# US Domestic Airline Network Temporal Analysis

**Alan Wang**[a,b,1]**, Brenden Cheung**[a,1]**, and Kun Xu**[a,1]

[a]Department of Mathematics, University of California, Los Angeles, CA 90095; and [b]Department of Statistics, University of California, Los Angeles, CA 90095

This manuscript was compiled on December 20, 2022

**Air is one of the most common types of domestic transportation in the United States. Due to the colossal amount of people connected through these transportation methods, it forms a vast transportation network. In this paper, we use complex network concepts and methods to analyze the development of the US Domestic Airline Network from 2009 to 2018. Methods and models employed in this paper include Centrality, Assortativity, Small-World Properties, Scale-Free properties, and Targeted Attack Tolerance. This study demonstrates the consolidation of the Hub-and-Spoke system of the industry and particularly focuses on the rise and fall of the small airports in the system.**

networks | airline | targeted attack

The airline industry plays a crucial role in connecting people and places around the world. In recent years, the growth of the global airline network has been staggering, with the number of annual passengers on all U.S. scheduled airline flights increasing by nearly 250 million between 2009 and 2018 (1). Despite the significant economic and social benefits of the airline industry, it is also a complex and dynamic system that is susceptible to a wide range of external factors such as economic downturns, natural disasters, and global pandemics.

In the paper "A study of the U.S. domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016" (2), Siozos-Rousoulis et al. conducted an analysis on the U.S. air transportation network to assess the temporal evolution of this air transport network's robustness. Through finding a number of network properties, they found trends that corresponded to the 9/11 attacks and restructuring of the flight network shortly afterwards.

In this paper, which was inspired by the paper written by Siozos-Rousoulis et al., we perform a similar analysis on the more recent development of U.S. airports, specifically looking at the years from 2009 to 2018. Using U.S. domestic flight data, we construct temporal networks with nodes corresponding to the airports and directed edges corresponding to the flights for each year. These edges are weighted by the number of flights in that year. From these air transportation networks, we investigate the networks' centrality measures, associativity, scale-free properties, resilience, and targeted attack tolerance to observe patterns and trends that have emerged over the past decade.

## Background

**Dataset.** We get the dataset "Airline Delay and Cancellation Data, 2009 - 2018, Flight info. of US domestic flights" from Kaggle (3). The dataset contains information on all domestic flights from 2009 to 2018, including origin, destination, airline, flight data and time, and delay information.

**Data Preprocessing.** The original dataset has a size of 800MB per year, and with 10 years in total, an 8GB dataset would be beyond our computation capacity. Therefore, before conducting any analysis, we clean the dataset. As the main focus of this project is to study the temporal structural changes in the domestic airline system, we only need information on ORIGIN, DESTINATION, AIRLINE, and DELAY. We use SQL to group this information together and calculate the number of flights from city A to city B by an airline, along with their average delay. In the end, for each year, we create a directed network using NetworkX (4), pointing from origins to destinations, with two edge attributes: number of flights and average delay.

After cleaning the dataset, the data size is reduced to 10MB per year. This is a significant reduction in size, which allows us to easily analyze the data using our computation resources.

## Methods and Models

**Centrality.** This study will use two indicators to measure the relative importance of different airports: node degree and betweenness centrality.

According to Newman, node degree (k) is the number of nodes that a node is connected to. As the node degree of node $i$ increases by one, node $i$ has one more connection with other nodes, which means node $i$ has greater importance. $k_i$ can be written as

$$k_i = \sum_{i=1}^{N} A_{ij}, \qquad [1]$$

where $k_i$ denotes the degree of node i and $A_{ij}$ is the adjacency matrix of the network (5).

Between centrality measures the relative importance of an airport to the aviation system. According to Siozos, the betweenness centrality (B) of node i is defined as

### Significance Statement

Due to the U.S. air transportation network's criticality to the mobility and functioning of local economies, it is imperative to assess its topological evolution and ensure that it is well-connected, efficient, robust, and secure. We have not only conducted empirical studies of the evolution of the network structure but also studied the system's resilience to targeted attacks. Our methodologies can be extended to different transportation networks to provide a general perspective of a system's vulnerabilities.

A.W., B.C., and K.X. designed research, performed research, analyzed data, and wrote the paper; A.W. and B.C. wrote the simulation code.

The authors declare no conflict of interest.

[1] A.W., B.C., and K.X. contributed equally to this work.

$$B_i = \frac{1}{(N-1)(N-2)} \sum_{i \neq j \neq k} \frac{g_{jk}(i)}{g_{jk}}, \qquad [2]$$

where $g_{jk}$ is the number of binary shortest paths between two nodes and $g_{jk}(i)$ is the number of binary shortest paths passing through node $i$ (2).

**Assortativity.** Assortativity measures the preference for a network's nodes to attach to others that are similar in some way. In the context of the airline network, assortativity measures how strongly airports with similar flight numbers are connected, where a high assortativity refers to that more flights fly from larger airports to other larger airports, shown in a **Hub-and-Spoke** system. On the other hand, a low assortativity shows a **Point-to-Point** system.

Point-to-point systems involve airlines operating flights directly between two destinations without any intermediate stops. This can be more efficient for passengers, as they can travel directly to their destination without having to change planes or transfer through a hub. It can also be more efficient for airlines, as they can operate smaller planes on fewer routes and avoid the costs and complexities of managing a hub. However, point-to-point systems can also be more expensive for airlines, as they may need to operate more flights to serve the same number of destinations, and they may also have to deal with higher operating costs and lower load factors on individual flights.

Hub-and-spoke systems, on the other hand, involve airlines operating flights through a central hub, where passengers can transfer between flights to different destinations. This can be more efficient for airlines, as they can operate larger planes on fewer routes and take advantage of the economies of scale and scope offered by a hub. It can also be more convenient for passengers, as they can connect to a wider range of destinations through a single hub. However, they have to transfer through a hub, deal with the inconvenience of connecting flights, and endure longer travel times. Additionally, hub-and-spoke systems can be more vulnerable to disruptions, as a delay can be easily propagated to an exponentially-large number of flights.

To quantify the change in assortativity and the development of the airline system over years, we use the idea of the Gini coefficient: the Gini coefficient is a term used to describe the level of inequality in a society. It is a commonly used measure of inequality, and it is calculated by taking the difference between the actual distribution of wealth within a society and a perfectly equal distribution and dividing it by the maximum possible difference under a perfectly equal distribution. The Gini coefficient ranges from 0 to 1, with 0 representing perfect equality and 1 representing perfect inequality.

We adopt the idea here: given the airline network context, we consider each airport as an individual in the society and the total number of outbound flights per year as its wealth, and we define the Gini coefficient ($G$) as

$$G = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |k_i - k_j|}{2MN} \in [0,1], \qquad [3]$$

where $k_i$ = weighted out-degree of node $i$, $M$ = total out-weights = $\sum_{i=1}^{N} k_i$, and $N$ = total number of nodes.

**Small-World Properties.** In the context of airline networks, small-world properties refer to the idea that any two airports in the world can be connected by a small number of flights. This is an important property of airline networks, as it allows for efficient and rapid transportation between distant locations. Such a study can provide valuable insights into the functioning and dynamics of the complex airline system.

One advantage of airline networks with small-world properties is that they allow for efficient and rapid transportation between distant locations. This is because, as mentioned above, any two airports in the world can be connected by a small number of flights. This allows passengers to reach their destinations quickly and easily, even if they are located on opposite sides of the world. Additionally, small world properties can also help airlines to better plan and manage their operations, as they can use this property to identify optimal routes and reduce the time and cost of travel.

On the other hand, one potential disadvantage of airline networks with small-world properties is that they may be more vulnerable to disruptions. Because these networks are characterized by a small number of flights connecting different airports, a disruption at a single airport or on a single route can have widespread effects on the entire network. This can lead to delays and cancellations, as well as increased costs for airlines and passengers. Additionally, small-world properties can also make it more difficult for airlines to respond to unexpected events, as they may not have alternative routes available to them.

To quantify the small-world properties, we define two metrics. The first one is the global clustering coefficient $C$ (5), which is the mean of the clustering coefficient per node:

$$C_i = \frac{2m_i}{k_i\,(k_i - 1)}, \qquad [4]$$

where $m_i$ = the number of edges that connect the neighbors of node $i$. Hence,

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \qquad [5]$$

The global clustering coefficient is the measure of the probability of the existence of triangles among the connected trios, and given the airline context, it is the probability that two airports are directly connected by flights, without needing passengers to transfer at a hub. A network with a high global clustering coefficient would have many airports that are connected to each other by direct flights, while a network with a low global clustering coefficient would have fewer direct flights and more "hub-and-spoke" connections through a smaller number of larger airports.

Secondly, we introduce the characteristic path length $L$:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} d_{ij}, \qquad [6]$$

where $d_{ij}$ is the shortest path length from airport $i$ to airport $j$.

The characteristic path length is calculated by taking the average of the shortest path lengths between all pairs of airports in the network: a low characteristic path length indicates that the network is well connected, with many direct flights between airports, making it easy for passengers to travel to their destinations with minimal stops or transfers. A high characteristic path length indicates that the network is less connected, with fewer direct flights and more transfers required to travel between airports.

**Scale-Free Properties.** A scale-free network is one in which the degree distribution follows a power law, with a few nodes having a disproportionately high number of connections compared to the rest of the network. It is defined as

$$f(x) \propto x^{-\alpha}, \qquad [7]$$

where $x$ is the degree and $f(x)$ is the fraction of nodes with a degree greater than or equal to $x$.

In the context of our airport network, finding the network to have scale-free properties would be significant because it would suggest that a small number of airports serve as hubs for a large number of flights. This could have important implications for the efficiency and resilience of the network. For example, if the network has a scale-free structure, it may be more efficient because many flights can be routed through the hub airports, reducing the need for flights to make long detours. However, it could also make the network more vulnerable, because if one of the hub airports experiences disruption, it could have a cascading effect on the rest of the network.

**Targeted Attack Tolerance.** Since 9/11, many studies on airline networks have focused on targeted attack tolerance. This refers to the network's ability to withstand targeted attacks on specific components, such as airports or individual flights, without losing its overall connectivity or functionality. We can also extend this concept of "attack" to include bad weather conditions, accidents, strikes, and natural disasters. This is an important consideration for the security and resilience of airline networks, as they are critical infrastructure and can be vulnerable to attacks from various sources.

To quantify the network's tolerance, we define two metrics: resilience (6) $R$ and the proportion of nodes in the giant connected component as we remove highly centralized nodes, which we define as $S$.

Resilience is defined as:

$$R = \frac{1}{N^2} \sum_{i \neq j}^{N} m(i,j), \qquad [8]$$

where $m(i,j)$ is the minimum number of nodes that need to be removed from the network such that $j$ is unreachable from $i$.

We assume that flights can only divert to airports that were originally reachable from their destination due to air traffic reasons. This metric quantifies how many airports need to be removed for the entire airline system to "shut down". A higher resilience allows the airline system to handle attacks more easily by redirecting.

The proportion of nodes in the giant connected component is defined as (5):

$$S = \frac{size(GCC)}{N}, \qquad [9]$$

where size(GCC) is the number of nodes in the giant connected component.

To simulate targeted attacks, we remove some number of nodes with the highest betweenness centrality along with all the edges that are associated with them and then recalculate S. This procedure is essentially shutting down the largest hub airports in the network and observing how the rest of our network holds despite removing these hub airports.

## Results

**Centrality.** After the 2008 Financial Crisis, the U.S. Air Transport Network witnessed a rapid increase in the number of airports. According to Fig. 1, from 2009 to 2014, especially from 2011 to 2014, more than 30 airports were constructed, implying a surge of 10 percentage points. However, in the meanwhile, the number of flight routes remained relatively stable at the level of 4600 routes.

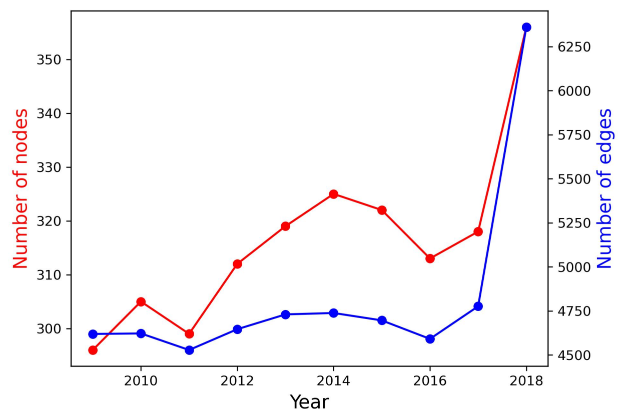


**Fig. 1.** The change in the number of nodes and edges from 2009 to 2018

The mean degree is the average of node degrees of all nodes in the network. As shown in Fig. 2, since the number of edges remains flat, the mean degree continues to decrease from over 31 in 2009 to slightly over 29 in 2014 as the number of nodes increases. Based on the analysis of the data, one possible explanation is that many relatively small regional airports are constructed after the financial crisis. Compared to existing airports, these new airports have fewer connections with other airports.

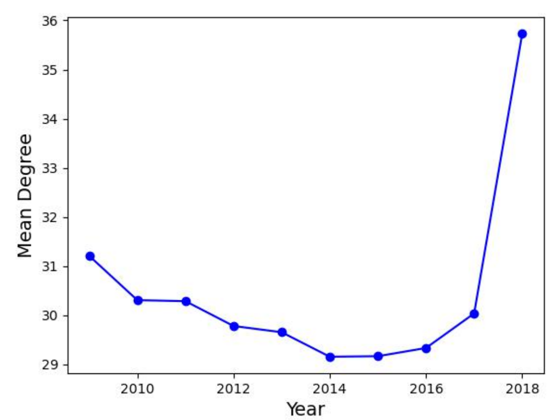


**Fig. 2.** The change in the mean degree from 2009 to 2018

However, from 2014 to 2016, there was a decrease of around 15 airports, and the mean degree began to rise up again. One possible explanation is that these smaller regional airports are not economical compared to larger ones. As a result of that, after five years of experiments, less efficient regional airports are shut down.

Betweenness centrality is a measure of the importance of a node in a network. The average betweenness centrality of a network is the average of the betweenness centrality of all nodes in the network. It measures the degree of centralization in any given network. As the number of nodes with a high betweenness centrality score increases, the average betweenness centrality will increase, implying that the network is becoming more centralized.

However, as smaller regional airports generally have lower betweenness centrality, when these airports are constructed, the average betweenness centrality will decrease, and the entire aviation network will become more decentralized. As shown in Fig. 3, after many smaller regional airports are constructed after 2011, the average betweenness centrality begins to drop from the peak value. After many smaller regional airports are shut down from 2014 to 2016, the average betweenness centrality has a short-term increase. After that, the U.S. aviation network continues its path to decentralization.

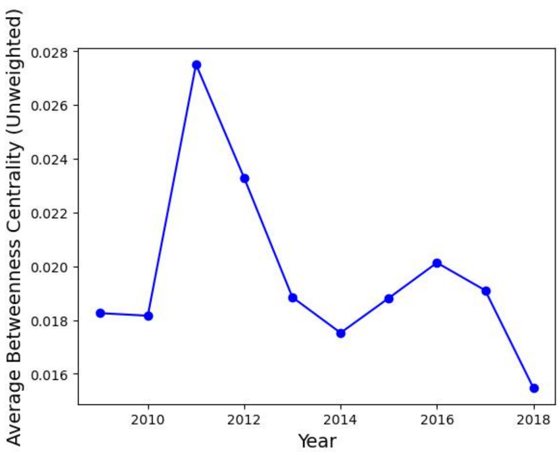


**Fig. 3.** The change in average betweenness centrality from 2009 to 2018

**Assortativity.** We present the change in the Gini coefficient in Fig. 4. As shown, the Gini coefficient displays an increasing trend, which suggests that the distribution of traffic among the different airlines in the network has become more unequal. This aligns with our analysis from the previous section: the disappearance of a Point-to-Point system and the consolidation of a Hub-and-Spoke system. However, we do notice an

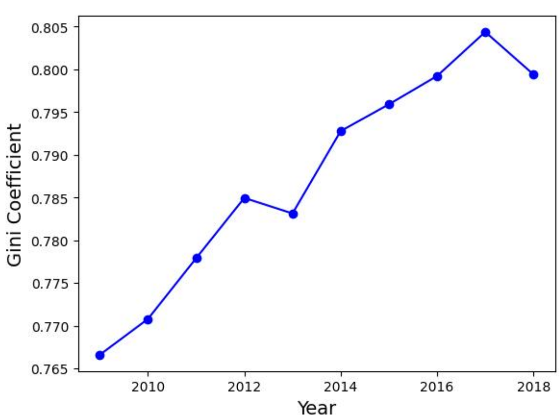


**Fig. 4.** The temporal evolution of the network's assortativity evaluated through the Gini coefficient.

abnormal decrease in the Gini coefficient from 2012 to 2013, which indicates that smaller airports were more dominant at that time. However, the Gini coefficient quickly bounces back. We again take a look at the centrality measure that started to increase in 2014, as shown in Fig. 3, and we are able to conclude that the first wave of small airports following the end of the recession quickly reached saturation in the market. The decrease in the Gini coefficient from 2012 to 2013 can be seen as a precursor to the closure of small airports and small airlines beginning in 2014.

**Small-World Properties.** We present the change in the clustering coefficient $C$ and the characteristic path length $L$. Theoretically, the two plots should have an inverse relationship, and they do closely match in Fig. 5.

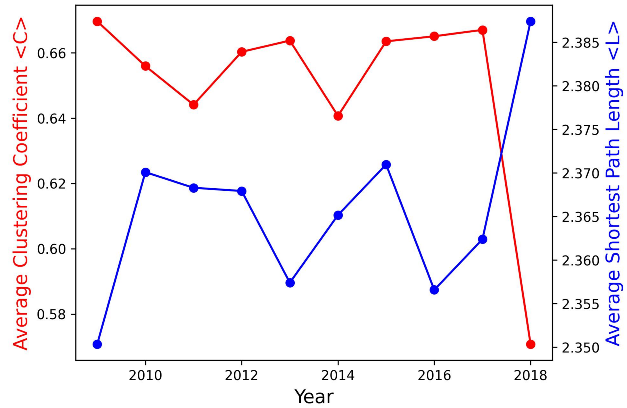


**Fig. 5.** The small-world properties of the network evaluated through the clustering coefficient $C$ and the characteristic path length $L$.

In the previous section, we concluded that there was an increase in the number of small airports between 2011 and 2014 and after 2017. However, while the characteristic path length decreased between 2011 and 2013, it showed an increasing trend after 2016. The explanation for this lies in the fact that prior to 2014, although there were many small airports, they were not reachable from large airports, hence the decrease in $L$. On the contrary, since 2017, with the strengthening of a more integrated hub-and-spoke system, smaller airports were no longer isolated from the hubs.

**Scale-Free Properties.** In Fig. 6, we plot the fraction of nodes having a degree $k$ or greater against the degree to determine if our air transportation network is scale-free for the network in 2010. We only show 2010's plot because when creating plots for each year, each plot was nearly identical, which makes sense since the change in degree distribution between years were not very significant. To identify a network as scale-free, this plot on a log-log scale should be linear. However, the line seems to curve down at around the $10^{1.6}$ degree mark, suggesting that the network is not scale-free.

In paper (2), Siozos-Rousoulis et al. found a very similar graph to ours, but claims that the network does exhibit scale-free properties because the degree distribution follows a power law with truncations, also known as a power law with an exponential cutoff, which is defined as,

$$f(x) \propto x^{-\alpha} e^{-\beta x} \qquad [10]$$

Note that the difference between this modified power law and the standard power law in equation 7 is that they multiplied an extra exponential $e^{-\beta x}$ to the original power law proportion. Although Siozos-Rousoulis et al. is correct that it can be modeled by a power law with truncations, it seems dubious since it still remains that there is a clear curvature in the plot.

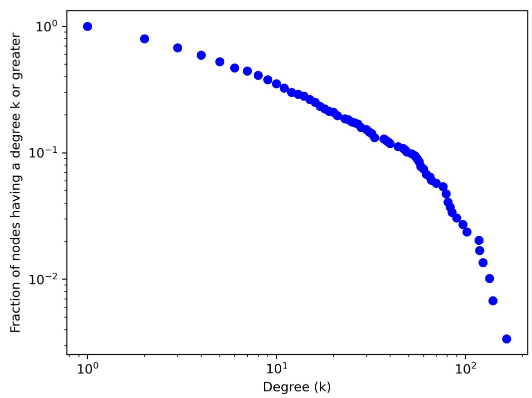


**Fig. 6.** An analysis on the degree distribution to determine scale-free properties.

**Targeted Attack Tolerance.** Figure 7 presents the plot of network resilience. The curve displays a convex shape, with its global minimum in 2004, which coincides with the rise and fall of small airports' development. The network's efficiency has shown a continuous decrease in resilience after the 2008 recession. However, with the integration of smaller airports into the hub-and-spoke system since 2017, there has been a noticeable increase in efficiency.

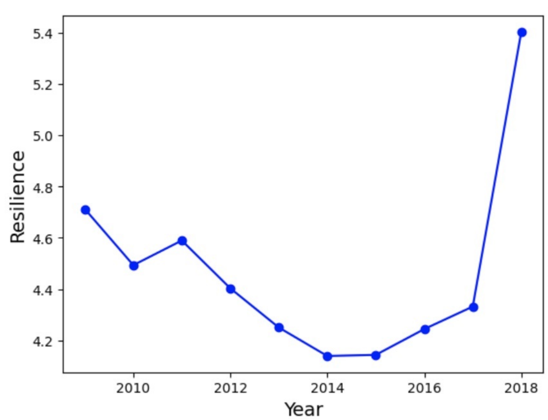


**Fig. 7.** The resilience of the network.

Figure 8 presents the plot of how quickly the network's proportion of the giant connected component decreases as we remove nodes of the highest betweenness centrality. Just looking at the years 2009, 2013, and 2018, it takes 11 nodes removed for the proportion of the giant connected component to be near 0. We can also see that in 2013, the rate that the proportion of the giant connected component decreases the fastest, followed by 2009, and then 2018. When the top five nodes with the highest betweenness were removed, the proportions of the giant connected component were 0.44, 0.50, and 0.56 for 2013, 2009, and 2018 respectively. These results suggest that the network's resilience to targeted attacks has fluctuated over time. The worse performance against targeted attacks from 2009 to 2013 can be attributed to the increase of smaller airports, since the removal of large hubs would then sever the connection to more airports. Conversely, the improved performance against targeted attacks from 2013 to 2018 can be attributed to the decrease of smaller airports, since there are fewer connections to sever after the removal of large hubs.

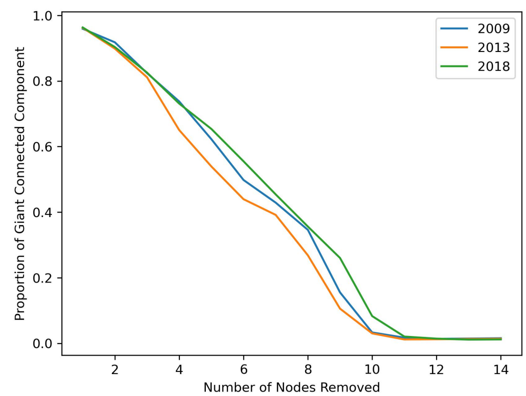


**Fig. 8.** Simulating a targeted attack on the U.S. air transportation network.

## Conclusions and Discussion

Due to the importance of domestic airlines in the US transportation system, it is imperative to assess its topological evolution and ensure that it is well-connected, efficient, and robust. In this paper, we study the temporal change of the US domestic airline network from 2009 to 2018. We connect the development of airports to the events affecting the industry.

Our main finding focuses on the development of small airports during the decade of 2009. From 2009 to 2014, following the end of the recession, there was an expansion of smaller regional airports. However, they remained relatively separated from larger airports. From 2014 to 2017, smaller airports were shut down due to market saturation. And since 2018, there has been a reinforcement of the hub-and-spoke system, with smaller airports opening and connecting to major hubs, resulting in a more interconnected network, which continues to be the status quo of today.

One limitation was that the data we used did not include the shortest air distances between airports. As a result, we were unable to calculate metrics such as efficiency, which generally measures how efficiently information is exchanged within a network. We considered trying to compute them with the data on the latitude and longitude of the airports, but given the time constraints and ample findings already, we did not end up doing it.

An area of study we could've considered delving further into is looking at the different airlines operating within the network. By studying the expansion of major carriers, we could have observed how different airlines are contributing to the evolution of the network over time. This different perspective of our network would have allowed us to identify additional trends and patterns.

Another area of study we could have looked into is how the network changes over different time periods, such as seasons, months, holidays, and major events. These are times when the number of flights may drastically increase or decrease, which could have significant effects on the network. By studying these changes, we could have gained insights into how the network responds to different types of disruptions and how it adapts to changes in demand.

Finally, when studying targeted attacks on the network, we could have clustered airports into large and small hubs. Since terrorists are more likely to attack major hubs due to their strategic importance, this would more accurately simulate a realistic targeted attack and help us better identify potential vulnerabilities and make recommendations for improving the security of the network.

Our work on the airline network is applicable to the analysis of similar transportation networks. The developed code has been made publicly available to facilitate the reproduction of the results and application on alternative networks.

## Code

Our code for the calculations and graphs included in this paper can be found here in this repository: https://github.com/cheungbrenden/math168project

**ACKNOWLEDGMENTS.** We thank Professor Porter for his helpful discussions and valuable suggestions. We thank Grace for her teaching of codes that are essential to the computation and visualization of our data.

1. 2018 traffic data for u.s. airlines and foreign airlines operating flights to and from the u.s. (https://www.bts.dot.gov/newsroom/2018-traffic-data-us-airlines-and-foreign-airlines-us-flights) (2020) Accessed: December 2, 2022.
2. L Siozos-Rousoulis, D Robert, W Verbeke, A study of the us domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016. *J. Transp. Secur.* **14**, 55–78 (2021).
3. Y Mu, Airline delay and cancellation data, 2009 - 2018 (https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018) (2019) Accessed: 1 November 2022.
4. A Hagberg, D Schult, P Swart, Exploring network structure, dynamics, and function using networkx (https://networkx.github.io) (2008).
5. M Newman, *Networks.* (Oxford university press), (2018).
6. EL Platt, *Network Science with Python and NetworkX Quick Start Guide: Explore and Visualize Network Data Effectively.* (Packt Publishing Ltd), (2019).