IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 26, 2020, accepted June 9, 2020, date of publication July 8, 2020, date of current version July 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003030

Hotspots Analysis Using Cyber-Physical-Social System for a Smart City

FARHAN AMIN, (Graduate Student Member, IEEE), AND GYU SANG CHOI, (Member, IEEE)
Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea

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

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2019R1A2C1006159, in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) support program supervised by the Institute for Information & communications Technology Promotion (IITP), under Grant IITP-2020-2016-0-00313, and in part by the Brain Korea 21 Plus Program funded by the National Research Foundation of Korea (NRF) under Grant 22A20130012814.

ABSTRACT Internet of things plays a vital role in providing various services to users. Significant volumes of data are generated from the communication between a large numbers of heterogeneous devices over the Internet. Big data technology is generally used to handle the large volume of data. Complex networks are graphs (networks) having non-trivial topological features, such as random graphs and lattices. Big data of complex networks concerns big data methods that can be used to analyze massive structural data sets, including considerably large networks and sets of graphs. This study is based on the critical phenomenon arising in complex networks that enable us to analytically predict the hotspots in smart cities. Hotspots are places with significantly high communication traffic relative to others. In this study, we propose a cyber-physical-social system for the analysis of high communication traffic hotspots using telecom data. The proposed model constructs a graph, and perform social network analysis on it. The process of hotspot extraction is performed, followed by social network analysis, which is conducted by quantifying the importance of each hotspot based on network metrics. These metrics aid in determining the importance of each hotspot in a telecom data network. Our objective is to prioritize different areas and detect hotspots quickly. Our results indicate that the proposed model has an efficiency comparable with that of state of the art methods. This research study will be helpful for urban planning and development, as well as in upgrading telecommunication infrastructure.

INDEX TERMS Cyber-physical systems (CPS), cyber-physical, social systems (CPSS), data analytics, smart city, urban planning, big data, hotpots, network traffic analysis, centrality measure, graph, complex networks.

I. INTRODUCTION

Population growth and the requirements of a comfortable life have resulted in tremendous growth in urban areas, leading to the urbanization [1]. Urban areas in general refer to modern towns and cities that are complex compared to rural areas in all aspects of life. Smart city is a contemporary urban concept that is essential for people to have a high quality of life. Generally, smart city refers to the interconnection of several subsystems of a city. It incorporates information and communications technology (ICT) [2], which is a derivative term from information technology (IT).

The primary objective of a smart city is to exchange information efficiently and provide smart governance [2]. Building a smart city is considered to be a challenging task in various aspects, as it requires intelligent choice and detailed planning [2].

The most important task in planning a smart city is to build an ICT structure. This simple communication concept has become considerably popular recently and has become a necessary part of our lives. Smart cities nowadays cover all features that use ICT to improve the efficiencies of the cities [2]. Moreover, the ICT paradigm enables smart cities to utilize infrastructure such as water resources, roads, and power supplies efficiently. ICT incorporates ubiquitous connectivity among users, services, and the environment [3].

Services include, but are not limited to, the weather, emergency response, shopping, transportation, etc. Another problem in smart cities is the security and efficiency of the internet of things (IoT) based on network devices.

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

FIGURE 1. Big data technology in a smart city.

Information related to smart devices are usually made available to users via underlying technologies such as IoT.

The objective of IoT is to connect things to the Internet and to control them from anywhere [3]. IoT provides timely and efficient information to users, leading new ways of thinking and making services smarter. According to a recent survey, a large amount of data has been generated in the last two years in both the private and public sectors [4]. This has spawned a new area of research named big data. Big data has become a widely studied topic in both industry and academia. Traditional methods for processing large and complex data are inadequate, and therefore big data analytics is more suited to the task. In addition, as data comprise numerous types of information such as integer, numerical, floating-point number, Boolean data, characters, and strings, it is challenging for traditional methods to store, analyze, gather, and process the data [5]. Processing and analysis in the big data paradigm leads to new ways of making decisions. Typically, data collection and processing techniques are very costly. Hence, it is desirable to incorporate a smart technology that can efficiently handle and analyze a large amount of data [6].

The use of big data technology to provide services in a smart city is shown in Fig. 1. Data is at first collected from several sources, and then stored in a certain place once it is received. The stored data is initialized and processed for analysis. The analysis is performed by using some graph-based analytics. After completing these processes, a unique solution for a smart city is provided. In this way, a certain quality of service is provided to people living in a smart city. Generally, graph analytics play a substantially important role in building and providing services in smart cities [7]. Graph analytics is grounded in graph theory [7]. Graph theory is considered a powerful tool in modeling highly connected systems such as computer systems, social network systems, biological systems, and complex systems [7]. In addition, graph theory models allow modeling component interactions in addition to device level logic. It is known that real-word data has a relational structure. In several cases, they are considerably complex and large, and therefore, it is difficult to understand their structure in an original format. This kind of data can be modeled as networks or graphs, and can be visualized easily. In this way, it becomes easy to understand the structure of a network. The use of big data and analytics changes strategic communication in companies, as well as the role of communicators. Companies increasingly drive their business decisions based on data. The use of big data in the telecommunication industry is gaining popularity in recent times. The use of big data to mine customer behavior is referred to as customer analytics [8]. Customer analytics increases the efficiency of big data by enabling organizations to predict buyer behavior by increasing their sales using inventory planning and market optimization [8]. In this direction, network deployment, demographic statistics, and call detail records (CDRs) are the main factors that need to be carefully investigated to make accurate predictions. There is a variety of open data sources available for these factors. Therefore, synthetic and real data for analysis need to be considered. These models usually do not capture large scale mobile networks accurately [9]. Heterogeneous cellular networks comprising different nodes, such as macro and microcells, have been used for analysis [9].

Interconnected objects in IoT are known as smart objects [10]. A large number of smart objects with numerous connections are used in the digital world. The excessive use of low-cost internet extensions and air interference has led to decreasing costs of smart devices. These devices are also known as physical devices because they can sense physical stimuli, aggregate data, and interact with other physical devices. Smart devices are defined by intelligent choices in establishing connections and interactions with other objects. The objects are connected based on certain protocols, and they operate autonomously. Examples include smartwatches, building automation systems, security systems, and intelligent health care systems, etc. [9]. When these objects connect via the Internet, they form cyber-physical systems (CPSs) [11]. CPSs are combined software and electronics platforms that to control and monitor their physical environments using objects such as actuators and sensors. These objects connect to the real world using the Internet [12]. Thus, the physical world merges with the virtual world to create what is known as cyberspace. Cyberspace refers to the combination of digitalized data, information, and communication, connected through the Internet.

Fig. 2 shows the structure of a CPS. It can be seen that several actuators and sensors are simultaneously connected

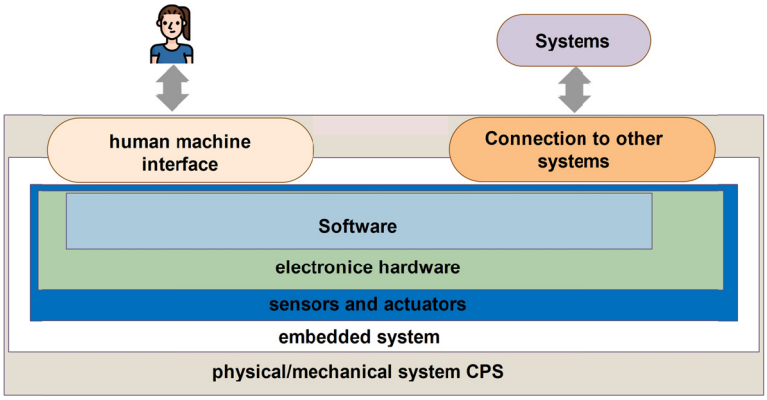


FIGURE 2. Schematic of a CPS.

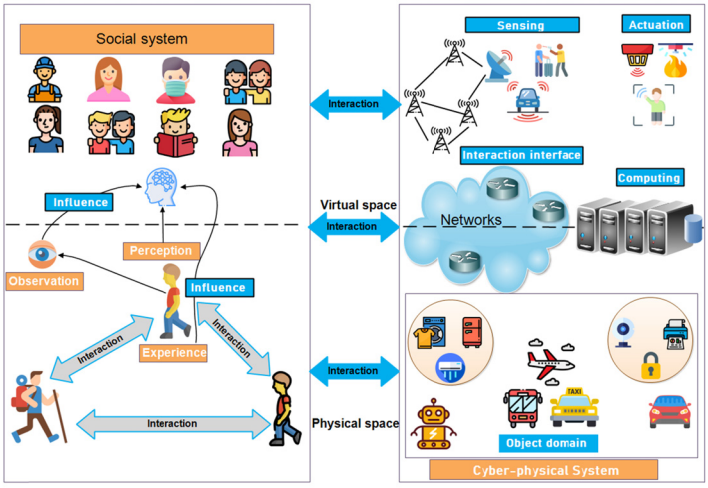


FIGURE 3. Schematic of a CPSS.

using software and electronics. Some interfaces are available to optimize these systems for human use. The most popular real world examples of CPSs are smartphones [13]. Smartphones connect the real world to the virtual via sensors such as the geolocation sensors, Internet, telephone, and wireless interfaces. There are several methods in which the human actors used in CPS and hence getting popularity. This acts as the foundation of cyber-physical-social systems (CPSSs) [14]. CPSSs use data on social behavior, and perform relationship analysis to provide relevant information services [11]. Various definitions have been proposed for CPSSs. According to one such definition, a CPSS is the integration of a CPS and a cyber social system (CSS), that enables smart interaction between the physical, cyber, and social spaces [12]. CPS is limited to multimedia entertainment, communicators, and business processing devices [15]. CSS refers to social networks such as Twitter, Facebook, and YouTube [16]. Most definitions agree that the presence of humans is an integral part of the system. From a systematic perspective, CPSS is considered an environment or a system, where machines and humans are involved in the virtual and physical world, respectively. The structure and the elements of a CPSS are shown in Fig. 3. In this figure, we can see three interconnected systems, namely, the cyber, physical, and social. In addition, the CPS and the social system term interact with each other in a certain environment, which is composed of the virtual and physical spaces, as shown in Fig. 3. The social system comprises various people with relations among them. These relations are formed based on their interactions and measured based on observation, perception, and personal experiences. The physical system is divided into two parts, sensors, and actuators. Several sensing devices, such as temperature sensors, sensors used in cars, and actuators, are shown in Fig. 3. These objects are connected using various technologies. CPS wireless and wired network technologies are used to process this data. This is represented by the virtual space.

CPSS has been an active research topic for more than a decade and plays a very important role as an interface between several objects to send and receive data, carry out actions. It also acts as a bridge between homogeneous and heterogeneous objects. Various concepts such as service-oriented architecture and web of things have been introduced to support heterogeneity [9].

In this work, we consider CPSS from a telecom perspective. Examples of applications are modern infrastructures for telecommunication, building automation, etc. Large quantities of data are generated in these systems, and the analysis of online and offline data volumes poses challenges. Several studies have been conducted in the past to explore these challenges. For example, in [17], the authors describe a case study of mobile network analysis and planning using a cellular data set. The authors used a large-scale data set comprising CDRs, including demographic and topological information of the any country, and investigated how they affect. They also discussed how the findings could be generalized to other case studies and scenarios such as constructing similar data sets. Similarly, Maria *et al.* [18] discussed the problems faced by telecom operators related in the communication service market [18]. Moreover, they discussed various examples and the right policies already proposed in telecom area of research.

It is understood that from the literature review that CPSSs play a very important role in providing a platform through which the sensors are connected. Potential users and services are also important aspects. Modern communication technologies and cutting-edge research focus on ultra-low latencies. Typically, telecom data comprise calls, SMS, and Internet data. These data are known as telecom transactions, and they pass through mobile devices. Telecom operators have a large collection of data such as SMS, internet service, customer profile, and location data. There are several challenges associated with these data such as efficient storage, parsing, and analysis. These challenges became more difficult especially with the advent of new techniques such as social network analysis (SNA) and machine learning algorithms. These methods require a large storage spaces and distributed processing solutions. Therefore, there is a need to build a big data model that can effectively process data having different sizes and complexities. In addition, the model has the functionality to provide fast calculation and also produces the processing time [19]. Therefore, in this study, a big data platform, containing numerous tools and capabilities to handle the challenges in processing telecom data has been proposed.

A. MOTIVATION

The primary motivation for this study is to find a new method to build a CPSS from telecom data using big data analytics. The proposed CPSS acts as a solution to the challenges that telecom operators face in the extraction of a considerable amount of data. Usually, high traffic areas or hotspots have a high activity density compared with the remainder of the city [20]. The identification of these hotspots is useful for telecom companies, so that plans to focus on providing better services in these areas can be made.

The second motivation is to provide a unique model based on real data that will be helpful for decision-makers in the telecom industry. Generally, the telecom companies always try to plan and offer the right services by using the most influential hotspots in the network. Because, it increases the service providing features.

B. BENEFITS FROM THIS STUDY

- The study provides an analysis of telecommunication data using big data analytics. Therefore, the results will be helpful for telecom companies to identify high communication strength areas. Telecom companies can thus pay more attention to provide better services in these areas.
- The proposed model aids in identifying highly congested areas in a city.
- Our methodology relates important network centrality features from graph theory to the field of communication.

C. CONTRIBUTIONS OF THIS STUDY

The contributions of our work are summarized below:

- The proposed graph analysis based CPSS model connects various attributes involved in telecom data. It is an accurate system providing in-depth knowledge of data generated by connected smart devices.
- We demonstrate that the use of network analysis metrics is a simple and helpful tool to identify high strength communication areas. Both higher and lower weight influencers were identified using network similarities, and it favors accurate analysis of telecom data.

The proposed CPSS model is based on three tiers, data collection, data processing, and application. The model extracts high traffic areas in a graph and then performs SNA. It applies network centrality metrics that quantify the importance of each node. Our model identifies low and high weight influencers. Finally, our model identifies ten areas that generate the highest traffic in a city. The proposed system was implemented and tested using a large graph analyzer (Network X), with a real dataset named Telecom Italia Big Data Challenge, provided by telecom Italia.

The remainder of the paper is organized as follows. In Section II, we discuss some state of the art studies in this direction. In Section III, we discuss the concepts of the proposed CPSS model and explain the process of graph building using this model. In Section IV, we discuss datasets and the architecture of our model. In addition, we discuss SNA features such as PageRank, degree centrality, and eigenvector centrality. In Section V, we discuss our results and the robustness and accuracy of the proposed model. We provide concluding remarks in Section VI.

II. RELATED WORK

A large number of CDR based methods have been proposed to build telecom networks. Customers are usually represented as nodes, and SMSes, interactions, and calls by edges [19].

Onnela *et al.* describe a detailed analysis of a network comprising a data set of detailed mobile call records [21]. The data set is quite large, having seven million call records. They considered customer call records as a weighted graph, and performed an analysis in terms of weighted clustering, degree strength, and weight distribution. Their objective was to measure the correlation between these quantities. This is considerably helpful in understanding the local structure

of a network. Similarly, Nanavati *et al.* discussed various graph-based properties such as neighborhood distribution and degree distribution [22]. They analyzed these properties by using the Indian telecom network data set. The data set comprised SMSes and call records.

Nattapon *et al.* experimented on CDRs of a telecom dataset in Thailand [23]. The objective of their experiment was to clean data using ''filters to filter anomaly number.'' They introduced a measure to capture the influencers based on their calling behavior. Ahmad *et al.* proposed a churn prediction SNA model by combining big data and machine learning techniques, along with feature selection [24]. Their SNA considered modern data warehousing, in which information is added by using switches and billing systems. They used several network centrality measures to provide an equality analysis between each pair of nodes. Furthermore, they performed a deep analysis in which each node pair interacted with other pairs using link attributes. They also implemented a score based on centrality measures that incorporated clustering. In this way, they could identify customers with tendencies to change to or leave a company.

Customer churn models are used by telecom companies and operators to detect customers who have a high tendency to leave a company (churn), and to provide suitable solutions, encouraging them to stay [19].

Centrality measures and SNA features were used to improve the results of churn prediction models, after representing CDRs data as a graph.

Modarresi *et al.* defined a graph-based analysis approach for enhancing resilience in a smart home [25]. In this study, the authors explored various network topologies in a smart home concerning complex networks. They presented a unique home network model, and performed a graph theoretic analysis on this model. They identified different metrics that are more applicable to this specific type of communication networks.

Mededovic *et al.* discussed the node centrality metrics, and concluded that they are significantly helpful in the analysis of hotspots in the telecom domain [26]. In this study, they performed an analysis of telecom data for two weeks. Their objective was to find the hotspots in that network and measure the interactions among them. They used one of the most popular node centrality metrics name Eigenvector. They ranked hotspots under various centrality metrics and identified high strength communication areas.

Seufert *et al.* proposed a simple Wi-Fi hotspot model for smart cities [27]. Initially, they targeted a specific area in an urban environment. After that they obtained the location of a Wi-Fi hotspot from a public Wi-Fi database and then identify the locations of ten cities including their features. In this study, they demonstrated that the different hotspot locations could be modeled with a uniform distribution of the angles, and the gamma distribution of the distances. This simple and accurate model of Wi-Fi hotspot locations was used to create spatial distributions of Wi-Fi hotspots in arbitrary cities; i.e., for performance evaluation of mechanisms that rely on the coverage and the throughput of Wi-Fi hotspots in cities.

Peiyan *et al.* discussed data forwarding in opportunistic networks [28]. They tried to explore the optimized sizes of hotspots in networks. They suggested a novel routing metric named Hoten, which supports human mobility. They used the relative entropy to design a utility function for public hotspots. They use their utility function in personal and public hotspots to evaluate the centrality of the nodes. They suggested considering nodes as parameters for accuracy.

Brdar *et al.* discussed knowledge retrieval from telecom data [29]. They presented several measures, mainly focused on graph theory and machine learning, and discussed various steps involved in knowledge recovery from raw telecom data. In addition, they discussed various contexts of different applications such as people moving across cities and home and work locations with timings. They highlighted concerns in this direction such as regulation privacy and real-time settings.

The studies discussed above are based on centrality calculation measures such as closeness and degree [30]. From the literature review, we identified that measures such as PageRank and the Katz centrality measure are useful for detecting influencers in medium or small-scale networks, but they become inefficient in large-scale networks, owing to complex theoretical calculations. Moreover, measures such as PageRank are incompatible with telecom data. PageRank is designed to rank web pages over the Internet [19].

To overcome these limitations, we propose a conceptual CPSS model to measure communication strength in smart cities. The proposed model is unique in all aspects because we have selected new measures to detect influencers. This makes our model compatible with telecom data, and applicable to large-scale networks.

In this study, we have used real telecom datasets from Milan, Italy. Our approach is more accurate and efficient than traditional methods that use network centrality measures, and this will be seen in our results and discussion sections. The details of the proposed conceptual model are discussed in the next section.

III. PROPOSED CONCEPTUAL CPSS MODEL

A. DESCRIPTION OF THE CPSS

In this section, we present our conceptual CPSS model. The model consists of three tiers, and each tier has different functionalities that enable read and write operations efficiently. Fig. 4 shows the conceptual framework of our proposed CPSS model. We discuss the functionality of each tier in detail in this section.

1) DATA COLLECTION TIER

Telecom data usually comprise CDR and customer data. Customer data comprise several elements such as customer name, age, address, sex, and customer ID. CDR data are produced by a telephone exchange; the elements are call type (incoming/outgoing), called number, calling number,

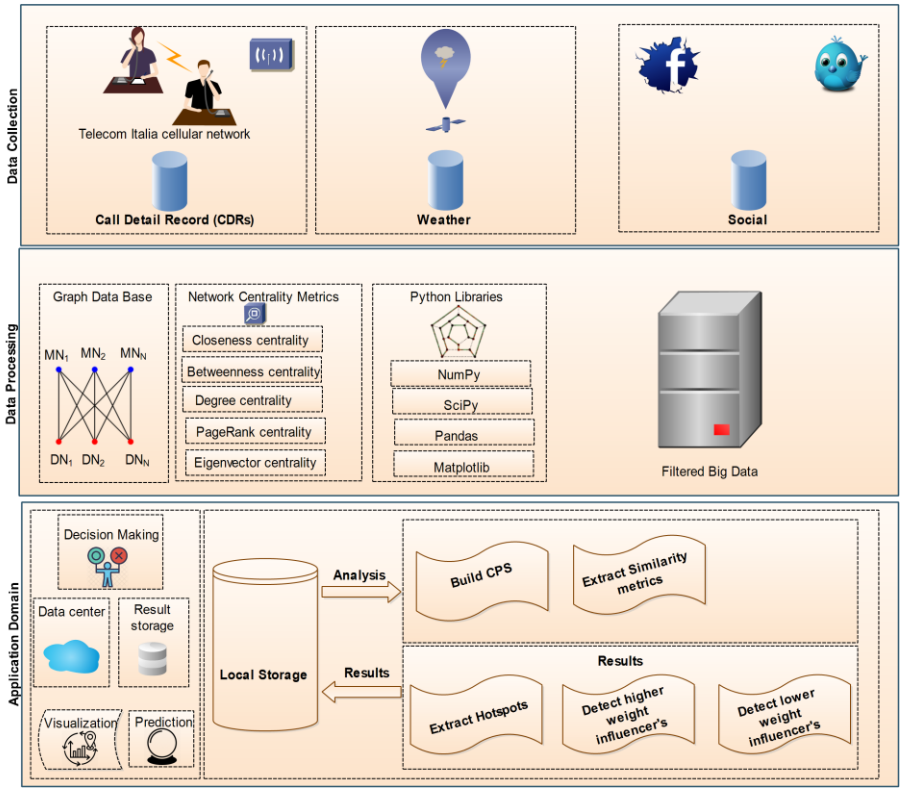


FIGURE 4. A conceptual framework of the proposed CPSS.

call duration, and switch ID. During pre-processing, some information is excluded to reduce the size of the data set. The data collection tier also collects data such as weather, environment, social factors, and user CDRs. The user CDR data set comprises call and SMS records, internet voice calls, etc.

2) DATA PROCESSING TIER

The data processing tier receives data from the data collection tier. This tier normalizes the data to a meaningful form, so that useful information can be extracted from that data.

Usually, the processing of a large amount of data requires more power and resources. As described previously, traditional methods of data processing are not feasible to analyze large data. Therefore, there is a need to develop a system that can efficiently pre-process the incoming stream of data. Our CPSS model provides an advanced way of handling big data.

The incoming stream of data is initially stored, and then handled by using Python data analysis library (pandas) as shown in Fig. 4. This storage system creates filtered big data with initial processing steps such as handling of redundant data and dealing with errors. Subsequently, the received data is converted into a graph database. Pandas perform cleansing, transforming, and the manipulation of the data [31], and Network X [31], generates a graph from the given dataset. The graph database can then extract the hotspots. The core package of Network X provides a complete data structure for the illustration of different graph types such as directed, in directed, and other graphs including loopholes. Network centrality metrics can then be applied. After these steps, the processed data is forwarded to the next tier.

3) GRAPH BUILDING

We consider Milan city as a graph G, where the vertices are the hotspots, and edges are communication strengths between these hotspots. The direct weighted graph G is denoted by $G = (V, E)$ with two types of weights for each edge. These weights depend on the durations of calls between each side of the edges.

This completed dataset sent for processing to a designated server. Once processing is done, the data is stored on local a disk for future use.

4) APPLICATION DOMAIN TIER

The application domain contains a cloud server, a storage device, and a data center. The results are ready for compilation once the hotspots are identified. The graph database sends the partially complete results stored in the result storage server to the data server. As both servers are essentially of the same kind, it depends upon the user if choosing two separate servers are feasible, keeping in view the cost of the system. The application domain tier handles database management and storage. Moreover, several activities such as storing analyzing etc. are performed by the objects that are also stored in these servers. Finally, our proposed model stores, analyzes, and displays the results to users.

IV. DATASETS AND MODEL ARCHITECTURE

In this section, we describe datasets and SNA features used in this work.

TABLE 1. Milano/trento datasets.

Domain	Descriptive details
Telecommunications	SMS, Call Internet; Incoming calls; Outgoing calls;
Weather	Weather station Data; Precipitation
Environment	Air Quality
News	Milano Today
Social	Tweets

TABLE 2. Telecommunication data analysis.

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

A. DATA DESCRIPTION AND PREPARATION

Here, we describe the datasets used, and the CPSS architecture based on feature extraction methods.

The polytechnic university of Milan complied a telecom data set named ''Telecom Italia Big Data Challenge'' [20]. The objective of this challenge was to stimulate the creation and the development of innovative methods and ideas in the field of big data.

This dataset comprises various geo-referenced datasets. In the first edition (created in 2014), data was provided from two cities in Italy, Milan and Trentino. A large number of participants from 100 universities across the country participated in the project. This open source dataset is unique in that it has been extracted from social networks, news, weather stations, telecom records, and electricity user data.

The dataset contains information from several companies providing various services in different areas of the two cities. Each company has different standards, and different frequencies' in providing these services. The details of this dataset are listed in Table 1. The ''telecommunication activity'' dataset comprises the following elements:

- he square ID represents the identification number of the Milan and Trento grids.
- The volume of incoming/outgoing connections for SMS.
- The approximate time of an event.
- The volume of incoming /outgoing connections for calls.
- The country codes, Internet traffic, etc.

We have used this dataset to identify high traffic areas.

The second dataset is called ''Milan/Trento to Trento/ Milan calls.'' The elements of this dataset are given below:

- Squire id1 is used to show the number of squares in Milan/Trento grid. It shows the origins of interactions.
- Square id2 is the square of Milan/ Trento grid. It shows the destinations of interactions.
- The approximate times of the events.
- Directional interaction strength: it shows the directional strength between both square Ids.

As the first data set only considers the areas, we used this dataset to enhance our model. Moreover, this dataset gives information about the sources and destinations of communications.

B. EXTRACTING HOTSPOTS

In this section, we discuss how the proposed model identifies high communication areas from a telecom dataset.

We started by using the ''mobile phone activity'' dataset. Elements of this dataset are: call records, Internet data, aggregated amount of connections for SMS, etc. Table 1 lists the dataset used in this study. In this table, six different datasets with different characteristics can be seen. The descriptive detail explains the purpose of that dataset.

Table 2 provides additional details of the telecommunication dataset used in our analysis. The size of the dataset is quite large as it covers both cities (Milan and Trento).

We have defined a parameter named threshold to identify areas with high communication activities. The parameter threshold shows the minimum amount of communication traffic for a certain area, and therefore it is a dynamic parameter.

Let i denote a high communication area.

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

I_i is the amount of communication for the area i. The parameter ω is computed by using the following equation.

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

Here, $Trif$ is the amount of communication in all areas of the smart city. P is a parameter used to find the cutoff threshold.

Fig. 5 is the first inductive visualization of high strength communication traffic areas of Milan city. Our proposed model started by identifying high telecommunication activity areas in three different areas of Milan city:

- Dumo (downtown)
- Bocconi (university area)
- Navigli (popular for nightlife)

The internet activity of people in these areas is plotted in Fig. 5. It can be seen that the number of connections in the Duomo area is higher than the other two areas, which may be attributed to the higher population there.

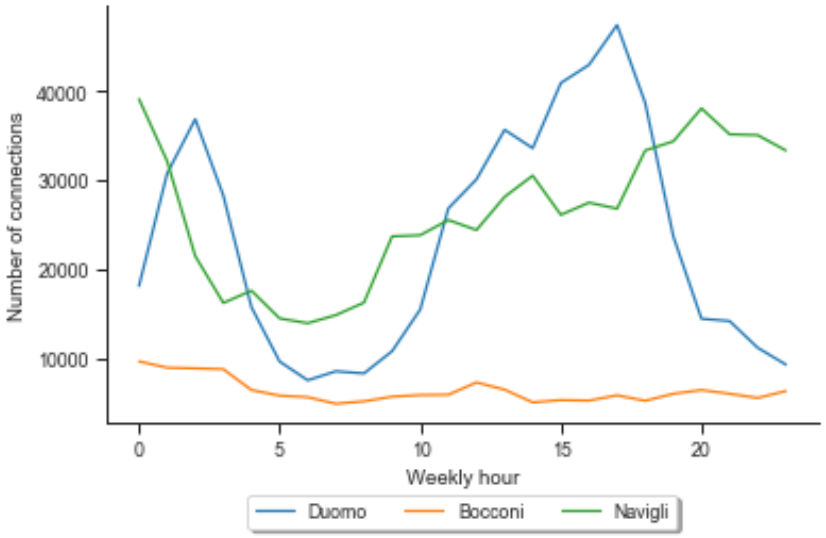


FIGURE 5. Extraction of hotspots: Telecommunication activity graph.

TABLE 3. Generated 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

It is worth noting that the peak of the plot for Navigli (nightlife) occurs about two hours later than the peak for Duomo (downtown). During the first three days (Friday, Saturday, and Sunday) of November, Bocconi has fewer phone calls. It may be because students go out of this area on Friday, and come back on Monday morning. Table 3 shows the detailed output generated from the dataset.

The plots indicate that the volume of calls reduces during the weekend.

After high communication traffic areas have been identified, the next step is to find the importance of each high communication traffic area by using different SNA features.

C. SNA FEATURES

Telecom companies can easily understand the structure of hotspots and predict the strength of relationships among nodes in a certain hotspot using SNA. We now apply network centrality measures to the three high traffic areas identified using our proposed CPSS model. The details of the SNA features are given below.

1) DETECTING LOWER WEIGHT INFLUENCERS

The concept of centrality in graph theory has been widely used to identify the most important nodes in networks [32], [33]. Usually, each node in the network is considered important, leading to the concept of centrality. Centrality measures favor hotspots with lower weights because the shortest path is a key component in centrality metrics.

2) CLOSENESS CENTRALITY

The closeness centrality shows how close a node is to all other nodes in a network. It can be calculated for one node from the average short path lengths among all pairs of nodes in the network. The node with the highest closeness can reach other nodes in the network by using a short path. Lower values of such paths indicate the most central nodes, meaning such nodes are closer to other nodes in the network.

Suppose $G(u)$ is the set of all nodes except node u, and d is the distance between two nodes in the network. For closeness centrality, we consider the shortest distance between two nodes. Then, the closeness centrality for a node u is defined as:

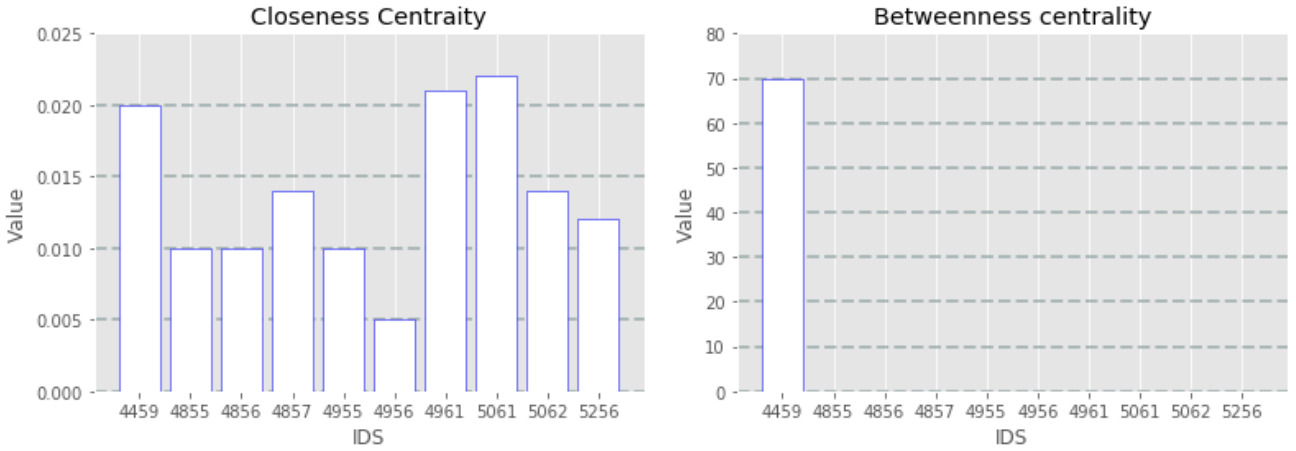
$$Closeness\,(u) = \frac{1}{\sum v \in G\,(u)\, d(u, v)} \tag{3}$$

3) BETWEENNESS CENTRALITY

This centrality measure also uses the shortest path in the network to find important nodes in the network. Betweenness centrality is calculated by selecting pairs of nodes and finding all short paths between them.

Let $\sigma(x)$ be the number of the shortest paths between nodes a and b. The betweenness centrality is then given by:

$$Between\,(x) = \sum_{a \neq b \neq x} \frac{\sigma(x)}{\sigma} \tag{4}$$



a) closeness centrality; b) betweenness centrality.

FIGURE 6. Important hotspots using: a) closeness centrality; b) betweenness centrality.

4) DETECTING HIGHER WEIGHT INFLUENCERS

The measures described above are not enough to calculate the influence score and the importance of a node. Therefore, we identified higher weight influencers. The description of the network analysis measures are given below.

5) DEGREE CENTRALITY

We start by defining the basic concept ''node degree.'' The degree defines the number of edges of a node connected to other nodes in the network [34]. The higher the degree of node spectacles, the more important that node is in the network [32]. If a node has a higher degree, more nodes can use it as an intermediate node for communication. Suppose N is the set of neighbors for a node. Then, the degree centrality for a node u is given as [35]:

$$Degree\,(u) = \sum_{y \in N(u)} d\,(u, v) \tag{5}$$

where $Degree\,(u)$ is the computed degree for node u_i.

6) EIGENVECTOR CENTRALITY

Sometimes individuals have multiple connections. It is necessary to identify these nodes (with multiple connections) in a network [36]. Similarly, by using this important node property, we easily access an important node in a network and link that node to other important nodes in the network. The eigenvector centrality measure is used to find the most important nodes in a network. It is an extended form of degree centrality.

All nodes not considered equivalent in the eigenvector centrality measure. A node is considered important if it is connected to a large number of nodes. Eigenvector measures a node's influence based on the number of links it has to other nodes in the network [37]. Popular examples of eigenvector centrality are the Katz and Google PageRank algorithm [38]. These measures used to identify important webpages on the Internet.

Let λ be an eigenvalue of the adjacency matrix. The adjacency matrix is a square matrix used to represent a finite graph [38]. Then, the eigenvector for a node u is defined as:

$$Eigenvector\,(u) = \frac{1}{\lambda} \sum_{y \in N(u)} d\,(u, v) \tag{6}$$

7) PAGERANK

PageRank is a kind of eigenvector centrality, first introduced by Brin and Page [39]. Originally designed to rank webpages on the Internet, it has evolved as an important tool in data mining, web algorithms, distributed systems, and distributed computing [40]. A node is considered important if it is linked to the other important nodes, or if it is highly linked to other nodes. It assigns a score to nodes based on their connections. PageRank differs from eigenvector in that it takes into account the link directions and weights. PageRank is applied in a graph as a random walk of the network.

Let q be the dumping factor, and n be the number of nodes. In PageRank, we consider the distance between two nodes instead of considering the intermediate nodes. The PageRank for a node u is given as:

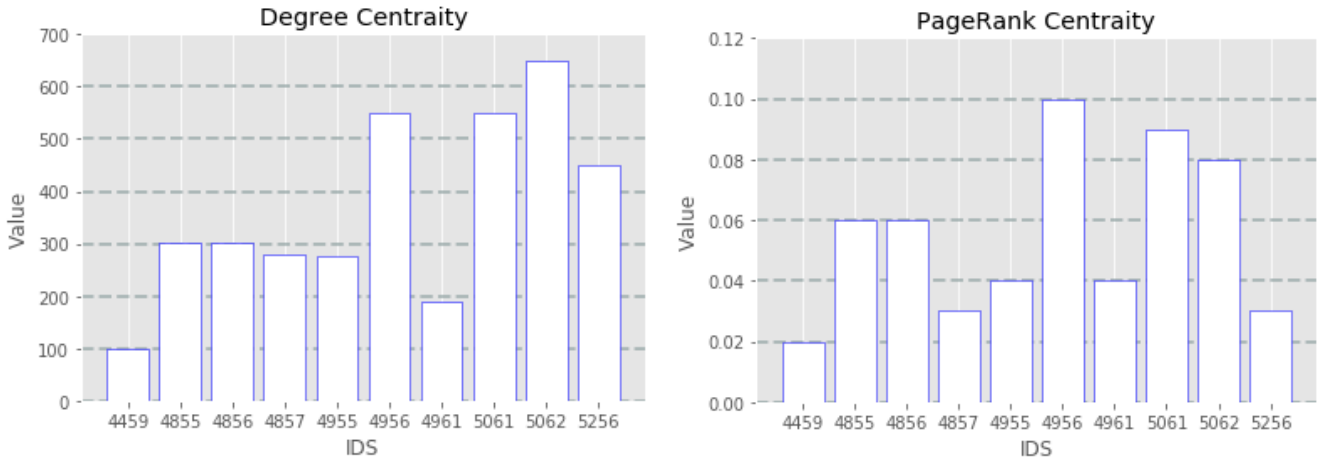
$$PageRank\,(u) = q. \sum_{v \in N(u)} \frac{PR\,(v)}{L(v)} + \frac{1-q}{n} \tag{7}$$

V. RESULTS AND DISCUSSION

Network X is a popular packages used for the analysis and exploration of networks. It has a core package that provides the data structure and algorithms for networks in directed and undirected graphs. In addition, the support of the Python programming language provides more flexibility, and makes this tool ideal for the representation of networks in various fields. It includes various useful Python libraries such as SciPy, NumPy, Matplotlib, and Pandas [31].

To measure the importance of each hotspot, we calculate each hotspot by using different SNA features.

We start our big data analysis by using the ''Telecom Italia'' dataset dated 2013-11-01 of Milan city. Initially, we identified high traffic areas based on Equations (1) and (2)



a) degree centrality; b) page rank.

FIGURE 7. Important hotspot using: a) degree centrality; b) page rank.

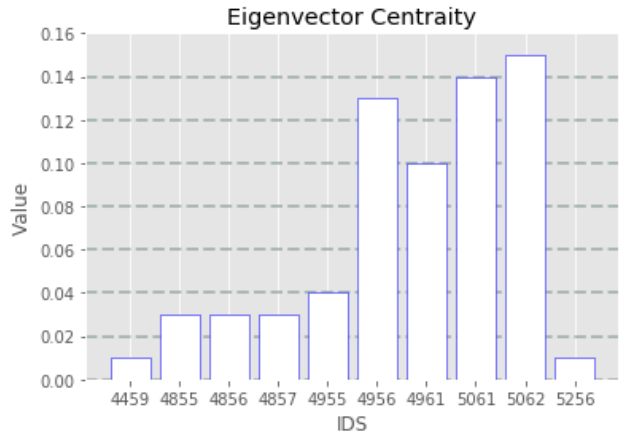


FIGURE 8. Important hotspots using Eigenvector centrality.

with $p = 0.75$. Our proposed model applies social network centrality metrics and identifies ten hotspots. These important hotspots were identified using the closeness and betweenness centralities, and shown in Fig. 6. In this figure, the x-axis is the IDs of hotspot, and the y-axis is the values of the corresponding centrality metric. It can be seen that ID number 4459 has the highest values for both the closeness and betweenness measures. In addition, the features of the network measures are helpful in the identification of low weights, because these measures always favor shortest paths in a network.

Important hotspots are also identified based on other centrality measures, and these are shown in Figs. 7 and 8. In these figures, ID numbers 5061 and 5062 have the highest centralities. It is due to the reason that all three measures, i.e. eigenvector, PageRank, and degree centrality favor hotspots with high weights. In addition, we find from adjacent IDs that the important hotspots are the neighbors to each other, and are located in the center of Milan city.

The results are summarized as follows:

- ID numbers 5061 and 5062 have the highest scores in closeness centrality. This behavior indicates that the closeness centrality measure always favors low weight hotspots because it supports short path lengths.
- Degree distribution, eigenvector, and PageRank centrality measures always favor traffic areas with high weights. High weights around a node usually indicate a high centrality value of that node.
- If we combine the above facts, it can be noticed that ID number 4956 has a lower centrality score in the first set, but has a higher centrality score in the second set. This notable observation demonstrates that we have considered one metric per family. Therefore, we focus on PageRank and closeness centrality in the next section.

A. ROBUSTNESS

The robustness of our proposed model is measured by comparing its consistency with the original dataset, and is shown in Fig. 9 (a) & (b) for the closeness and PageRank centralities, respectively.

We continue our analysis using the same dataset for the second week, along with the ten hotspots identified in the first week. We compared values for week 1 and week 2.

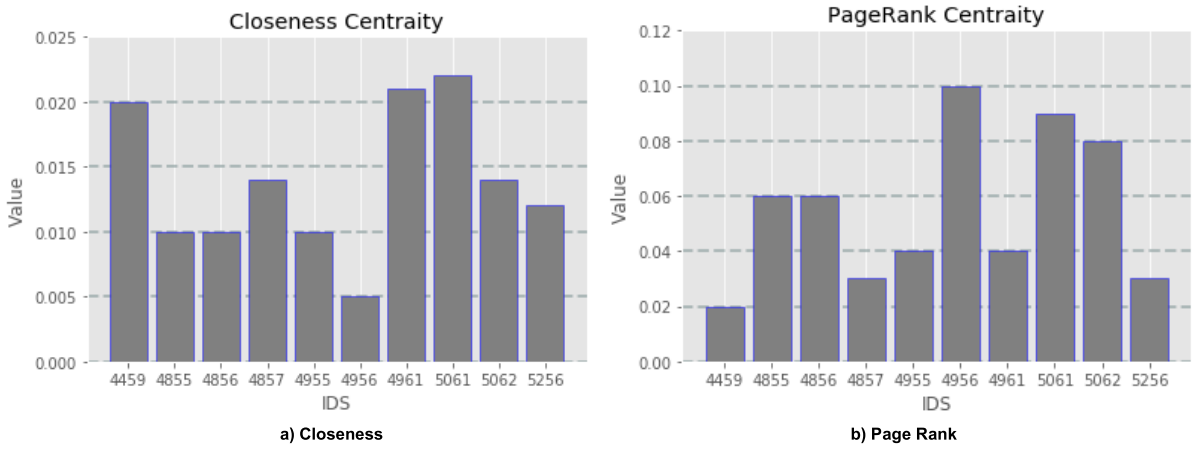
We noticed that the rankings of hotspots for both weeks were the same by using both the network metrics. In addition, the difference for the first week is less than 8% for each hotspot for closeness centrality, and less than 8% for PageRank centrality.

B. ACCURACY

A social network was built, and weighted according to the marketing needs of companies concerning the number of outgoing calls, and the Internet traffic.

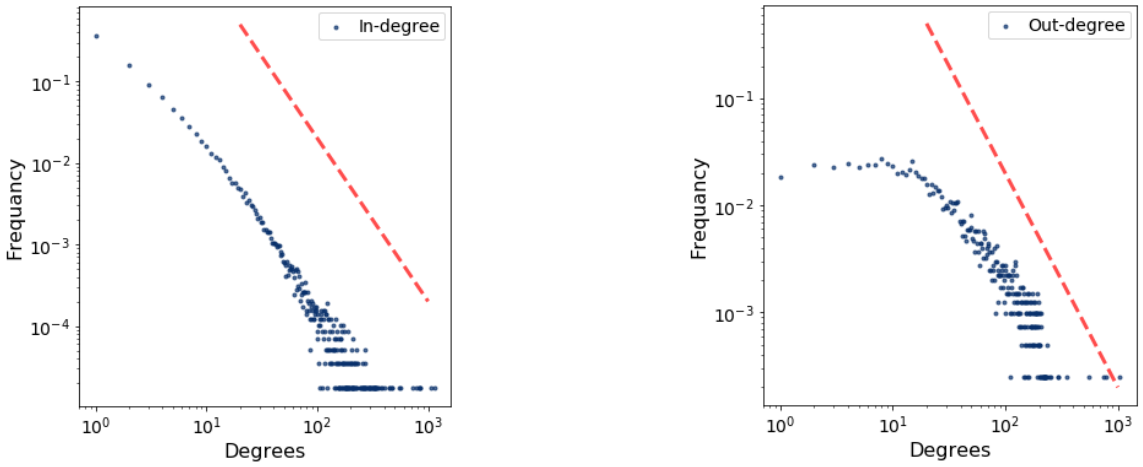
Fig. 10 shows the frequency distribution of in-degree and out-degree. The frequency distribution is a fraction of nodes in the network with a different type of node degrees. All three graphs are in Figs. 10 and 11 have been plotted on a logarithmic scale.

The x-axis in these graphs shows the degrees, and the y-axis represents the frequencies of the nodes. The logarithmic scale is used to show a large range of values.



a) Closeness b) Page Rank

FIGURE 9. a) Closeness b) Page Rank.



a) In-degree b) Out-degree

FIGURE 10. Frequency distribution of a) In-degree b) Out-degree.

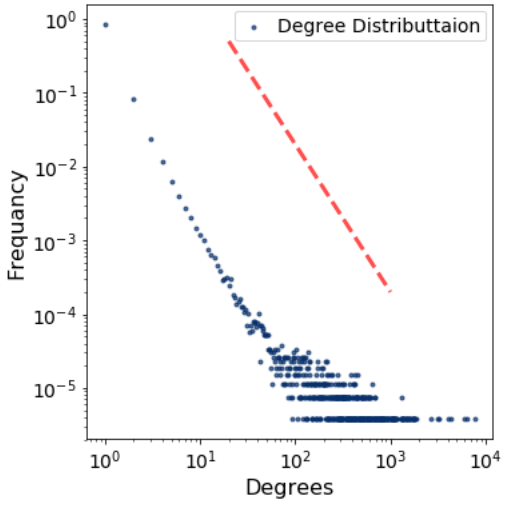


FIGURE 11. Degree distribution.

The red vertical line indicates a power law. The power-law distribution of the degrees shows that the behavior with exponent 3.017 for in-degree, and out-degree is 3.01 and 3.084 for degree distribution.

These results show that the distribution of various types of degrees usually gives similar results. Because few nodes have large degrees, they may be suitable for the individual a telecom company service provider.

The evaluation of the proposed model demonstrates that the SNA features used gave results that are more accurate in the identification of hotspots, compared to results achieved by traditional methods.

VI. CONCLUSION

We proposed a method for building a CPSS model on big data platforms by using telecom data, and by extracting hotspots and performing SNA. The analysis was performed using social network features such as degree and closeness centrality. The proposed CPSS model extracts high traffic areas from cities and applies network centrality features on it to find the importance of each node. In this way, our proposed model identifies the ten highest traffic areas in a smart city. The use of SNA is considerably helpful

for telecom companies because it is a very simple implementation to identify and rank high traffic areas in a smart city. In this way, the structure of hotspots can be understood easily, and the strengths of relationships among nodes in a certain hotspot can be found. Our big data analysis model has shown that the ranking of hotspots remains practically the same under various network centrality metrics. In addition, we identified social network metrics with better applicability in telecommunication data networks. As future work, we are interested in doing more research by performing the analysis of the whole dataset every week. The proposed method will be helpful for researchers working on learning machine-learning techniques, especially for traffic forecasting. In addition, our method will also be useful in geolocalized tweets.

REFERENCES

[1] N. M. Kumar, S. Goel, and P. K. Mallick, "Smart cities in India: Features, policies, current status, and challenges," in *Proc. Technol. Smart-City Energy Secur. Power (ICSESP)*, Mar. 2018, pp. 1–4.

[2] 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.

[3] 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.

[4] S. Din, M. M. Rathore, A. Ahmad, A. Paul, and M. Khan, "SDIoT: Software defined Internet of Thing to analyze big data in smart cities," in *Proc. IEEE 42nd Conf. Local Comput. Netw. Workshops (LCN Workshops)*, Oct. 2017, pp. 175–182.

[5] P. G. Hunter, E. G. Schellenberg, and A. T. Griffith, "Misery loves company: Mood-congruent emotional responding to music," *Emotion*, vol. 11, p. 1068, Oct. 2011.

[6] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, Jun. 2016.

[7] 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. (ISC)*, Sep. 2018, pp. 1–5.

[8] A. A. Khade, "Performing customer behavior analysis using big data analytics," *Procedia Comput. Sci.*, vol. 79, pp. 986–992, Jan. 2016.

[9] 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.

[10] P. Yadav and S. Vishwakarma, "Application of Internet of Things and big data towards a smart city," in *Proc. 3rd Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU)*, Feb. 2018, pp. 1–5.

[11] 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.

[12] 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.

[13] J. Shen, C. Wang, A. Wang, Q. Liu, and Y. Xiang, "Moving centroid based routing protocol for incompletely predictable cyber devices in Cyber-Physical-Social distributed systems," *Future Gener. Comput. Syst.*, vol. 108, pp. 1129–1139, Jul. 2020.

[14] B. A. Yilma, Y. Naudet, and H. Panetto, "Introduction to personalisation in cyber-physical-social systems," in *Proc. Workshops Move Meaningful Internet Syst. (OTM)*, Cham, Switzerland, 2019, pp. 25–35.

[15] 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.

[16] S. Wang, Y. Guo, Y. Li, and C.-H. Hsu, "Cultural distance for service composition in cyber–physical–social systems," *Future Gener. Comput. Syst.*, vol. 108, pp. 1049–1057, Jul. 2020.

[17] 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.

[18] 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*, G. Dzemyda, J. Bernatavičienė, and J. Kacprzyk, Eds. Cham, Switzerland: Springer, 2020, pp. 97–110.

[19] 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.

[20] 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, Dec. 2015, Art. no. 150055.

[21] 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," *J. Phys.*, vol. 9, no. 6, p. 179. Jun. 2007.

[22] 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.

[23] 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.

[24] 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*. [Online]. Available: http://arxiv.org/abs/1904.00690

[25] A. Modarresi and J. Symons, "Modeling and graph analysis for enhancing resilience in smart homes," *Procedia Comput. Sci.*, vol. 160, pp. 197–205, 2019.

[26] E. Mededovic, V. G. Douros, and P. Mahonen, "Node centrality metrics for hotspots analysis in telecom big data," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2019, pp. 417–422.

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

[28] P. Yuan and H. Ma, "Opportunistic forwarding with hotspot entropy," in *Proc. IEEE 14th Int. Symp. 'World Wireless, Mobile Multimedia Netw.' (WoWMoM)*, Jun. 2013, pp. 1–9.

[29] 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 *High-Performance Modelling and Simulation for Big Data Applications*, J. Kołodziej and H. González-Vélez, Eds. Cham, Switzerland: Springer, 2019, pp. 163–185.

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

[31] A. Hagberg, P. Swart, ad D. S. Chult, "Exploring network structure, dynamics, and function using NetworkX," Los Alamos Nat. Lab. (LANL), Los Alamos, NM, USA, Tech. Rep. LA-UR-08-05495; LA-UR-08-5495 TRN: US201006%%1254, 2008.

[32] 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. (TENCON)*, Oct. 2018, pp. 2086–2091.

[33] 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.

[34] F. Amin, R. Abbasi, A. Rehman, and G. S. Choi, "An advanced algorithm for higher network navigation in social Internet of Things using small-world networks," *Sensors*, vol. 19, no. 9, p. 2007, Apr. 2019.

[35] D. Wei, Y. Li, Y. Zhang, and Y. Deng, "Degree centrality based on the weighted network," in *Proc. 24th Chin. Control Decis. Conf. (CCDC)*, May 2012, pp. 3976–3979.

[36] 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.

[37] 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.
[38] 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*. Cham, Switzerland: Springer, 2020, pp. 122–137.
[39] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.
[40] A. D. Sarma, A. R. Molla, G. Pandurangan, and E. Upfal, "Fast distributed pagerank computation," in *Proc. Int. Conf. Distrib. Comput. Netw.*, 2013, pp. 11–26.

FARHAN AMIN (Graduate Student Member, IEEE) received the B.S. degree in computer science from Gomal University, Dera Ismail Khan, Pakistan, in 2007, and the M.S. degree in computer science from International Islamic University Islamabad (IIUI), Pakistan, in August 2012. He is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, College of Engineering, Yeungnam University, Gyeongsan, South Korea. He is an invited Reviewer in the IEEE, ACM, the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (TETCI), *MDPI Sensor*, and *Future Generation Computer Systems* (FGCS). His research interests include big data analytics, data science, the Internet of Things (IoT), the social Internet of Things (SIoT), and human behavior analysis using big data. He is a member of ACM. He was a recipient of the fully-funded scholarship for Masters Studies and Ph.D.

GYU SANG CHOI (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pennsylvania State University. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics Company Ltd., from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently a Professor. His research interests include data mining, deep learning, and parallel computing, while his prior research has been mainly focused on improving the performance of clusters. He is a member of ACM.