol. 105, pp. 1028–1042, Apr. 2020.

[12] F. Amin and G. S. Choi, "Intelligent service search model using emerging technologies," *Comput., Mater. Continua*, vol. 77, no. 1, pp. 1165–1181, 2023.

[13] B. A. Yilma, Y. Naudet, and H. Panetto, "Introduction to personalisation in cyber-physical-social systems," in *Proc. Move Meaningful Internet Syst., OTM Workshops*, 2018, pp. 25–35.

[14] S. Wang, D. Wang, L. Su, L. Kaplan, and T. F. Abdelzaher, "Towards cyber-physical systems in social spaces: The data reliability challenge," in *Proc. IEEE Real-Time Syst. Symp.*, Dec. 2014, pp. 74–85.

[15] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Appl. Sci.*, vol. 11, no. 23, p. 11202, Nov. 2021.

[16] P. D. Francesco, F. Malandrino, and L. A. DaSilva, "Assembling and using a cellular dataset for mobile network analysis and planning," *IEEE Trans. Big Data*, vol. 4, no. 4, pp. 614–620, Dec. 2018.

[17] M. Visan, A. Ionita, and F. G. Filip, "Data analysis in setting action plans of telecom operators," in *Data Science: New Issues, Challenges and Applications*. Cham, Switzerland: Springer, 2020, pp. 97–110.

[18] F. Amin and G. S. Choi, "Hotspots analysis using cyber-physical-social system for a smart city," *IEEE Access*, vol. 8, pp. 122197–122209, 2020.

[19] Y. Ye, H. Zhu, T. Xu, F. Zhuang, R. Yu, and H. Xiong, "Identifying high potential talent: A neural network based dynamic social profiling approach," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 718–727.

[20] N. R. Al-Molhem, Y. Rahal, and M. Dakkak, "Social network analysis in telecom data," *J. Big Data*, vol. 6, no. 1, p. 99, Dec. 2019.

[21] Y. Zhang, M. Hannigan, and Q. Lv, "Air pollution hotspot detection and source feature analysis using cross-domain urban data," in *Proc. 29th Int. Conf. Adv. Geographic Inf. Syst.*, Beijing, China, Nov. 2021.

[22] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the province of trentino," *Sci. Data*, vol. 2, no. 1, Oct. 2015, Art. no. 150055.

[23] J. Wu, E. Frias-Martinez, and V. Frias-Martinez, "Spatial sensitivity analysis for urban hotspots using cell phone traces," *Environ. Planning B, Urban Analytics City Sci.*, vol. 48, no. 9, pp. 2623–2639, Nov. 2021.

[24] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. D. Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, "Analysis of a large-scale weighted network of one-to-one human communication," *New J. Phys.*, vol. 9, no. 6, p. 179, Jun. 2007.

[25] L. Cai, H. Wang, C. Sha, F. Jiang, Y. Zhang, and W. Zhou, "The mining of urban hotspots based on multi-source location data fusion," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 2061–2077, Feb. 2023.

[26] A. A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the structure and evolution of massive telecom graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 703–718, May 2008.

[27] N. Werayawarangura, T. Pungchaichan, and P. Vateekul, "Social network analysis of calling data records for identifying influencers and communities," in *Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2016, pp. 1–6.

[28] A. Kasem Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning and social network analysis in big data platform," 2019, *arXiv:1904.00690*.

[29] A. Modarresi and J. Symons, "Modeling and graph analysis for enhancing resilience in smart homes," *Proc. Comput. Sci.*, vol. 160, no. 2, pp. 197–205, Aug. 2019.

[30] E. Mededovic, V. G. Douros, and P. Mähönen, "Node centrality metrics for hotspots analysis in telecom big data," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2019, pp. 417–422.

[31] M. Seufert, T. Griepentrog, V. Burger, and T. Hoßfeld, "A simple WiFi hotspot model for cities," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 384–387, Feb. 2016.

[32] P. Yuan and H. Ma, "Opportunistic forwarding with hotspot entropy," in *Proc. IEEE 14th Int. Symp.*, Jun. 2013, pp. 1–9.

[33] S. Brdar, O. Novović, N. Grujić, H. González–Vélez, C.-O. Truică, S. Benkner, E. Bajrovic, and A. Papadopoulos, "Big data processing, analysis and applications in mobile cellular networks," in *Lecture Notes in Computer Science*. Cham, Switzerland: Springer, 2019, pp. 163–185.

[34] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, Dec. 1966.

[35] A. Hagberg, P. Swart, and D. S Chult, *Exploring Network Structure, Dynamics, and Function Using Network*. Los Alamos, NM, USA: Los Alamos National Lab, 2008.

[36] F. Amin, A. Ahmad, and G. S. Choi, "Community detection and mining using complex networks tools in social Internet of Things," in *Proc. IEEE Region 10 Conf.*, Oct. 2018, pp. 2086–2091.

[37] H. Li, "Centrality analysis of online social network big data," in *Proc. IEEE 3rd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2018, pp. 38–42.

[38] F. Amin, A. Ahmad, and G. Sang Choi, "Towards trust and friendliness approaches in the social Internet of Things," *Appl. Sci.*, vol. 9, no. 1, p. 166, Jan. 2019.

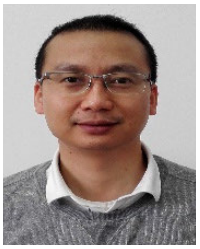
[39] P. Howlader and K. S. Sudeep, "Degree centrality, eigenvector centrality and the relation between them in Twitter," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 678–682.

[40] F. Amin, J.-G. Choi, and G. S. Choi, "Community detection based on social influence in large scale networks," in *Web, Artificial Intelligence and Network Applications*. Caserta, Italy: Springer, 2020, pp. 122–137.

[41] Z. Wang, Y. Zhao, J. Xi, and C. Du, "Fast ranking influential nodes in complex networks using a k-shell iteration factor," *Phys. A, Stat. Mech. Appl.*, vol. 461, pp. 171–181, Nov. 2016.

[42] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Contro*. Hoboken, NJ, USA: Wiley, 2015.

FARHAN AMIN (Member, IEEE) received the Ph.D. degree from the Department of Information and Communication Engineering, College of Engineering, Yeungnam University, Gyeongsan, South Korea, in October 2020. He was an Assistant Professor with the Department of Computer Engineering, Gachon University, South Korea. Since March 2022, he has been an Assistant Professor with the Department of Information and Communication Engineering, Yeungnam University. He has more than ten years of teaching and research experience. He has delivered various keynote speeches, invited talks, invited lectures, and short courses. He has authored over 40 publications (books, book chapters, journal publications, and conference publications). He has various Korean patents. His research interests include the Internet of Things, social Internet of Things, big data, data science, and machine learning aspects in emerging technologies. He is a member of ACM. He was a recipient of the fully-funded scholarship for the master's and Ph.D. studies.



LIGANG HE (Member, IEEE) is currently a Reader with the Department of Computer Science, University of Warwick, U.K. His research interest includes parallel and distributed computing. He has published more than 150 papers in the research area.

GYU SANG CHOI (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Pennsylvania State University. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently an Assistant Professor. His research interests include data mining, deep learning, computer vision, storage systems, parallel and distributed computing, supercomputing, cluster-based web servers, and data centers. He is mainly working on text mining, reinforcement learning, and deep learning, with computer vision, while his prior research has been mainly focused on improving the performance of clusters. He is a member of ACM.