Forecasting Flight Delays Using Clustered Models Based on Airport Networks

Mehmet Güvercin, Nilgun Ferhatosmanoglu, and Bugra Gedik

***Abstract*— Estimating flight delays is important for airlines, airports, and passengers, as the delays are among major costs in air transportation. Each delay may cause a further propagation of delays. Hence, the delay pattern of an airport and the location of the airport in the network can provide useful information for other airports. We address the problem of forecasting flight delays of an airport, utilizing the network information as well as the delay patterns of similar airports in the network. The proposed "Clustered Airport Modeling" (CAM) approach builds a representative time-series for each group of airports and fits a common model (e.g., REG-ARIMA) for each, using the network based features as regressors. The models are then applied individually to each airport data for predicting the airport's flight delays. We also performed a network based analysis of the airports and identified the *Betweenness Centrality (BC)* score as an effective feature in forecasting the flight delays. The experiments on flight data over seven years using 305 US airports show that CAM provides accurate forecasts of flight delays.**

***Index Terms*— Flight delay estimation, airport networks, graph partitioning, hubs, betweenness centrality, REG-ARIMA, airport clustering, time series clustering, graph theory.**

I. Introduction

MAJOR factors of flight delays include technical problems of airplanes, weather conditions, scheduling conflicts, overuse of airport capacities, and delay propagation between flights. While these factors are traditionally well studied, patterns due to the structure of airport networks have not been enough understood yet. In this paper, we investigate whether the position of an airport in the transportation network and information about similar airports improve the estimation of delay patterns. We aim to forecast flight delays by incorporating *airport network information* and *similarity of delay patterns of airports* into the estimation models. Accurate forecasting of flight delays is essential both for optimization of airline operations and airport capacity planning.

The network of airports is first represented as a graph structure with each airport as a node, and the number of flights between two airports as the weight of the edge between the nodes. A set of graph-based features is extracted for each airport. In particular, we adapt the measures of hub score, betweenness centrality, articulation point, in-degree, and weighted-in-degree into the context of air transportation networks.

We then use graph features and time-series patterns of delays to quantify similarities between airports, and cluster the airports based on these similarities. We finally model each cluster of airports with regression with autoregressive integrated moving average errors (REG-ARIMA) using the extracted features as regressors. The clusters are used to develop a joint model of airports for flight delay estimation. The information aggregated in the clusters helps to remove noise and handle outliers. We refer this approach as CAM, Clustered Airport Modeling, which makes use of graph based features for delay time-series estimation.

An extensive set of experiments is presented on millions of domestic flights between 305 airports in the United States over seven years. Developing a joint model for a cluster of airports based on graph features and delay patterns is shown to improve the estimation accuracy for individual airports. The betweenness centrality, which quantifies how important an airport is in the routes of other airport destination and arrival pairs, is found to be effective both for clustering the airports and as a regressor in the REG-ARIMA model.

The presented network based analysis results can contribute to understanding the airport networks and their effect on delays. In particular, the analysed measures of articulation points and betwenness centrality can provide simple explanations to better understand network based delay behaviors. The proposed approach of using delay information of similar airports can help aviation system operators perform more effective planning and budgeting.

The organization of the paper is as follows. We discuss the related work in Section 2. We explain the proposed methodology in Section 3, including the network based analysis, time series representation of flight delays, and the clustered airport modeling. In Section 4, we present the performance evaluation and results. We conclude in Section 5.

II. Related Work

The proposed approach is related to the areas of flight delay estimation, time series modeling, and network analysis. We summarize the related work and how our method is placed in the literature for each these areas.

Manuscript received June 21, 2018; revised December 19, 2018, July 8, 2019, and December 11, 2019; accepted January 24, 2020. Date of publication May 11, 2020; date of current version May 3, 2021. This work was supported in part by the Scientific and Technological Research Council of Turkey under Grant 112M950, and in part by the Bilkent University. The Associate Editor for this article was K. Wang. *(Corresponding author: Mehmet Güvercin.)*

Mehmet Güvercin and Bugra Gedik are with the Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: mehmet.guvercin@bilkent.edu.tr).

Nilgun Ferhatosmanoglu is with the Department of Industrial Engineering, University of Turkish Aeronautical Association, 06800 Ankara, Turkey.

Digital Object Identifier 10.1109/TITS.2020.2990960

A. Flight Delay Estimation

Flight delay prediction has attracted significant attention both in practice and research literature [1]. Carriers and customers get affected by excess travel times, departure and arrival delays. Around 19% of the US domestic flights have a delay more than 15 minutes [2]. The causes of flight delays are studied from the perspectives of airlines and customers [3]. Airline hubbing and peaking airport concentration due to over-scheduling flights, besides other logistic and economic factors, are found to cause delays. Barnhart *et al.* study passenger delays as a factor for flight delays and derive findings for its causes [4]. They analyze flight cancellation and missed connections and develop a discrete choice model to estimate historical passenger travels. A taxonomy of flight delay prediction problems and a review of prediction approaches are presented in [1]. Carriers aim to consider airport network effect while deciding to postpone or cancel flights, such as giving priority to flights that start or end in hub airports [5]. Our work introduces a graph based approach to flight delay estimation by incorporating the network features of the airports and analyzing them as groups in the network.

B. Time-Series Modeling

The proposed method involves regression with ARIMA modeling and clustering. There is extensive work in these areas in the data mining and statistics literature. ARIMA is widely used for many applications, such as forecasting the electricity price [6], predicting the frequency and severity of accidents [7]. Seasonal ARIMA models (SARIMA models) are also well established in the literature [8]. Time series clustering can be applied either over raw data, or models or features built over raw time series data [9]. Clustering is used to combine forecasts of time-series data [10]. Forecasting on multiple time series is recently studied via clustered models based on time series similarities, which helps to improve the scalability of forecasting methods [11]. Another line of research is to design Long short-term memory (LSTM) type recurrent neural networks [12] for a variety of machine learning problems on time-series data. Adapting these methods for multiple (delay) time-series data considering an underlying (airport) network structure is an interesting problem.

C. Network Analysis

Our work utilizes network analysis to better understand the air transportation networks and forecasting flight delays. We contribute to the literature by linking the graph features of the airports to their exhibited delay patterns. Network analysis has made a significant impact in Web and social networks [13]–[15], starting from the early work by Freeman on measuring the structural centrality [16]. White and Borgatti adapt the centrality measures on undirected graphs to directed graphs [17]. Authoritative and hub scores of node sources in a hyper-linked environment are extensively studied in the literature [13], [18].

The network structure of airports has attracted some attention in air transportation research [19]. Santos and Robin analyze the variables, including the hub-airport variable, that explain the flight delays at the European airports [20]. Kim and Hansen introduce a non-parametric approach to estimate the effects of demand changes and throughput changes on delay [21]. Delay propagation of flights is modeled by considering both local congestion in individual airports and propagation of these delays over connected airports [22]. This approach aims to model the stochastic nature and time-varying behavior of airports. A network-based model is introduced to simulate the effects of aircraft ground movements in apron taxiways to gate assignment operations [23]. Airport network is also used as an exploratory variable to obtain the global delay state of the entire system [24].

The recent US Federal Aviation Administration (FAA) Strategic Plan discusses to increase the throughput capacity of airports and congested air corridors [25]. Our approach can provide a network based insight to help prioritize certain airports in terms of the planned capacity enhancements. It can complement some of the tasks in FAA's research plan via a better understanding of the air transport networks. Among the functionalities of the Aviation Environmental Design Tool (AEDT), released by the FAA Office of Environment and Energy (AEE), are to model multiple airports in a single study, airplane taxi delay and sequence modelling [26].

III. Proposed Methodology

The arrival delay of a flight is the amount of time of being late to its destination. A practical task is to build a model that can forecast whether and how much a flight will be delayed. In this paper, we propose a methodology to utilize the airport network information in forecasting the flight delays. We illustrate our approach using the flight information of 305 US airports for 7 years, collected from Research and Innovative Technology Administration (RITA), absorbed into OST-R. The data set includes the records of millions of commercial domestic flights, each with a set of attributes, e.g., the year, month, day of month, flight number, origin, destination, scheduled time, arrival delay.

The flight delay of an airport is represented by an aggregate time series of all its flights' delays. We develop models, based on regression with ARIMA errors, that are built on groups of airport time series, as opposed to modeling each airport individually. Our intuition is that the underlying causes of delays can be similar for the airports that have similar features or similar delay patterns. By clustering the delay time series, the model of each airport that might suffer from sparseness or outliers can be enriched with data of other airports. Hence, we use the airport interaction network and the similarities of the airports' delay patterns to cluster the airports and develop a joint representative model for each group of airports.

We generate an airport network and adapt the methods for network analysis to analyze the interactions (flights) between the airports. Graph based features are used to represent each airport and to cluster them based on the similarities of these features. They are used as also regressors in REG-ARIMA.

Section III.A presents how the airports are clustered and the group models are generated for flight delay estimation.

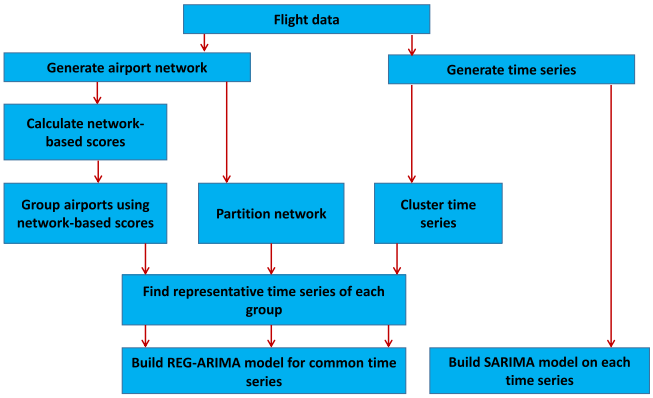


Fig. 1. Flowchart of clustered airport modeling.

In Section III.B, we analyze the flights between airports and generate the airport interaction graph. We discuss how node and interaction features are extracted from this network, and illustrate their potential relationship with the flight delays. Section III.C presents the time series modeling of flight delays.

A. Clustered Airport Modeling

Our approach, Clustered Airport Modeling (CAM), clusters the airports and builds a common REG-ARIMA (or SARIMA) model for the aggregate time series of flight delays for each cluster. Figure 1 shows the steps of CAM. It first constructs the airport network, consisting of airports as the nodes and their flight relations as edges, and extracts the node features for each airport. CAM then clusters the airports (via k-means, PAM) using the node features, graph partitioning methods, and delay time series patterns of airports.

To generate the arrival delay time series for each airport, we divide a day into periods (i.e., eight three-hour periods for our experiments) and use the delayed flights in a specific period to calculate the corresponding value. The time point value of the (three-hour) period in the corresponding airport's time series is the maximum (or median) of delays. The signal features of airports' time series are also explored in clustering, besides the graph based features and graph partitioning methods.

CAM applies the clustered modeling using REG-ARIMA (or SARIMA) forecasting model. A common regressors set is generated for each cluster. We use the each graph-based feature as regressor in regression model. While generating a common time series of a cluster for each time point t_i we find the maximum (median) time series in that cluster and assign the value of time point t_i of maximum (median) time series to the time point t_i of the common time series. To determine the ith value of regressor r which is the corresponding regressor of feature f, we use ith value of feature that belongs to selected maximum (median) time series. A REG-ARIMA model is developed based on the common time series and regressors set for each cluster. For each cluster we have a common regression model and use this model to find the residual time series, and we build a SARIMA model for each residual. To estimate the future values, we use the predicted values of the common regression model and the predicted values of the specific residual time series' SARIMA model. As a baseline, we also build a SARIMA model for each airport's time series individually.

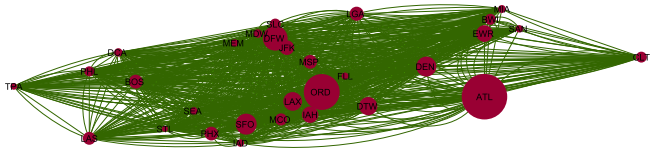


Fig. 2. Top-30 airports of US aviation system considering number of delays.

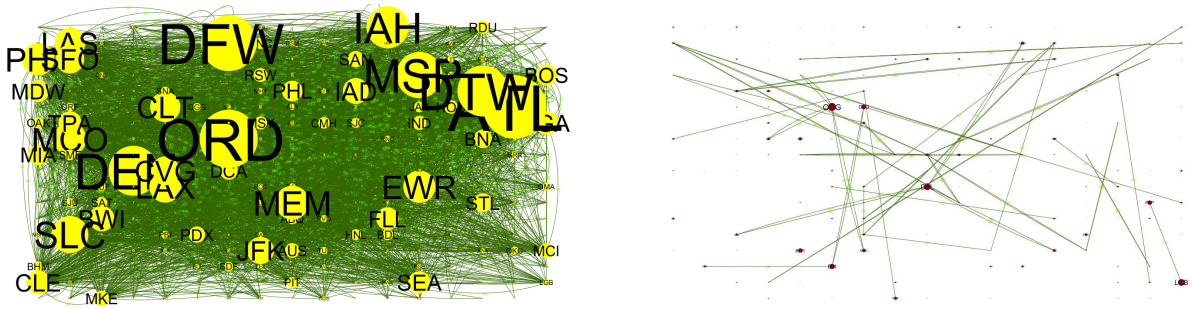
1) CAM: Graph-Theoretic Clustering: The graph-based features that we explore are: hub score, betweenness centrality, articulation point, in-degree and weighted in-degree. These features are fed to a clustering algorithm to obtain the airport clusters. We build SARIMA and REG-ARIMA models on a representative time-series for each cluster. We refer to these approaches as graph theoretic clustered SARIMA modeling (GTC-SM) and graph theoretic clustered REG-ARIMA modeling (GTC-RAM), respectively.

2) CAM: Graph Partitioning: Graph partitioning can also be used to group the airports [27]. A partition in a network can be defined as a set of nodes with dense connections internally and sparser connections to outside of the partition. We identify partitions of airports and treat each group of airports as a hard partition. Several methods have been developed especially in the social network literature for partitioning and community detection, such as edge betweenness community [28], walk trap community [29], spin glass community [30], leading eigenvector community [27] fast greedy community [31]. Partitioning algorithms show similar performance in our case so we select the walk trap community algorithm. We refer this approach as graph partitioned SARIMA modeling (GP-SM) and graph partitioned REG-ARIMA modeling (GP-RAM) in our performance evaluation.

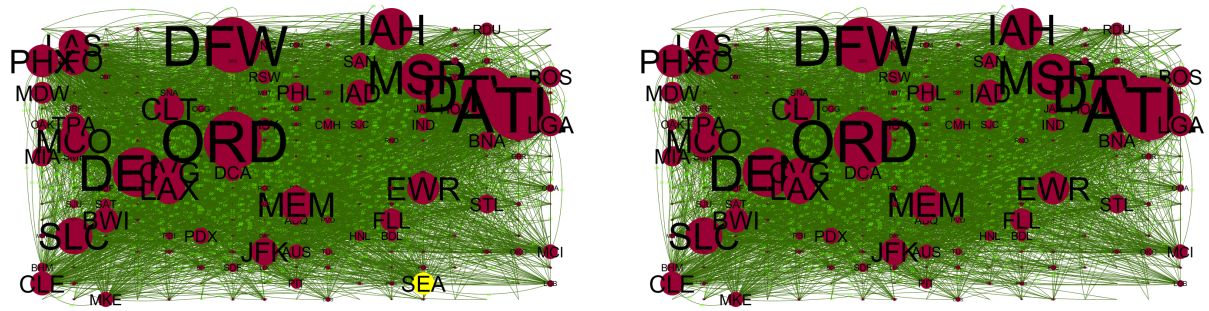
3) CAM: Time Series Clustering: Another approach we explore for clustering is to utilize signal information of airports' delay time series. We extract features using time series transformation methods, namely Discrete Fourier Transform (DFT) [32] and Discrete Wavelet Transform (DWT) [33]. We call these approaches as TSDFT-RAM (time series clustered model using DFT) and TSDWT-RAM (using DWT).

B. Airport Network Analysis

Flights between airports are represented by an airport interaction graph: Each airport corresponds to a node and each flight between two airports corresponds to an edge between the nodes. The edge weight is calculated using total number of flights from the origin airport to the destination (i.e. 1/w, w: is the total number of flights from the origin to destination). The data sets are collected from http://www.transtats.bts.gov/. We use this interaction network to analyze the topological properties of the airports within the global airport network. We visualize the airport network graph of with 305 nodes and 4622 edges using Cytoscape [34]. Every year has its own flight



(a) Left side: "ATL" first neighbors; right side: without "ATL" first neighbors



(c) Left side: "MWH" first neighbors; right side: without "MWH" first neighbors

Fig. 3. Connectivity level and number of delays.

delays, so the airport network edges have different network scores. While doing experiments, we used their own network status for modeling. However, for the graph theoretical analysis later on we presented the scores for only one of the years since aggregating them would not be reasonable.

Delay vs. airport size and connectivity Figure 2 illustrates the top-30 airports which have the highest number of delays. The node sizes are shown according to number of delays (e.g., ATL has the highest number of delays). It is not surprising that the number of delays is correlated by the size of airport and the connectivity level of the nodes, as illustrated in Figure 3. Selected airports are "ATL, "MWH" which have the highest, and the lowest number of delays in the US. Almost all of the airports are connected to "ATL" as a first neighbor, as depicted in Figure 3(a) in yellow and just one airport is connected to "MWH", as depicted in Figure 3(b). We observe that the airports that have a high flight density is prone to have flight delays.

Graph-based features. The topological features include: the *hub score* of the airport, the *betweenness centrality* of the airport, and *articulation point(s)* on the graph. The node score features include *in-degree* and *weighted in-degree* of the airports.

Hub Score is the left-singular vectors of the Singular Value Decomposition (SVD) of the adjacency matrix A of a graph, which is used to represent the relative importance of a node in a network [13]. An airport with a high hub score is the origin of many flights to important and large-scale airports, and is naturally more important than the nodes with low hub score. Airports with similar hub scores may be expected to show similar behavior in terms of their arrival delays. Top 5 normalized hub scores can be seen in Table I.

Betweenness centrality of node v in a directed graph $G = (V, E)$ can be represented as:

$$b(v) = \sum_{s \neq v \neq t \in V} \frac{p_{st}(v)}{p_{st}} \tag{1}$$

where node s to node t represented as p_{st} and the number of shortest paths that pass through node v total number of shortest paths from represented as $p_{st}(v)$. In order to apply idea of betweenness centrality to airport network context, we took the edge values as $1/w$ where w is number flights between corresponded nodes.

Algorithm 1 Find-Articulation-Points

```
dfsnum(v) ← −1, for all v
dfscounter ← 0 r ← |V|
for i ← 1 to r do
    v ← V_i
    if dfsnum(v) ≠ −1 then
        DFS(v)
```

Betweenness centrality (BC) of an airport can quantify its use as a popular transfer node between other airports in the network. An airport with high BC is in the path of many arrival-destination pairs and may denote some relationship for connecting flights. Being a central airport in the network naturally increases the density of the flight traffic. Considering percentage of intersected nodes in Figures 2 and 4 one can say that BC can serve as a potential indicator for delay behavior.

Articulation point of a graph is a node whose removal causes other nodes to be unreachable. Let $G = (V, E)$ be a directed graph, articulation points of graph G can be found

TABLE I
TOP-5 AIRPORTS IN THE CONTEXT OF NODE SCORES

Rank	Airport	Hub score	Airport	Between.
1	Hartsfield-Atlanta I.	1.00	Hartsfield-Atlanta I.	1.00
2	Chicago O'Hare I.	0.883	Dallas-Fort Worth I.	0.777
3	Dallas-Fort Worth I.	0.775	Chicago O'Hare I.	0.654
4	San Francisco I.	0.745	Salt Lake City I.	0.546
5	Denver I.	0.737	Detroit Metropolitan	0.539
Rank	**Airport**	**In-degree**	**Airport**	**W. in-degree**
1	Hartsfield-Atlanta I.	1.000	Hartsfield-Atlanta I.	1.000
2	Chicago O'Hare I.	0.924	Chicago O'Hare I.	0.755
3	Dallas-Fort Worth I.	0.886	Dallas-Fort Worth I.	0.645
4	Detroit Metropolitan	0.810	Denver I.	0.575
5	Denver I.	0.791	Los Angeles I.	0.481

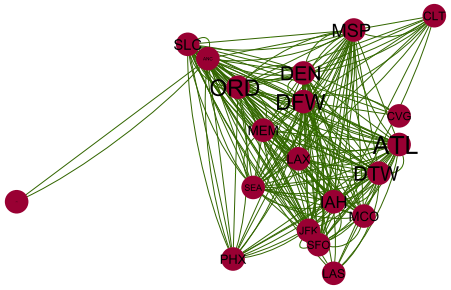


Fig. 4. Top-20 airports with highest betweenness centrality scores.

by Algorithm 1 and Algorithm 2. *dfsnum* is a variable that keeps the information whether node v discovered or not. Also *dfscounter* counts *dfs*s for a specific node. We have identified 19 articulation points in the US airport data set that follows this definition.

Articulation points of an airport network may be expected to have similarly high traffic behavior, causing high number of flight delays. 12 of 19 articulation points of the airport network are indeed among the top 19 airports ordered by the number of delayed flights. The majority of articulation points have high flight delays.

Algorithm 2 DFS(v)

```
dfsnum(v) ← dfscounter
dfscounter ← dfscounter + 1
low(v) ← dfsnum(v)
foreach edge (v, x) do
    if dfsnum(x) == −1 then
        DFS(x)
        low(v) ← min{low(x), low(v)}
        if low(x) ≥ dfsnum(v) then
            v is an art. point
    else if x is not parent of v then
        low(v) ← min{low(v), dfsnum(x)}
```

The number of neighbor airports and the number of flights are naturally related to the traffic and arrival delays. The number of airports that have flights to a node is the in-degree of an airport node. The in-degree of node v in a directed graph $G = (V, E)$ can be calculated as:

$$id(v) = |\{u : (u, v) \in E\}| \tag{2}$$

The weighted in-degree of node v, the number of incoming flights, in a directed graph $G = (V, E)$ is:

$$wid(v) = \sum_{\{u:(u,v)\in E\}} w_{uv} \tag{3}$$

Table I lists the top 5 airports in terms of the presented scores. The top-30 highest delayed airports are illustrated in Figure 2. Figure 5 displays the correlation between the graph-based scores and the number of delayed flights. Red points in the figure represent outlier points in that score set and straight lines are obtained through linear regression on scores except outlier points. Scores presented in Table I and Figure 5 are normalized to the largest value.

C. Time Series of Flight Delays

Arrival delays of an airport can be represented as a time series, a sequence of numerical points in successive order, over the differences between the actual arrival times and the scheduled arrival times. The value of each time point is the maximum (or median) arrival delay, in minutes, of the incoming flights to the airport for the corresponding period. For the purpose of experimentation, we produce the time series for one-year data of length 2920 (8 points for each day) for 305 distinct airports. We do this for seven years that are utilized in experimental section. We then estimate the delays of a period of three-hours in a day. Figure 6 illustrates the delay behavior of time series of the selected 8 big airports for a day. IATA codes of airports are used instead of airports' names in graphs (ATL, SFO, LAX, etc.) Each delay point versus time in the graphs represents the maximum (median) delay occurred in a period of three-hours of a day and delays are measured in minutes. The figure shows the maximum and median based delay behaviors of airports, and illustrates that there are airports with similar delay behaviors to each other.

For the proposed approach we use three types of forecasting models: Multiple regression models, Seasonal Autoregressive Integrated Moving Average (SARIMA) family of models and Regression with ARIMA Errors (REG-ARIMA) models.

Multiple regression model represents a dependent variable y by using k multiple independent variables $x_1, x_2, \ldots, x_k$ as

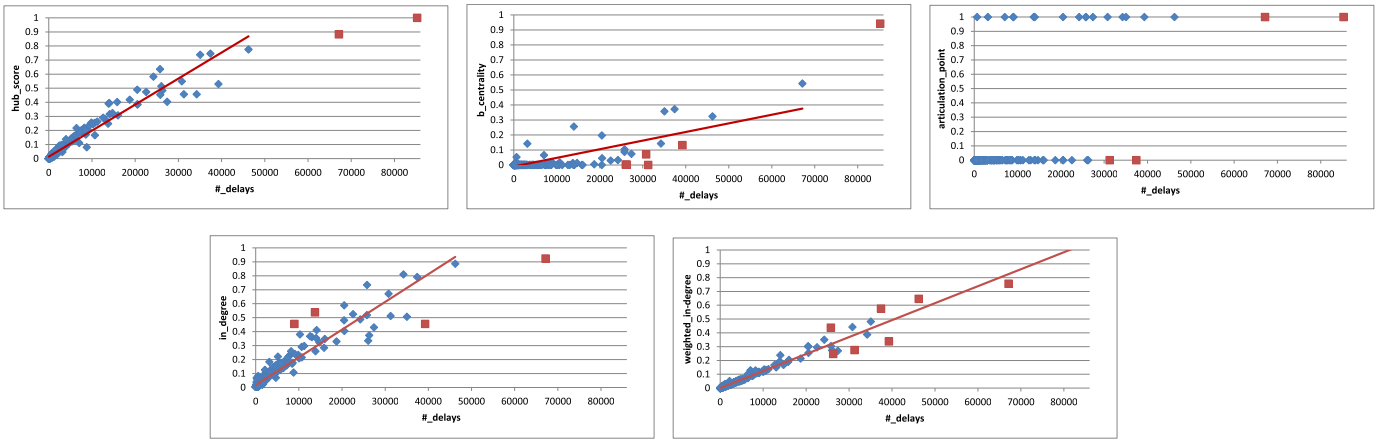


Fig. 5. Graph-based scores vs. number of delays.

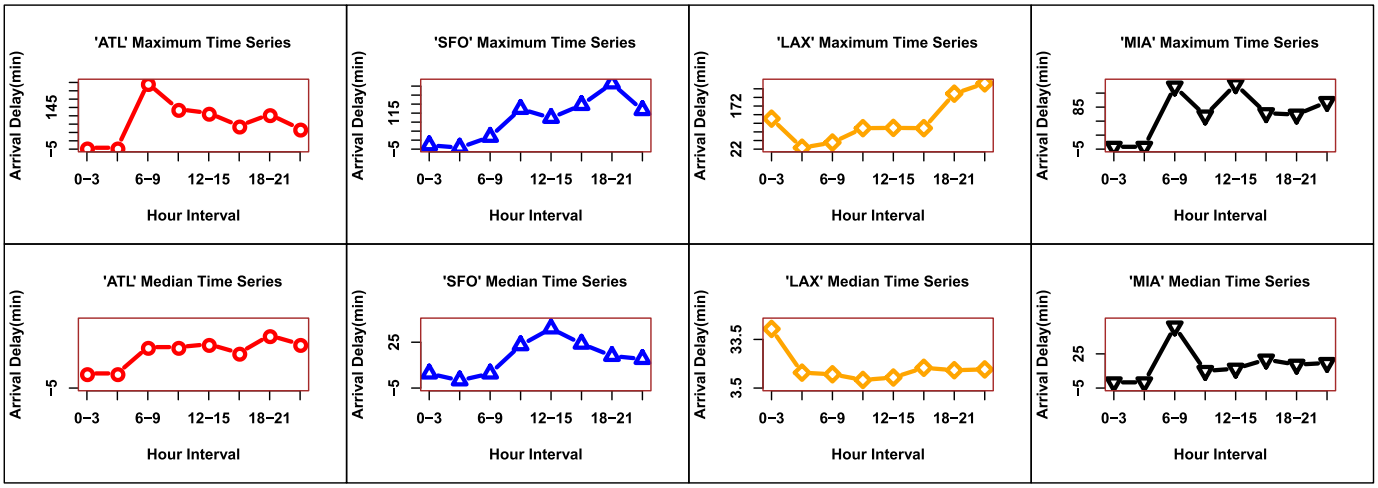


Fig. 6. Example time series of maximum and median arrival delays for a day.

in the form of Equation 4. Building a regression model is the problem of finding the model's coefficient set $b_1, b_2, \ldots, b_k$.

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_k * x_k \quad (4)$$

SARIMA model represents a time point of a time series as the linear combination of its past time points. A $SARIMA$ model $SARIMA(p, d, q)(P, D, Q)$ is represented by its autoregressive order p, differencing order d, moving average order q, seasonal autoregressive order P, seasonal differencing order D and seasonal differencing order Q. Building a $SARIMA$ model on given data series aims to determine the order of model and a vector of parameters. Further discussions on SARIMA model operators, stationarity of time series, and how to estimate the parameters of a SARIMA model can be found in [11].

REG-ARIMA model is a combination of a regression and an Autoregressive Integrated Moving Average (ARIMA) model. To build a REG-ARIMA model on time series X, one builds a regression model on X where residual time-series N of the regression model follows a SARIMA or an ARIMA model. The first part of the REG-ARIMA model is formulated as in Equation 5 and N is the remaining time series on which a SARIMA model will be built.

$$X = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_k * x_k + N \quad (5)$$

IV. Experimental Evaluation

We evaluate how the proposed network/clustering based methodology change the accuracy of forecasting the flight delays, compared to modeling each airport individually. We compare the different variants of incorporating the network/clustering information with the baseline of fitting an individual model for each time series of delays.

We used the data set provided by RITA (Research and Innovative Technology Administration), absorbed into OST-R, that contains 7 years of flight records in the United States for the years 2006 to 2012. The data include attributes such as origin, destination, arrival time, scheduled arrival time, etc. RITA coordinates the U.S. Department of Transportation research programs.

We constructed the network of the 305 airports in the data set, and generated the flight arrival delay time series of each airport. Note that the arrival delay is defined as the difference between the scheduled arrival time and the actual arrival time, both in local time. The forecasting methods are implemented to predict the results for three-hour periods. A delay time-series of length 2920 for each year is used for each airport. We took the first 2680 time points to build the models and made the forecasts for the remaining 240 points.

TABLE II
CORRELATION COEFFICIENTS BETWEEN FEATURES

	HScore	Betw.	APoint	InDegree	WInDegree
HScore	1.000	0.695	0.618	0.953	0.969
Betw.	0.695	1.000	0.601	0.681	0.822
APoint	0.618	0.601	1.000	0.683	0.663
InDegree	0.953	0.681	0.683	1.000	0.948
WInDegree	0.969	0.822	0.663	0.948	1.000

We present the accuracy results for every week (4 weeks for 240 points) and compare the accuracy performances using the measures of Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) in Equation 6, and Equation 7 respectively.

$$MAPE = \frac{1}{h}\left(\Sigma_{i=1}^{h}\left|\frac{x_{n+i} - f_i}{x_{n+i}}\right|\right) \quad (6)$$

$$MAE = \frac{1}{h}\left(\Sigma_{i=1}^{h}|x_{n+i} - f_i|\right) \quad (7)$$

where h is the forecasting period, x_{n+i} is the i-th future time point, and f_i is the i-th forecast.

A. Approaches in Comparison

The baseline method, ISM (Individual SARIMA model) fits an individual model to each time series of delays. We refer to our methods that follow different variants of graph partitioning, clustering, and time-series analysis within the proposed methodology as follows: Graph theoretic clustered SARIMA modeling (GTC-SM), graph theoretic clustered REG-ARIMA modeling (GTC-RAM), graph partitioned SARIMA modeling (GP-SM), graph partitioned REG-ARIMA modeling (GP-RAM), time series clustered model using DFT (TSDFT-RAM), time series clustered model using DWT (TSDWT-RAM) and individual REG-ARIMA modeling (RAM).

GTC-SM and GTC-RAM first cluster the airports using the graph-based features and use common SARIMA and REG-ARIMA models respectively for the clusters. RAM builds a REG-ARIMA model for each time series. GP-SM and GP-RAM partition the airport network using graph partitioning, and fit a common SARIMA and REG-ARIMA models respectively to the representative time series of each partition. TSDFT-RAM uses DFT to extract the time series features to obtain the airport clusters and fit a common REG-ARIMA model for each cluster. TSDWT-RAM uses DWT for extracting the features.

Although we did experiments for all approaches stated above, to simplify presenting results we illustrated only the results of approaches (GTC-RAM, TSDFT-RAM, TSDWT-RAM) that perform consistently better than the baseline.

B. Validation of Approaches Using RAM

Table II shows the correlation of the pairs of graph-based features. Hub score, in-degree, and weighted-in-degree are highly correlated with each other, while betweenness centrality and articulation have less correlation with them and with each other.

TABLE III
REGRESSION MODELS' SUMMARIES

Model	P-value	Adjusted R-squared.
1	2.20E-16	0.224
2	2.20E-16	0.250
3	2.20E-16	0.064
4	2.20E-16	0.107
5	2.20E-16	0.146
6	2.20E-16	0.150
7	2.20E-16	0.100
8	2.20E-16	0.300
9	2.20E-16	0.158
10	2.20E-16	0.224
11	4.43E-10	0.030
12	2.20E-16	0.081
13	0.014	0.004

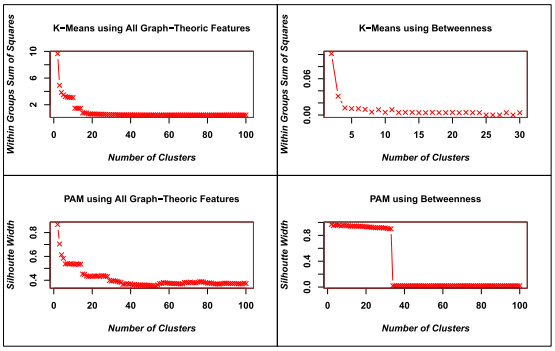


Fig. 7. Cluster quality behavior changing according to number of clusters.

Table III illustrates the model summaries for the case of the number of clusters around 50. The summaries of the clusters whose regression models are based only on the intercept, and the clusters with size of 1, are not presented in the table. Model number defines 15 out of 50 models. The p-values for all the models are found to be less than 0.05, i.e., all regression models are statistically significant.

C. The Number of Clusters

We use k-means [35] and Partitioning Around Medians (PAM) [36] in our experiments. To determine the number of clusters, we utilize the within-cluster sum of squares and silhouette width as quality measures for k-means and PAM, respectively. The plots for the number of clusters vs. the cluster quality are presented in Figure 7. We can see an elbow behavior on all plots which can be used in determining the number of clusters. These represent the qualities of the graph-theoretic clustering. The same procedure is applied for the time series clustering in our experiments. We continue with k-means in experimental evaluation.

D. Accuracy Evaluation

We compare the quality of forecasts of the proposed approaches with those of the baseline ISM. We examine both the performance improvement via different grouping methods (graph-theoretic clustering, graph partitioning, time series clustering) and via different time series modeling approaches (SARIMA modeling, REG-ARIMA modeling). We test the performance on maximum and median time series of seven

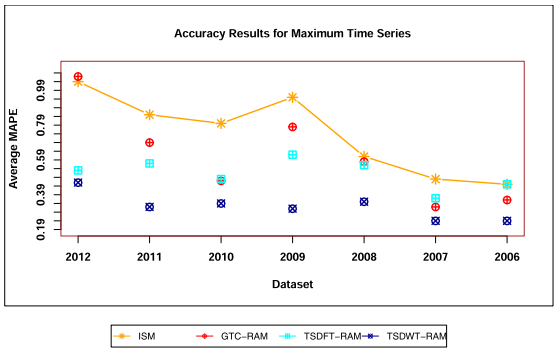


Fig. 8. Accuracy comparison of proposed approaches using all features for maximum time series.

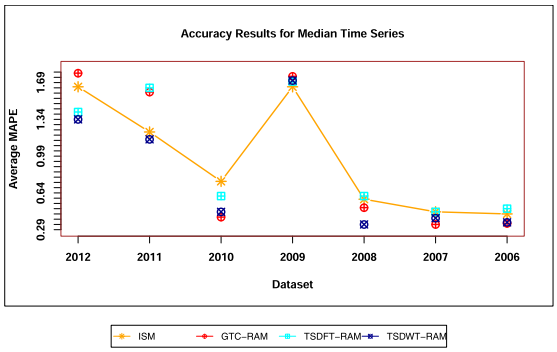


Fig. 9. Accuracy comparison of proposed approaches using all features for median time series.

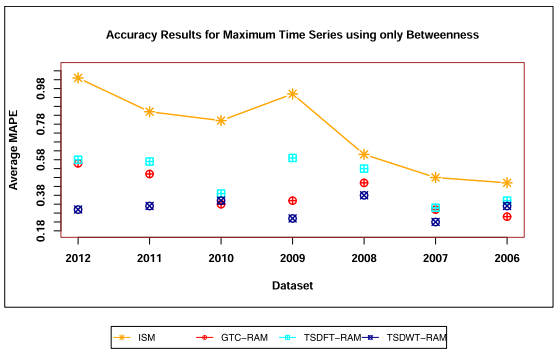


Fig. 10. Effect of using only betweenness centrality feature on accuracy for maximum time series.

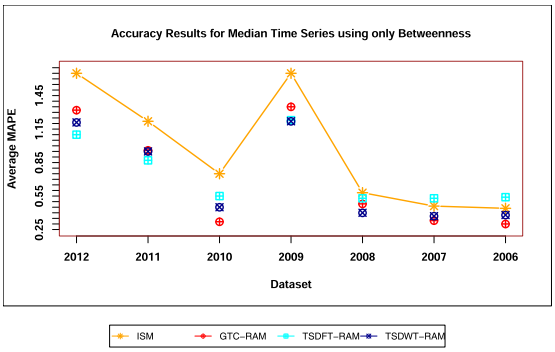


Fig. 11. Effect of using only betweenness centrality feature on accuracy for median time series.

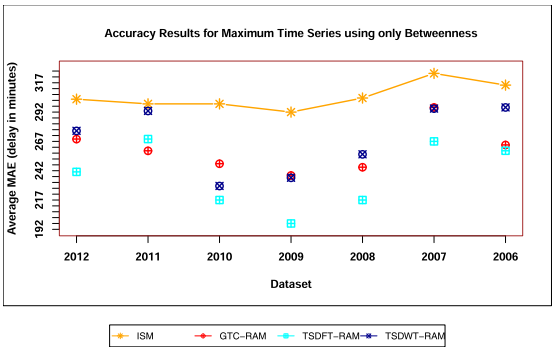


Fig. 12. Performance of methods for maximum time series measured by MAE.

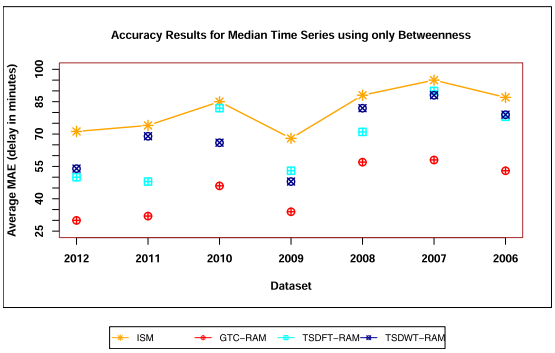


Fig. 13. Performance of methods for median time series measured by MAE.

different years' data sets. Maximum time series are composed of maximum delays of each 3-hour slots of days and median time series are created by using median delay values of each 3-hour slots of days. Individual modeling and prediction are done on local times, combining is done according to UTC time. We note that the graph features are utilized both to cluster airports and used as regression variables in REG-ARIMA.

1) Forecasting Models Using All Graph-Theoretic Features: Figure 8 and 9 show the MAPE results where all graph-theoretic features are included in both GTC stage and REG-ARIMA. The yellow star-shaped line represents the baseline ISM. More successful models compared to ISM exist below this yellow star-shaped line.

Experimental results show that the proposed approaches have significant improvements compared to the baseline model ISM. In particular, GTC-RAM, TSDFT-RAM and TSDWT-RAM result in outstanding improvements over ISM. We can summarize improvements of the proposed approaches as follows.

a) On maximum time series: TSDWT-RAM shows an average of 55% improvement in terms of average MAPE of forecasts over the baseline. The improvements range from 43% to 62%, for the years of 2008 and 2011, respectively. GTC-RAM shows from 5% to 41% improvements, for 2008 and 2010. On the yearly average, GTC-RAM makes a 25% improvement. TSDFT-RAM provides an average of 18% improvement over the baseline in terms of the forecast accuracy.

b) On median time series: TSDWT-RAM shows an average of 18% improvement in terms of average MAPE of forecasts over the baseline. The improvements range from 13% to 45%, for the years of 2007 and 2008, respectively. GTC-RAM shows from 14% to 45% improvements, for 2008 and 2010. On the yearly average, GTC-RAM makes a 27% improvement.

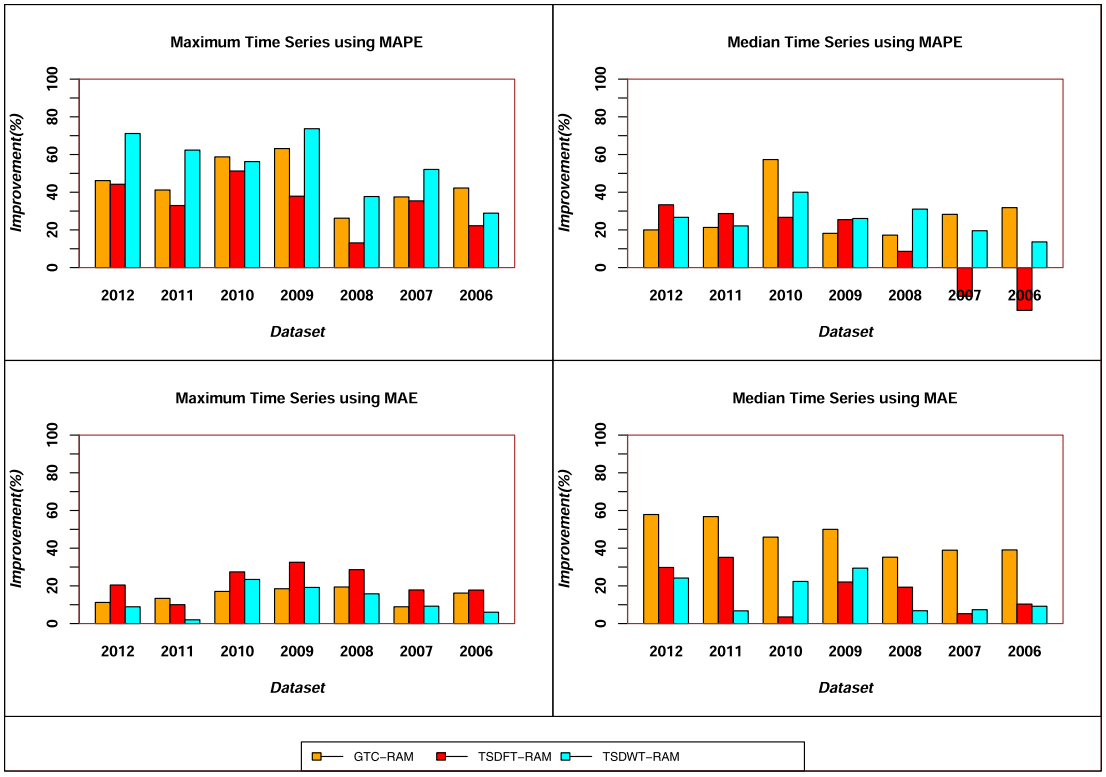


Fig. 14. Accuracy improvements using only betweenness.

Clustering makes the forecasting models more robust to the outliers in the time series. A further improvement is achieved by using a REG-ARIMA model where the graph-based features are used as the regressor variables. For many airports, ISM has a high MAPE that is significantly more than 1. The clustered model reduces the MAPE to values significantly smaller than 1. We have also checked the specific cases where the individual model performs better than the clustered model for an airport. According to clustering, some of the big airports may not belong to any cluster (e.g. ORD) or some small airports (e.g. VLD, CLT) may belong to a cluster on which prediction model performs worse compared to individual modeling. In all of these cases, the MAPE is significantly less than 1 for both types of models.

2) Identifying Important Features for Forecasting: We repeat the experiments using a subset of the features, as opposed to using all. This helps us understand which features are the most important for accuracy improvements. We find out that "betweenness centrality" gives the best accuracy result among all cases on this setup. Note that betweenness centrality of a node in the graph measures degree of being the center for shortest paths. A node with higher betweenness centrality may correspond to a transfer center or a hub in the airport network.

Results of the accuracy improvements when the feature subset containing only "betweenness centrality" is used on maximum and median time series are presented in Figure 10 and 11. Accuracy improvements are illustrated in Figure 14 and summarized below.

a) On maximum time series: On yearly average, GTC-RAM makes 45% improvement in terms of average MAPE of forecasts over the baseline. The improvements range from 26% to 63%, for the years of 2008 and 2009, respectively. TSDFT-RAM shows from 13% to 51% improvements, for 2008 and 2010. On the yearly average, TSDFT-RAM shows a 33% improvement. TSDWT-RAM provides same level accuracy when only betweenness centrality topological feature is used compared to case where all topological features are used. MAPE of this model ranges from 28% to 73%, and yearly average is 54%.

b) On median time series: On yearly average, GTC-RAM shows a 28% improvement in terms of average MAPE of forecasts over the baseline. The improvements range from 17% to 57%, for the years of 2008 and 2010, respectively. TSDFT-RAM does not have improvement for years 2006 and 2007, so its yearly average keeping out these years is 25%. Yearly average of TSDWT-RAM is also 25%.

We also evaluated the methods using MAE measure. We present MAE results only for the three top performing methods using MAPE. Accuracy results of these methods measured by MAE are shown in Figures 12 and Figure 13 for the maximum and median time series respectively. The summary of improvements are illustrated in Figure 14 The performance behavior of the methods is similar with the evaluation by the MAPE.

Betweenness Centrality (BC) score of an airport is found to be a factor in understanding the delays associated with the airport. The BC does not always have a high correlation with the number of flights. The airports that are central in the paths of potential travel itineraries are vulnerable to further delay. Similarly, most of the *articulation points* of the airport network are found to be among the highest delayed airports.

BC and articulation have less correlation with other measures such as the hub score, and with each other. Several airports have highly similar graph based features in the airport network. For example, ATL and ORD are consistently in the same clusters based on the graph centrality measures. This may help to gather more information about their delay patterns using additional data from each other.

V. Conclusion

While airport networks contain rich information, they have not been explored enough for some essential tasks in air transportation, such as forecasting flight delays. In this paper, we incorporated airport network information and utilized graph based scores, such as betweenness centrality (BC) and articulation points in forecasting of arrival delays. The position of the airports in the network and the airports' delay time-series similarities are investigated as potential parameters to augment the models for forecasting flight delays.

We introduced the Clustered Airport Modeling (CAM) that uses a REG-ARIMA model enhanced with the results of clustering. The CAM approach includes grouping and modeling steps that make use of the airport network. The network is used for both graph-based clustering of airports and as an exploratory variable for the prediction model. Our experiments show that CAM provides more accurate results than a baseline (SARIMA) model applied individually for each airport. BC score is found to be an effective regressor in the clustered REG-ARIMA. When we compare ISM and CAM with LSTM (Long-Short Term Memory), which is usually considered as the state of the art in forecasting, CAM is found to be as good as LSTM, which are both more successful than ISM. This observation suggests that using network structure in similar forecasting problems is a promising direction to pursue.

To the best of our knowledge, this is among the first to utilize the airport network for forecasting flight delays. Our work may inspire other types of analysis based on air transportation networks. The trajectory of the delays can be analyzed by differentiating the airports that cause the delay propagation and those that are the victims of the propagation. This line of work can help policy makers to analyze airport networks and improve traffic flow management.

Acknowledgment

The authors would like to thank the anonymous reviewers and the editor for their valuable inputs.

References

[1] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A review on flight delay prediction," 2017, *arXiv:1703.06118*. [Online]. Available: http://arxiv.org/abs/1703.06118

[2] V. Martinez, "Flight delay prediction," M.S. thesis, Dept. Comput. Sci., ETH Zürich, Zürich, Switzerland, 2012. [Online]. Available: https://www.research-collection.ethz.ch/handle/20.500.11850/153312

[3] N. G. Rupp, "Further investigations into the causes of flight delays," Dept. Econ., East Carolina Univ., Tech. Rep., 2007.

[4] C. Barnhart, D. Fearing, and V. Vaze, "Modeling passenger travel and delays in the national air transportation system," *Oper. Res.*, vol. 62, no. 3, pp. 580–601, Jun. 2014.

[5] N. G. Rupp and G. M. Holmes, "An investigation into the determinants of flight cancellations," *Economica*, vol. 73, no. 292, pp. 749–783, Nov. 2006.

[6] R. Ghodsi, M. Zakerinia, and M. Jokar, "Neural network and fuzzy regression model for forecasting short term price in ontario electricity market," in *Proc. 41st Int. Conf. Comput. Ind. Eng.*, Accessed: Apr. 20, 2016.

[7] F. Van den Bossche, G. Wets, and T. Brijs, "A regression model with arima errors to investigate the frequency and severity of road traffic accidents," in *Proc. Electron. 83rd Annu. Meeting Transp. Res. Board*, Washington, DC, USA, 2004.

[8] C. Dritsaki, "Forecast of SARIMA models: An application to unemployment rates of Greece," *Amer. J. Appl. Math. Statist.*, vol. 4, no. 5, pp. 136–148, Jan. 2016.

[9] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[10] M. Kumar and N. R. Patel, "Using clustering to improve sales forecasts in retail merchandising," *Ann. Oper. Res.*, vol. 174, no. 1, pp. 33–46, Feb. 2010.

[11] İ. Gür, M. Güvercin, and H. Ferhatosmanoglu, "Scaling forecasting algorithms using clustered modeling," *Int. J. Very Large Data Bases*, vol. 24, no. 1, pp. 51–65, Feb. 2015.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[14] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proc. ACM SIGKDD Workshop Mining Data Semantics (MDS)*, New York, NY, USA, 2012, pp. 3:1–3:8. [Online]. Available: http://doi.acm.org/10.1145/2350190.2350193

[15] B. Hoppe and C. Reinelt, "Social network analysis and the evaluation of leadership networks," *Leadership Quart.*, vol. 21, no. 4, pp. 600–619, Aug. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1048984310000901

[16] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Netw.*, vol. 1, no. 3, pp. 215–239, 1979.

[17] D. R. White and S. P. Borgatti, "Betweenness centrality measures for directed graphs," *Social Netw.*, vol. 16, no. 4, pp. 335–346, Oct. 1994.

[18] F. Fouss, M. Saerens, and J.-M. Renders, "Links between Kleinberg's hubs and authorities, correspondence analysis, and Markov chains," in *Proc. 3rd IEEE Int. Conf. Data Mining*, 2003, pp. 521–524. [Online]. Available: http://dblp.uni-trier.de/db/conf/icdm/icdm2003.html#FoussSR03

[19] M. Zanin and F. Lillo, "Modelling the air transport with complex networks: A short review," *Eur. Phys. J. Special Topics*, vol. 215, no. 1, pp. 5–21, Jan. 2013.

[20] G. Santos and M. Robin, "Determinants of delays at European airports," *Transp. Res. B, Methodol.*, vol. 44, no. 3, pp. 392–403, Mar. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191261509001313

[21] A. Kim and M. Hansen, "Deconstructing delay: A non-parametric approach to analyzing delay changes in single server queuing systems," *Transp. Res. B, Methodol.*, vol. 58, pp. 119–133, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191261513001744

[22] N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transp. Res. C, Emerg. Technol.*, vol. 27, pp. 60–75, Feb. 2013.

[23] Y. Cheng, "Solving push-out conflicts in apron taxiways of airports by a network-based simulation," *Comput. Ind. Eng.*, vol. 34, no. 2, pp. 351–369, Apr. 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835297002829

[24] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transp. Res. C, Emerg. Technol.*, vol. 44, pp. 231–241, Jul. 2014.

[25] *FAA Strategic Plan, FY 2019-2022*. Accessed: Jun. 2019. [Online]. Available: https://www.faa.gov/about/plans_reports/media/FAA_Strategic_Plan_Final_FY2019-2022.pdf

[26] *Federal Aviation Administration 2016 National Aviation Research Plan (NARP), Report of the FAA to the US States Congress*. Accessed: Jun. 2019. [Online]. Available: http://www.faa.gov/go/narp

[27] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 3, p. 36104, Sep. 2006.

[28] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.
[29] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006.
[30] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," 2006, *arXiv:cond-mat/0603718*. [Online]. Available: https://arxiv.org/abs/cond-mat/0603718
[31] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, Dec. 2004, Art. no. 066111. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0408187
[32] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," *ACM SIGMOD Rec.*, vol. 23, no. 2, pp. 419–429, 1994.
[33] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in *Proc. 15th Int. Conf. Data Eng.*, 1999, pp. 126–133.
[34] P. Shannon, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.
[35] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.
[36] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith, "Clustering rules: A comparison of partitioning and hierarchical clustering algorithms," *J. Math. Model. Algorithms*, vol. 5, no. 4, pp. 475–504, Dec. 2006.
[37] E. Conti, S. Cao, and A. J. Thomas, "Disruptions in the U.S. airport network," 2013, *arXiv:1301.2223*. [Online]. Available: http://arxiv.org/abs/1301.2223
[38] C. H. Q. Ding, H. Zha, X. He, P. Husbands, and H. D. Simon, "Link analysis: Hubs and authorities on the world wide Web," *SIAM Rev.*, vol. 46, no. 2, pp. 256–268, Jan. 2004.
[39] S. Ho and M. Xie, "The use of ARIMA models for reliability forecasting and analysis," *Comput. Ind. Eng.*, vol. 35, nos. 1–2, pp. 213–216, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835298000667
[40] E. W. Weisstein. (Dec. 15, 2017). *Graph Eccentricity*. [Online]. Available: http://mathworld.wolfram.com/GraphEccentricity.html

Mehmet Güvercin graduated from Computer Engineering, Istanbul Technical University. He received the M.S. degree in computer science from Bilkent University, where he is currently pursuing the Ph.D. degree. His research is on forecasting and scalable forecasting related to networks.

Nilgun Ferhatosmanoglu received the B.S. degree in industrial engineering from Bilkent University, and the Ph.D. degree in industrial and systems engineering from Ohio State University. Her current research interests are in the areas of statistical modelling, data analytics, and optimization. She received a TUBITAK Career Award for her project on network-based airline data analytics and forecasting.

Bugra Gedik received the Ph.D. degree in computer science from the Georgia Institute of Technology. He is currently an Associate Professor with the Department of Computer Engineering, Bilkent University, Turkey. His research interest is in data-intensive distributed systems.