Data Visualization and Delay Prediction of Airline Flight Network.

July 20, 2022.

Sarthak Sharma,
M.S. (Computational Science),
Department of Scientific Computing,
Florida State University.

Contents

Introduction

Introduction

- On-time performance is crucial for successful flight operations of any airline.
- Flight departs from origin airport and arrives at destination airport.
- Important flight events:
  - **OUT**: Boarding of passengers and luggage is completed. Aircraft doors are closed. Aircraft “disconnects” from origin airport’s terminal.
  - **OFF**: Aircraft flies up from runway at origin airport.
  - **ON**: Aircraft lands on runway at destination airport.
  - **IN**: Aircraft connects to terminal at destination airport. Doors open. Passengers deboard. Luggage taken out by staff.

Introduction (continued)

- Delay:

$$D = t_{act} - t_{sch}$$

where t_{act} = event's actual time, and t_{sch} = event's scheduled time.

- Delay can be positive, zero, or negative.
- **Flight's departure delay:**

$$D_{dep} = O_{act} - O_{sch}$$

where O_{act} = flight's actual OUT time, and O_{sch} = flight's scheduled OUT time.

- **Flight's arrival delay:**

$$D_{arr} = I_{act} - I_{sch}$$

where I_{act} = flight's actual IN time, and I_{sch} = flight's scheduled IN time.

Introduction (continued)

Causes of flight delays:

- Bad weather.
- Airport circumstances.
- Arrival delay of one or more feeder flights (from which passengers are obtained).
- Technical problems in aircraft.
- Circumstances during flight (e.g., medical or other emergency of passenger or crew, necessitating temporary unplanned halt of aircraft at intermediate airport).
- Busy or temporarily unavailable runway at destination airport, causing wait before landing, etc.

Introduction (continued)

Copa Airlines:

- National airline of Republic of Panama.
- Hub airport: PTY, located at Panama City. Most flights have PTY as either origin or destination.
- **Inbound** flights arrive at PTY airport.
- **Outbound** flights depart from PTY airport.
- Most passengers of Copa Airlines are “transit” passengers (transfer from inbound flights to outbound flights at PTY airport).
- Thus, most inbound flights are “feeders” to outbound flights.
- Challenge: to prevent delay propagation in flight network.

Introduction (continued)

Resources required in commercial aviation industry:

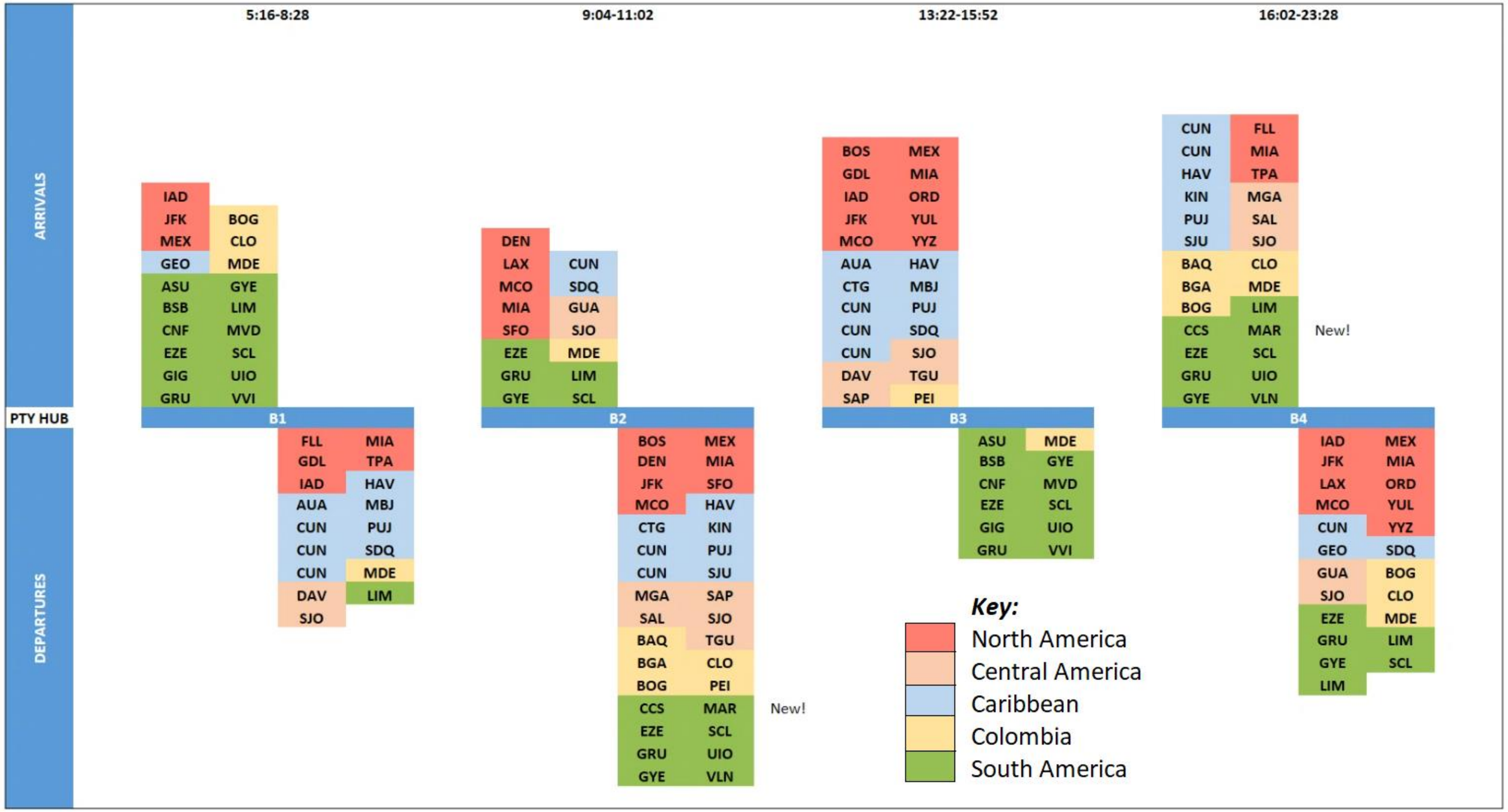
- Aircraft,
- Airports,
- Flight crew,
- Passengers,
- Engineers and technical staff,
- Aviation fuel,
- Ground staff,
- Security arrangements,
- Financial resources (capital),
- Other resources.

Introduction (continued)

Banks at PTY airport: B1, B2, B3, B4, B5, B6.

Depend on:

- **Time** of inbound flight's arrival or outbound flight's departure.
- **Region** of inbound flight's origin or outbound flight's destination.



Introduction (continued)

Flights considered for analysis: flights satisfying all the following criteria:

- Flights of Copa Airlines (and not the subsidiary Wingo / Copa Colombia).
- Origin different from destination.
- PTY airport as either origin or destination.
- No missing or negative values in data provided by Copa Airlines.

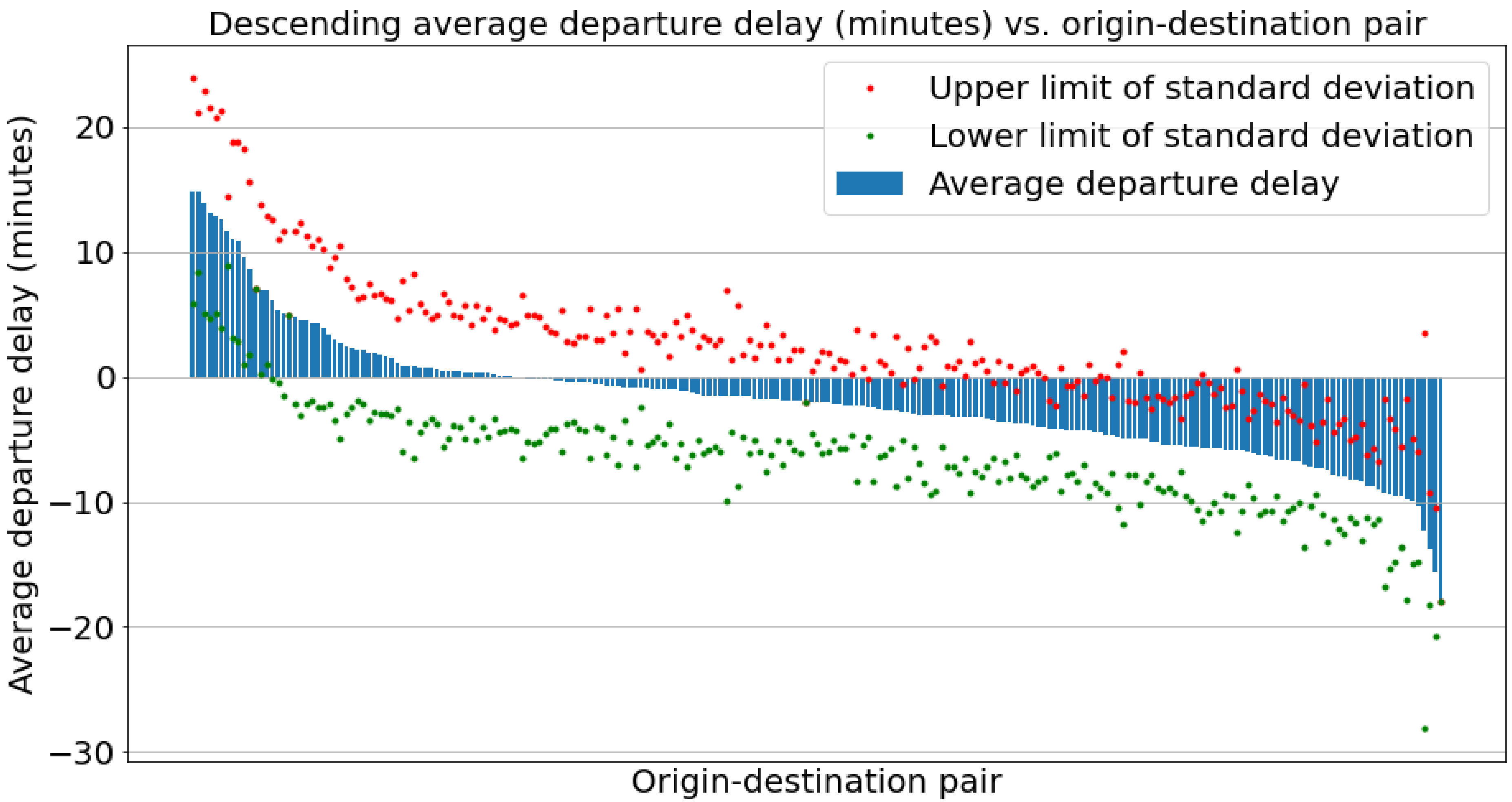
Flight's unique identification: concatenation of the following:

- Scheduled departure date (GMT) in yyyy-mm-dd format.
- Origin and destination codes.
- Scheduled departure time (GMT) in 24 hours format hh:mm.
- Flight number.

No two flights will have the same combination of the above.

Delay Probability Analysis

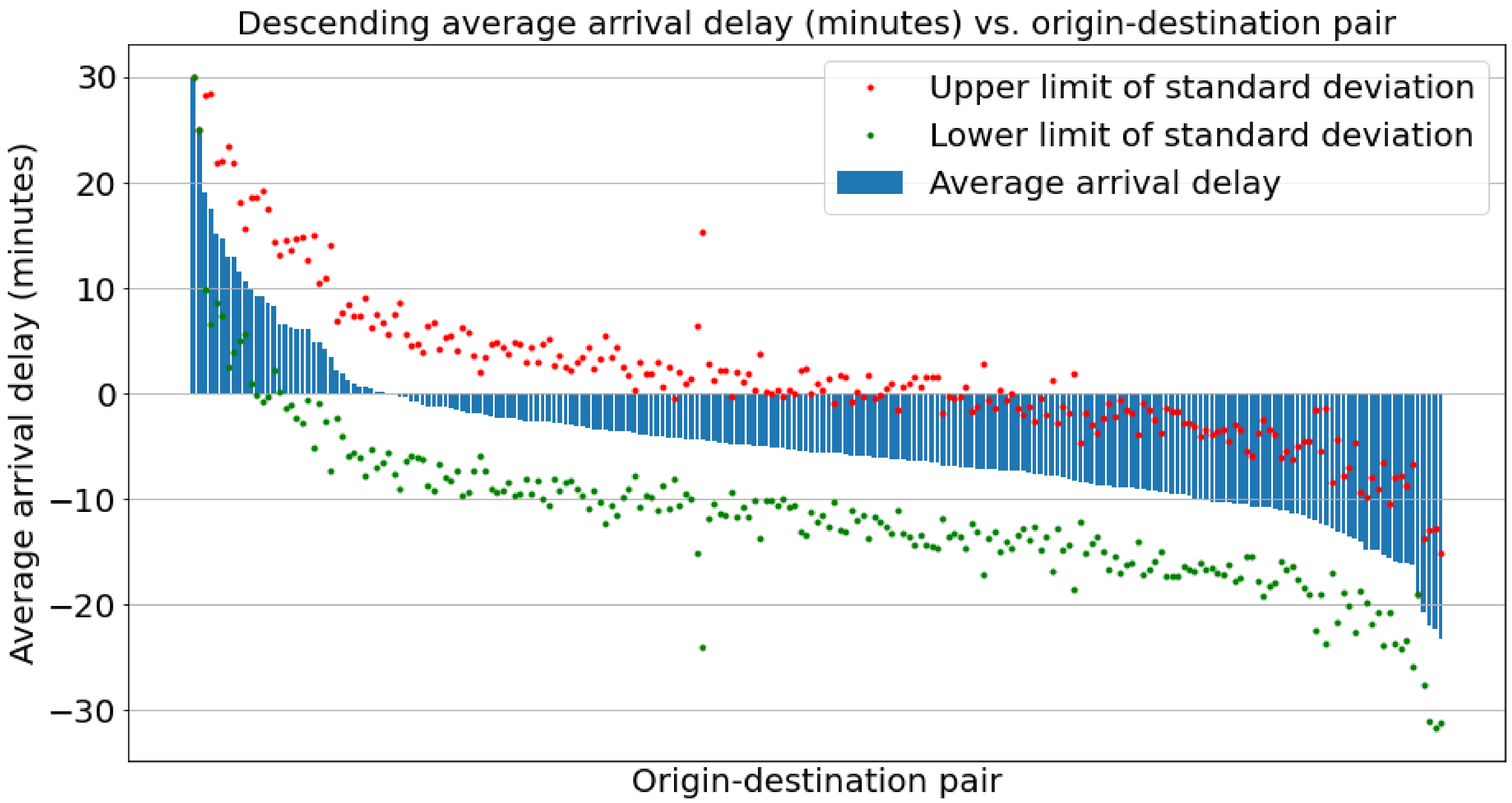
Delay Probability Analysis



Preliminary analysis

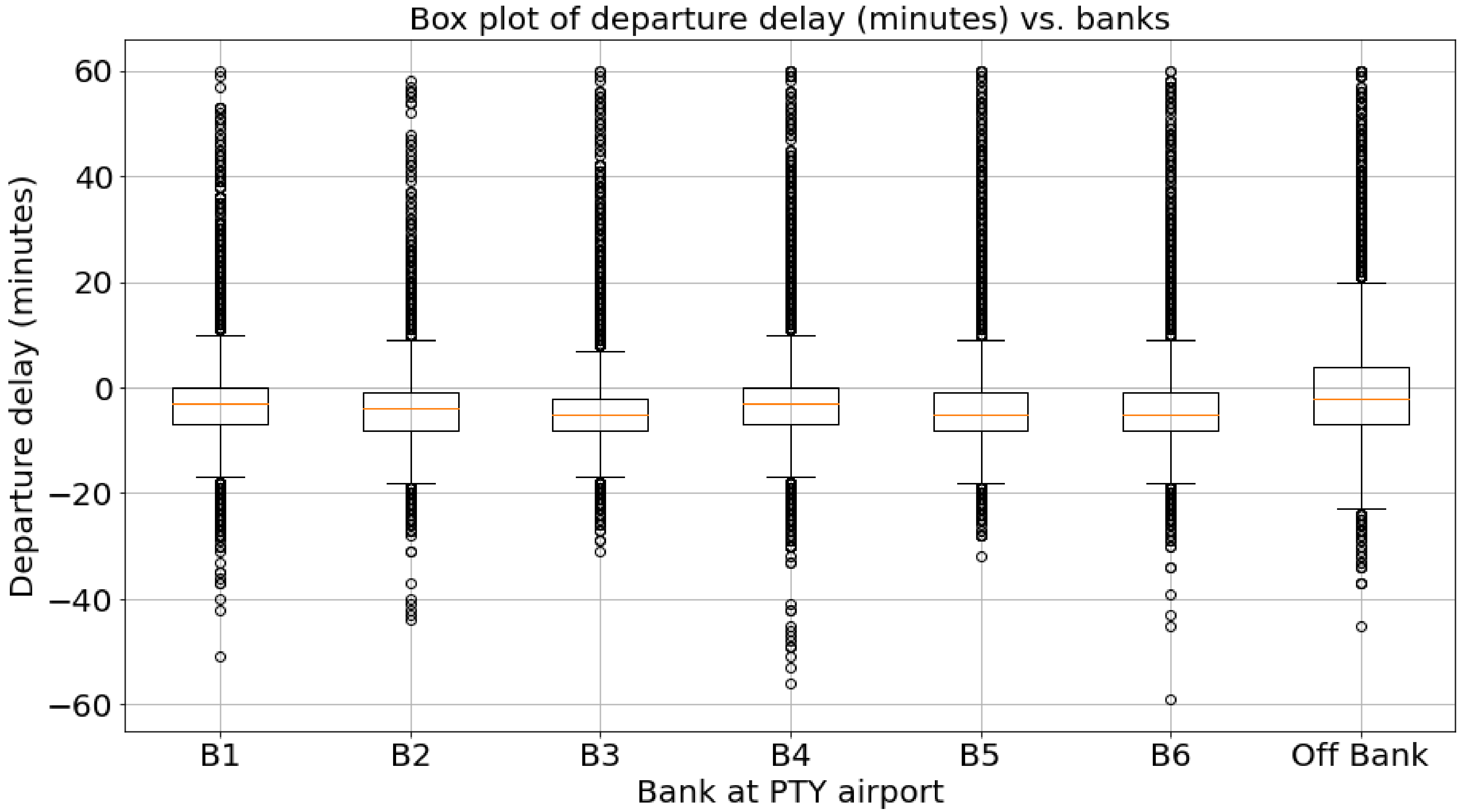
- Removed outliers (delays with magnitude > 60 minutes).
- Plotted average delays of different flight routes, with upper and lower limits of standard deviation.

Delay Probability Analysis (continued)



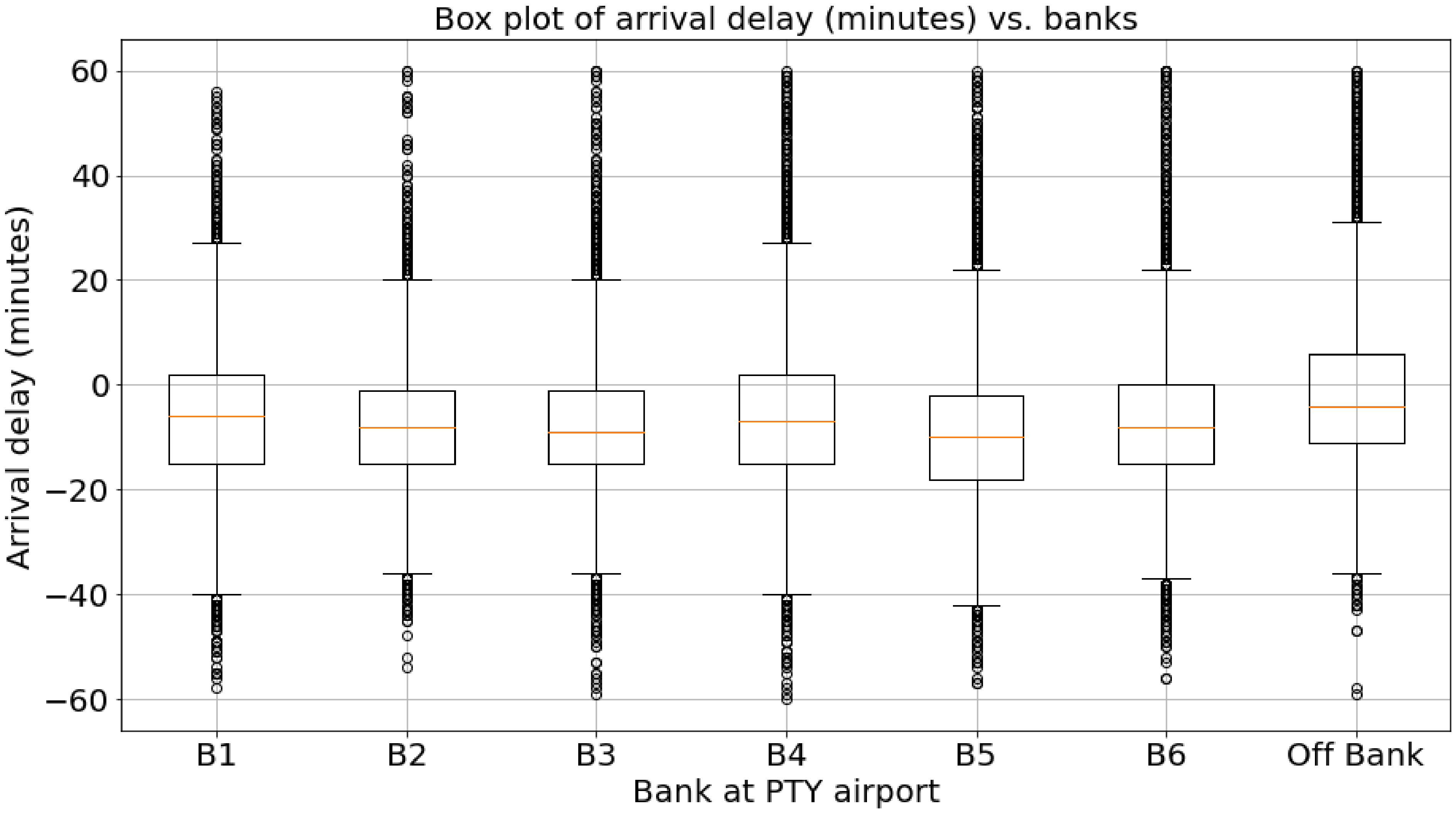
- Most routes of Copa Airlines have negative average delays (in both departure and arrival).
- Larger fraction of flights arrive early, compared to flights departing early.

Delay Probability Analysis (continued)



- Next, we plot the average delays of Copa Airlines flights at different banks at PTY airport.
- Average departure delay can be widely varied (and positive for banks B1, B4, and B6).

Delay Probability Analysis (continued)



- Average arrival delay is small and negative (i.e., flights arriving early) for all banks.
- Only for off-bank flights, both average departure delay and average arrival delay are large.

Delay Probability Analysis (continued)

Delay probabilities and their correlation, if any, with flight duration:

- Delay categories:

negative or zero delay,	1 to 15 minutes,	16 to 30 minutes,
31 to 45 minutes,	46 to 60 minutes,	more than 60 minutes.

- For each origin-destination pair (OD), the **delay probability** in a delay category is:

$$P_C = \frac{N_C}{N_{OD}}$$

where, N_C is number of flights of that OD in that delay category,
N_{OD} is total number of flights of that OD,
provided that total number of flights of that OD is not less than a certain value (e.g., 30), since if there are too few flights, then the calculated probability may not be reliable.

Delay Probability Analysis (continued)

- **Pearson's correlation coefficient** between delay probability and flight duration, to see if these two quantities are correlated in any way:

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\left[\sum_{i=1}^{n}\{(x_i - \bar{x})^2\}\right]\left[\sum_{i=1}^{n}\{(y_i - \bar{y})^2\}\right]}}$$

where, x_i and y_i are data samples of the two quantities (x and y),

n is number of data samples in each quantity (must be same for both x and y),

$\bar{x}$ is average of all x_i,

$\bar{y}$ is average of all y_i.

$r \in [-1, +1]$.

$r = -1$ denotes perfect negative correlation (as x increases, y decreases linearly).

$r = +1$ denotes perfect positive correlation (as x increases, y also increases linearly).

$r = 0$ denotes that the two quantities are not correlated in any way.

Delay Probability Analysis (continued)

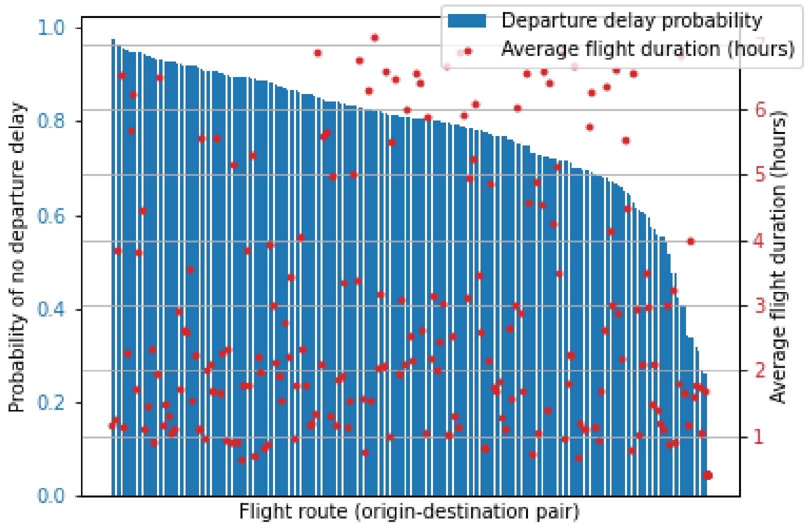
Average **departure** delays:

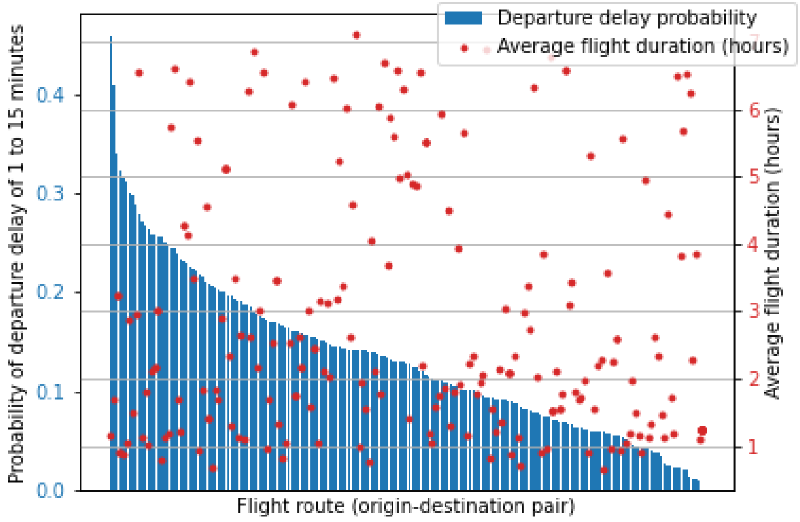
Departure delay category (minutes)	Correlation coefficient between departure delay probability and average flight duration	Fraction of flights in this departure delay category
≤ 0	0.17098	83.6 %
1 to 15	0.07274	11.5 %
16 to 30	0.00012	2.3 %
31 to 45	0.01290	0.8 %
46 to 60	− 0.12173	0.3 %
≥ 61	0.02813	1.3 %

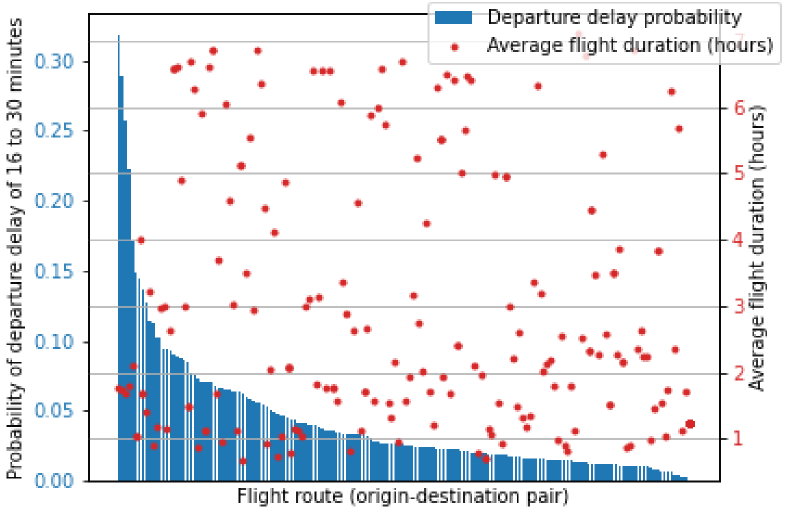
For each departure delay category, we see that there is not a strong correlation between departure delay probability and average flight duration.

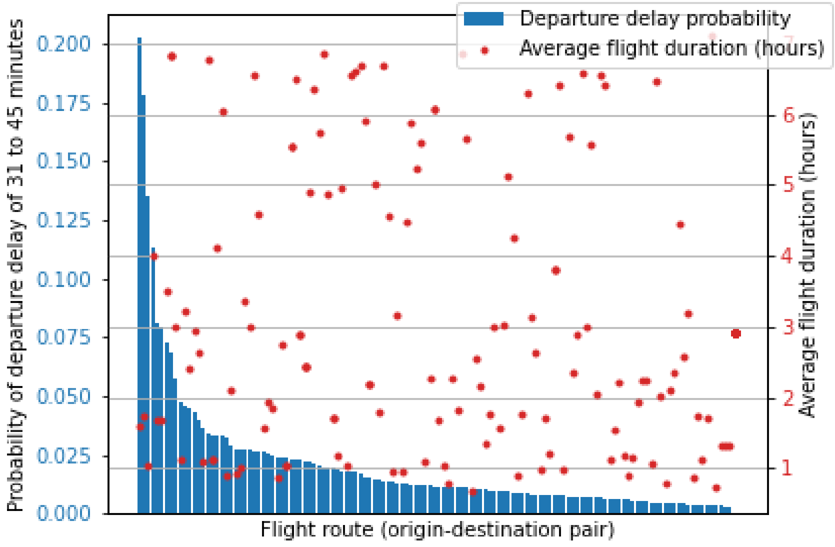
Delay Probability Analysis (continued)

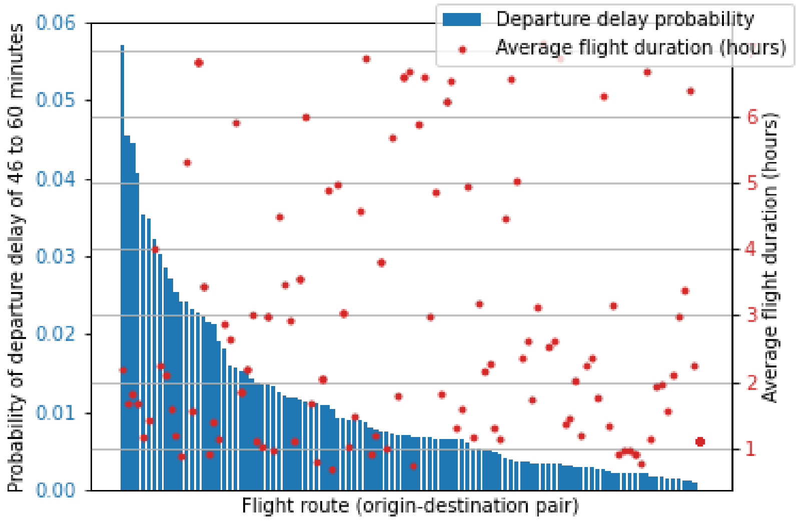
Departure delay analysis:

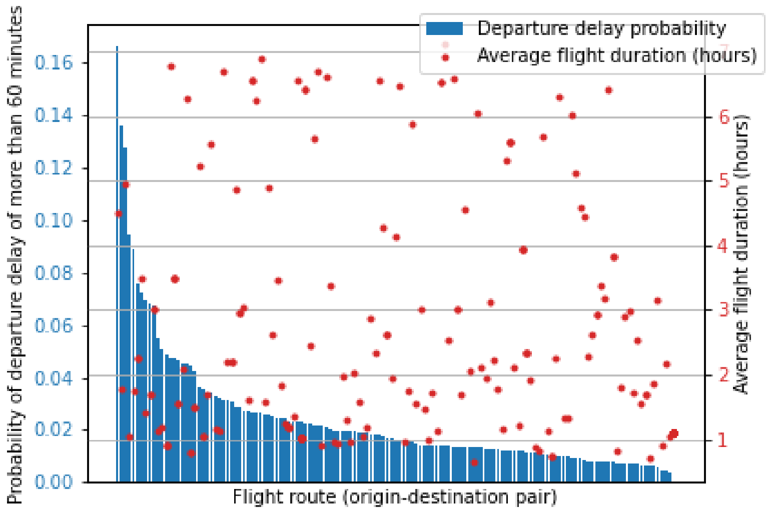












Delay Probability Analysis (continued)

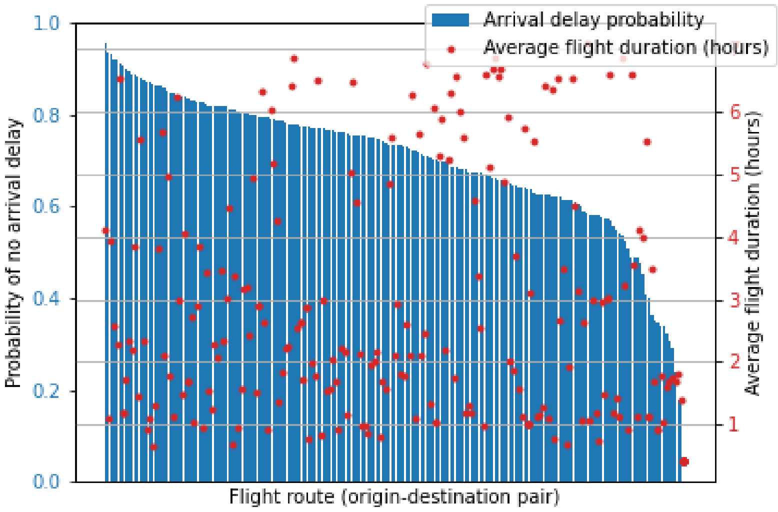
Average **arrival** delays:

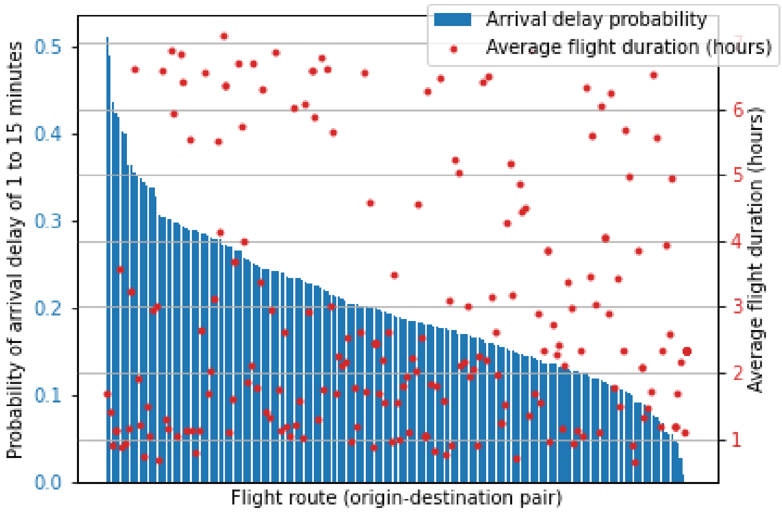
Arrival delay category (minutes)	**Correlation coefficient between arrival delay probabilities and average flight durations**	**Fraction of flights in this arrival delay category**
≤ 0	0.16510	75.2 %
1 to 15	0.07465	18.5 %
16 to 30	0.08033	3.7 %
31 to 45	−0.05864	0.9 %
46 to 60	−0.06731	0.4 %
≥ 61	0.02198	1.3 %

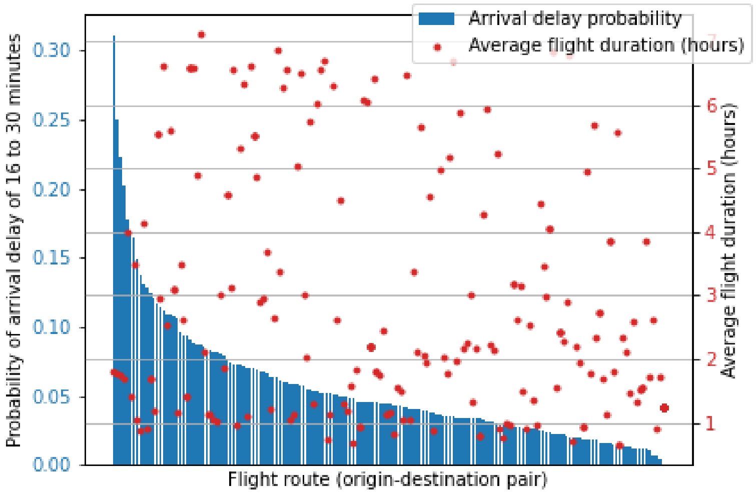
For each arrival delay category, we see that there is not a strong correlation between arrival delay probability and average flight duration.

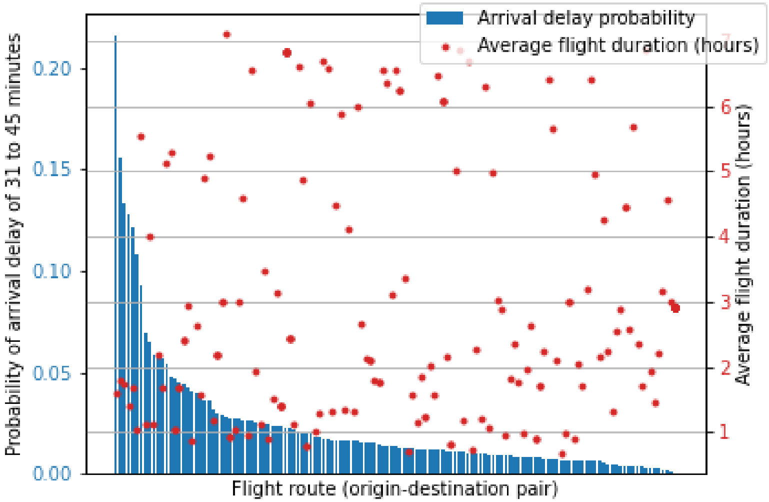
Delay Probability Analysis (continued)

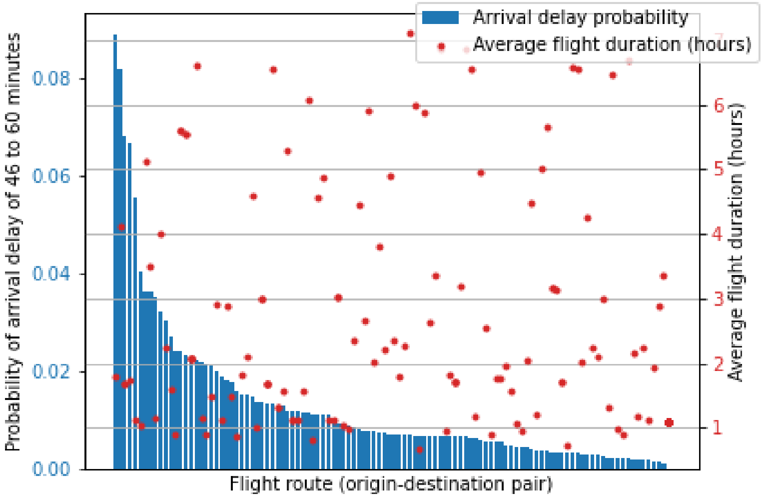
Arrival delay analysis:

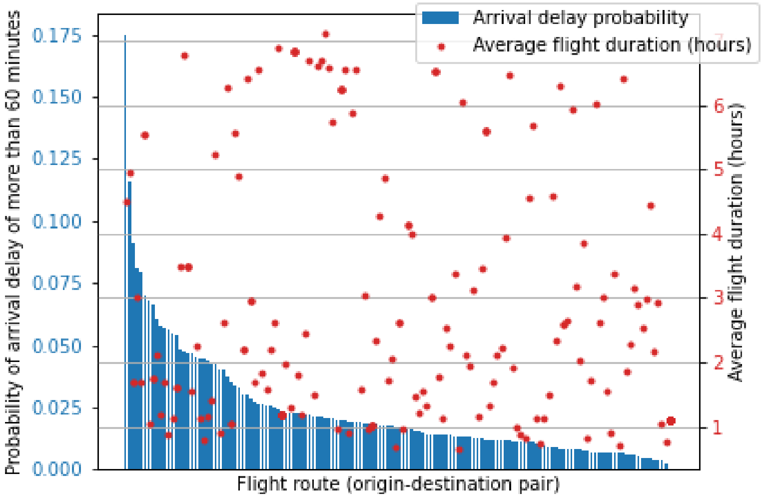






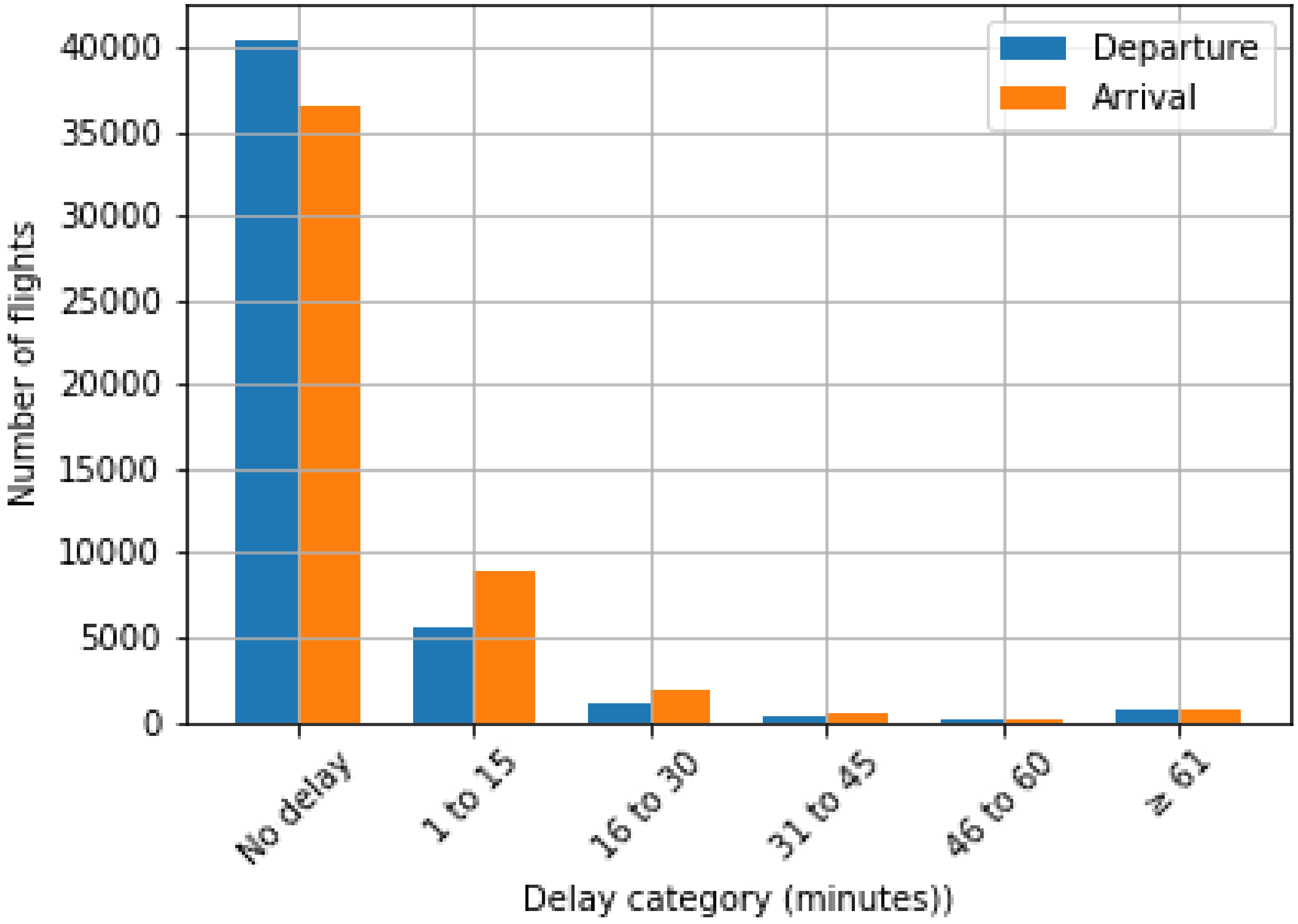






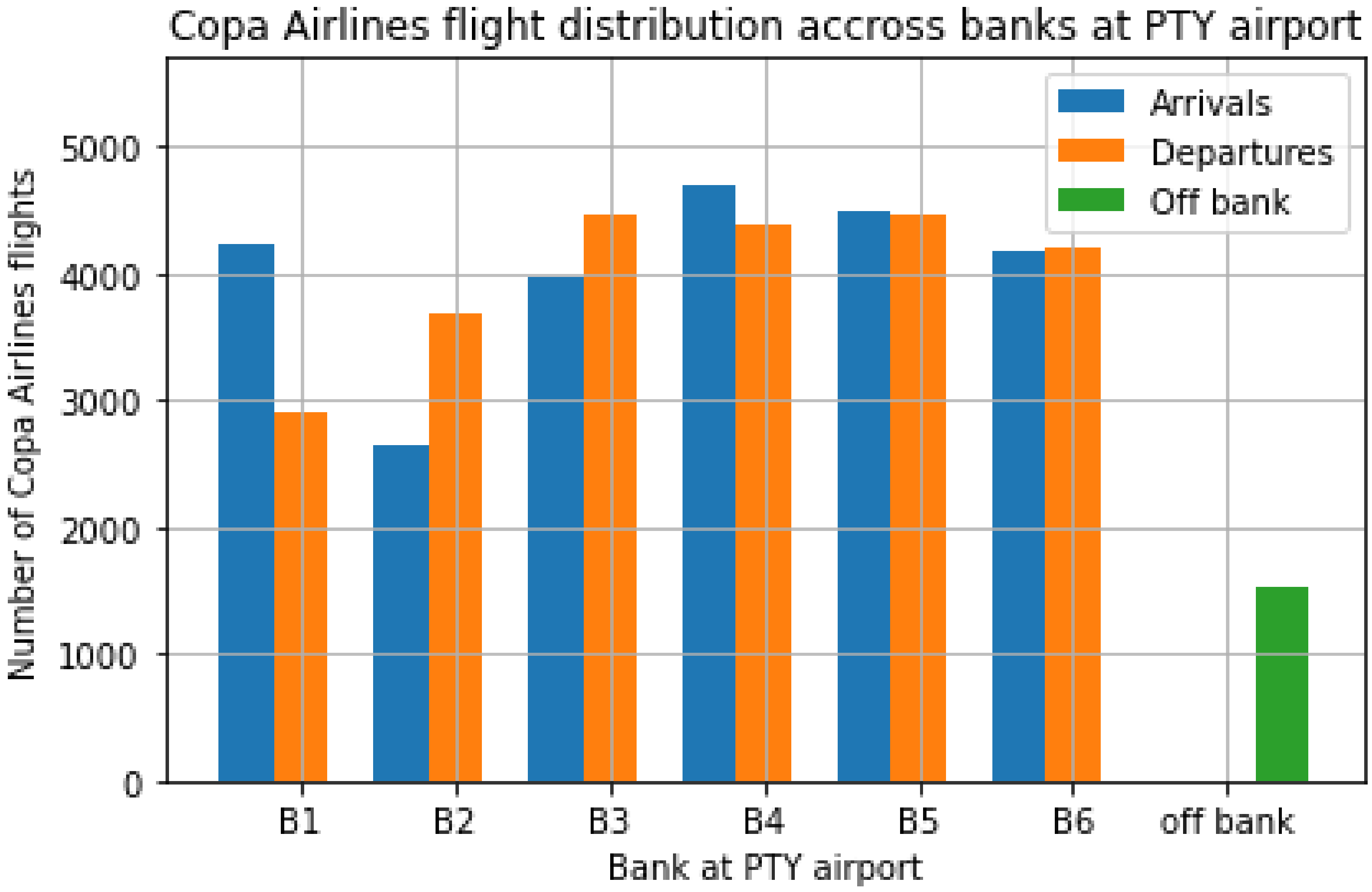
Delay Probability Analysis (continued)

Most flights of Copa Airlines have zero or negative delay, for both departure and arrival.



Data Visualization

Introduction (continued)



- At PTY airport, aircraft may arrive at one bank, and depart from another bank.
- Some aircrafts may not be assigned to any bank.

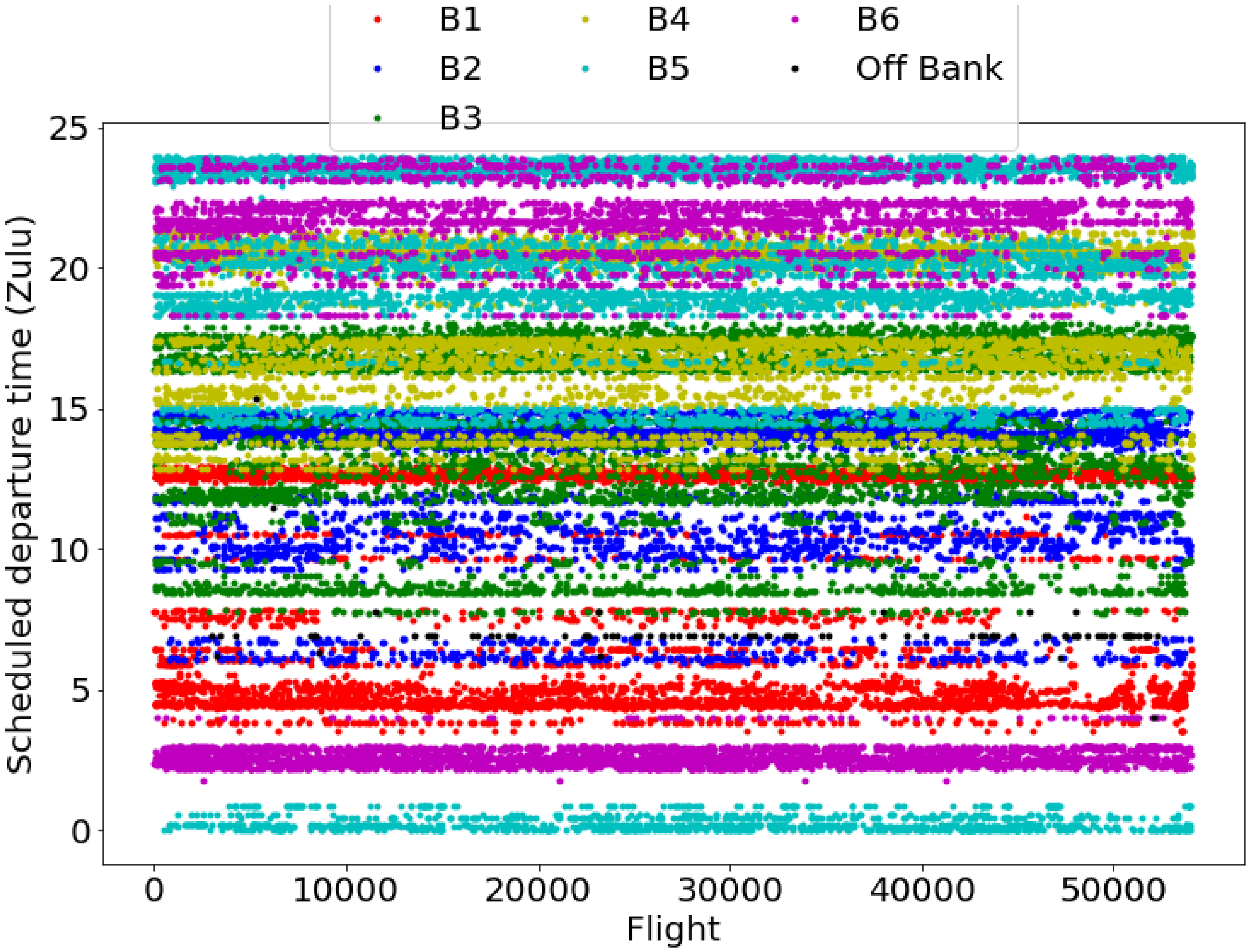
Off bank: 1522 flights.

Bank	Arrival	Departure
B1	4221	2894
B2	2639	3679
B3	3980	4464
B4	4707	4396
B5	4495	4476
B6	4181	4208
Total	*24223*	*24117*

Data Visualization

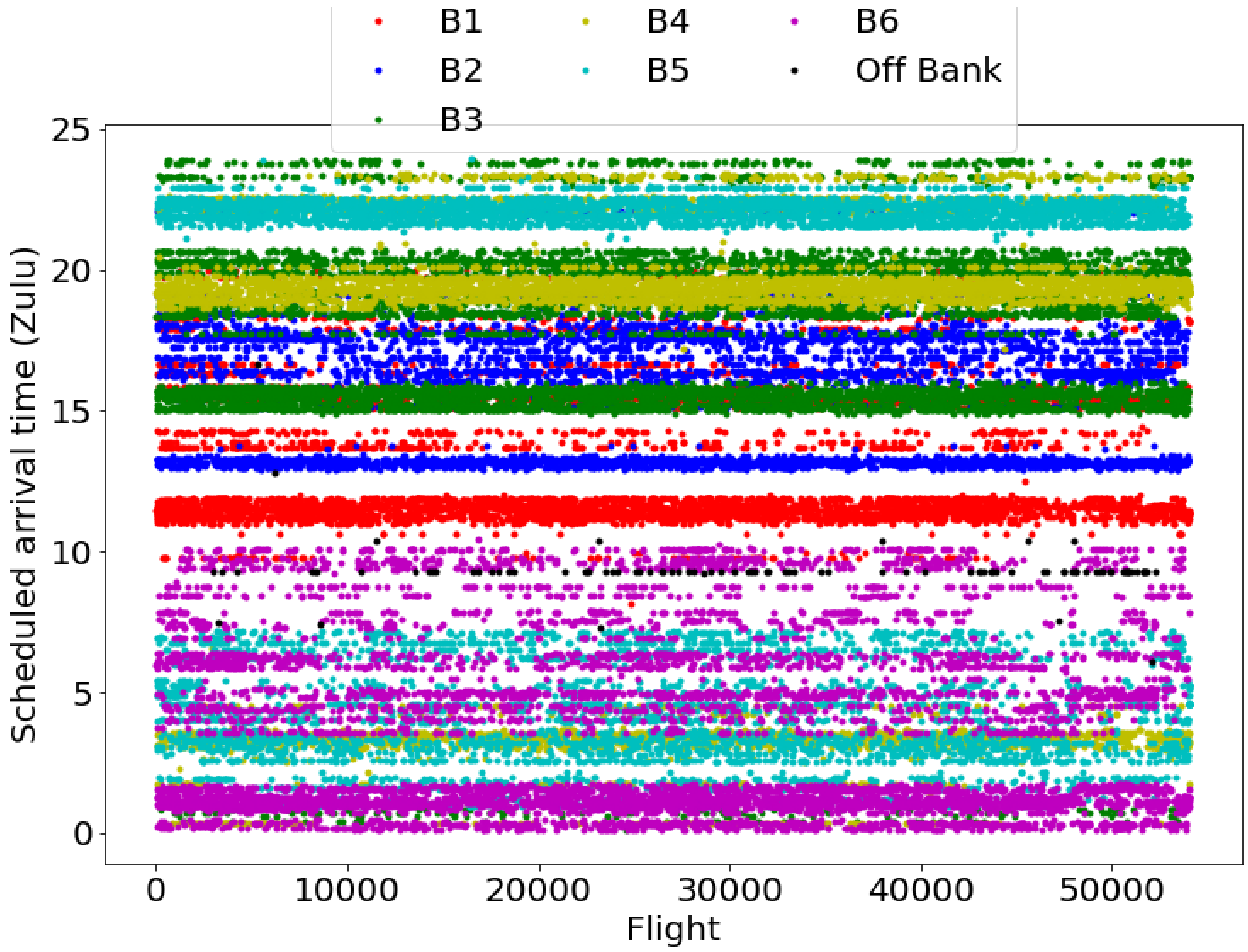
Visualization of scheduled departure and arrival times with respect to different banks at PTY airport:

- For departure:

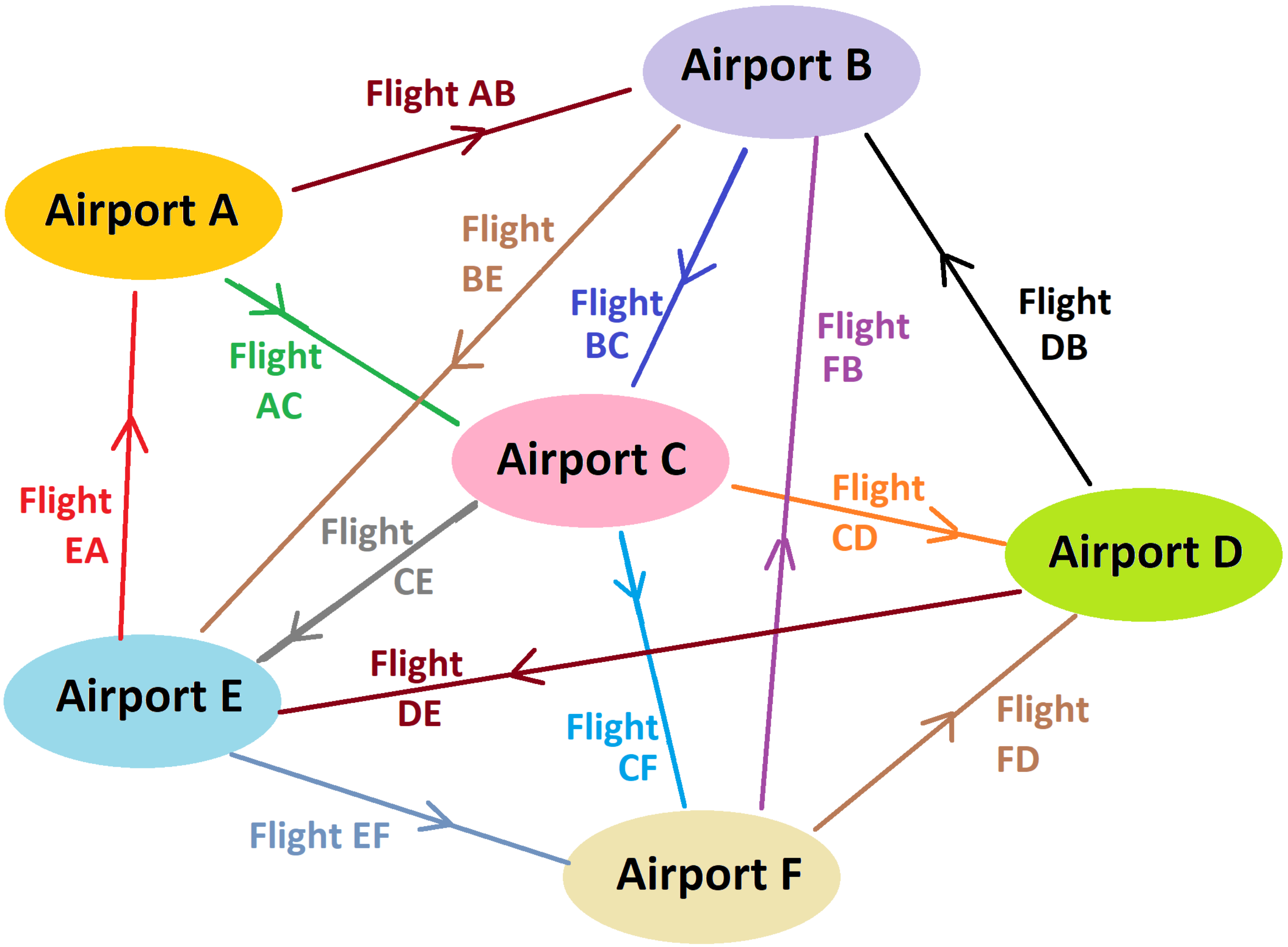


Data Visualization (continued)

- For arrival:

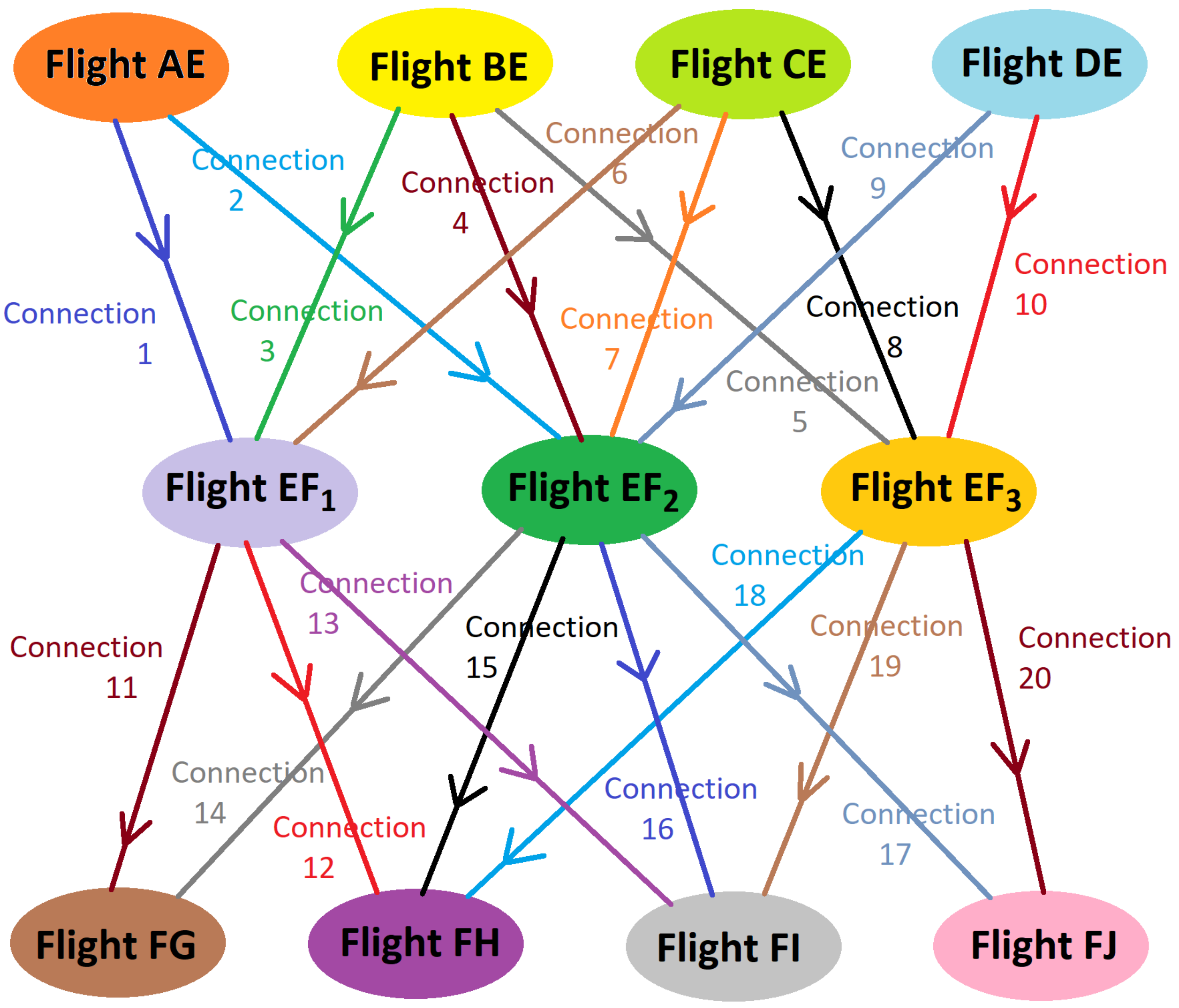


Data Visualization (continued)



Visualizing flight networks:

- Airports as nodes, and flights as links between nodes.
- Seems more intuitive, as it is closer to the real-world representation.



Data Visualization (continued)

- Flights as nodes, and transfer of resources (passengers, crew, luggage, aircraft) as links between nodes.
- Not analogous to the real world.

Data Visualization (continued)

We are using the second method (flights as nodes, and connections between those flights as links between nodes). Reasons:

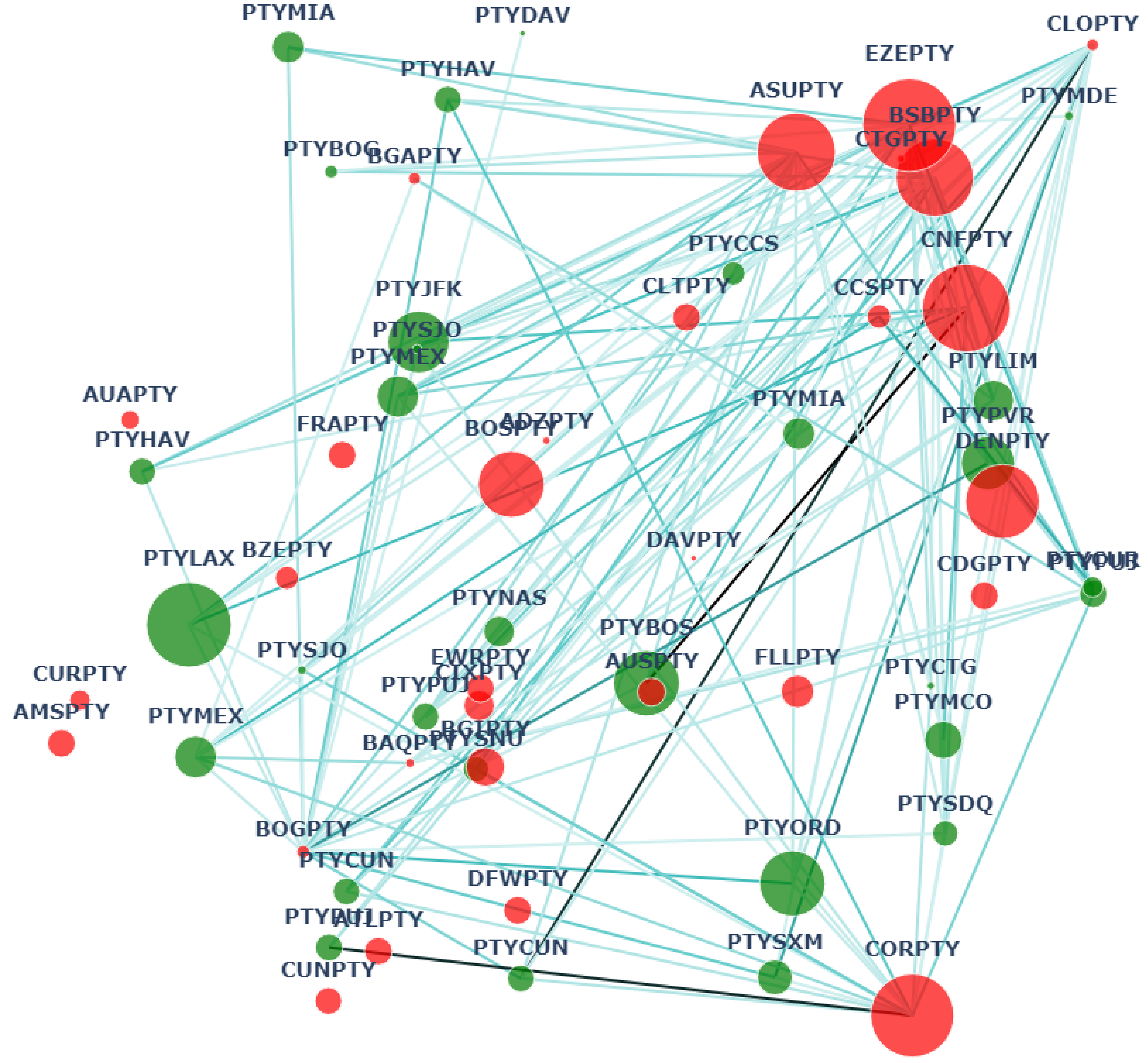
- For a flight network like that of Copa Airlines, where a particular airport (e.g., PTY) is the hub, it makes more sense to visualize flights as nodes.
- Having airports as nodes will just create a hub-and-spoke visualization for every visualization case, where airports other than the hub would be connected to a single node (the hub) via a single edge.
- Having flight connections as edges may reveal important data about transfers of resources between different flights, aircraft maintenance between two consecutive flights, etc.
- Above may be difficult when using the first method (airports as nodes and flights as edges).

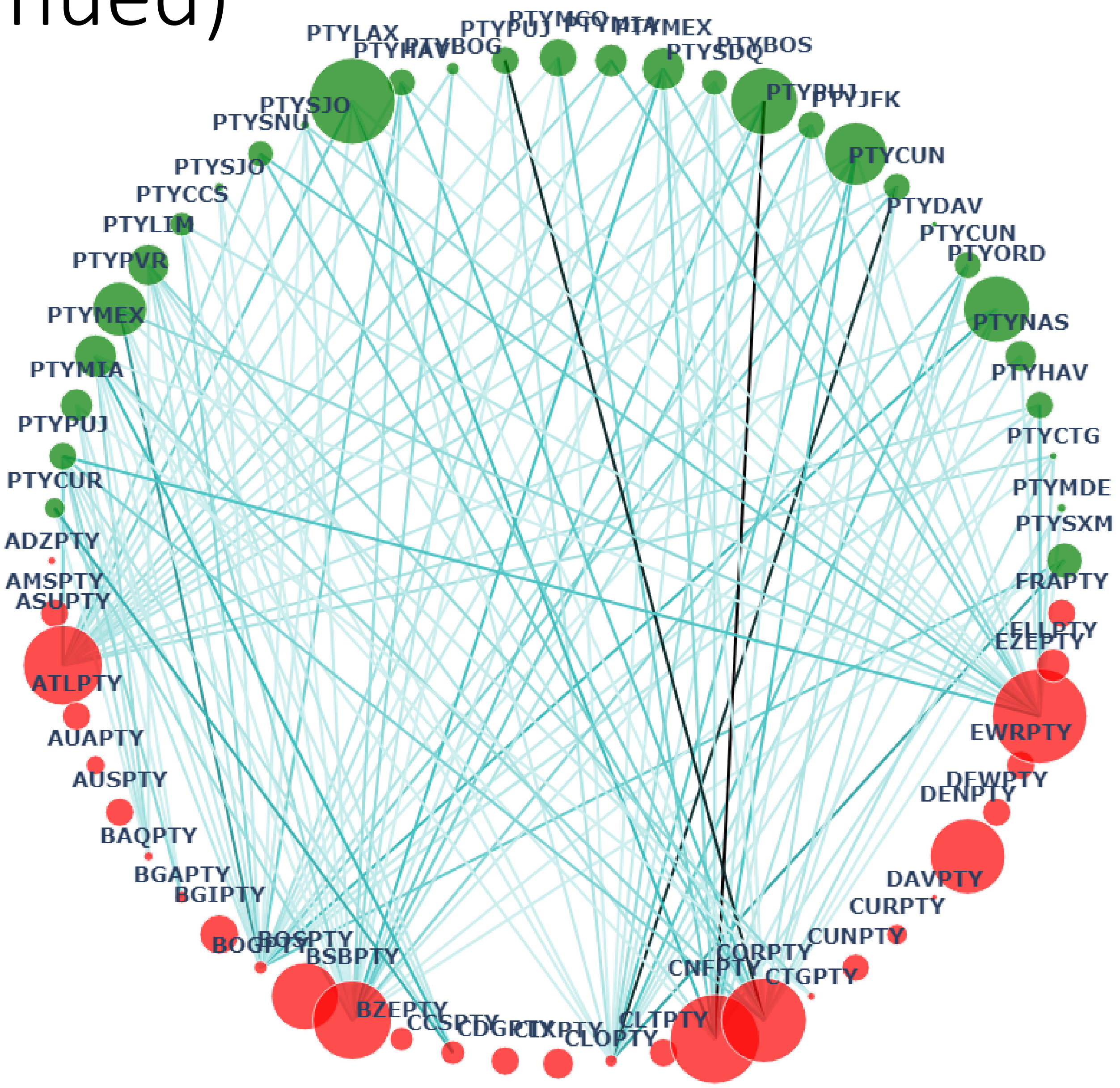
Data Visualization (continued)

For first few visualization exercises:

- Using Python, Plotly, Networkx, Pandas, etc. in Jupyter Notebook.
- Nodes represent flights.
- Node color is red for inbound flights (to PTY), and green for outbound flights (from PTY).
- Node size is directly proportional to flight duration (but can be some other quantity).
- Edges represent transfer of passengers from inbound (feeder) flights to outbound (non-feeder) flights.
- Edge color is directly proportional to number of passengers transferred (dark for more passenger transfer, light for less passenger transfer).
- Thickness is constant for all edges, but can be changed to vary according to some parameter.

Data Visualization (continued)



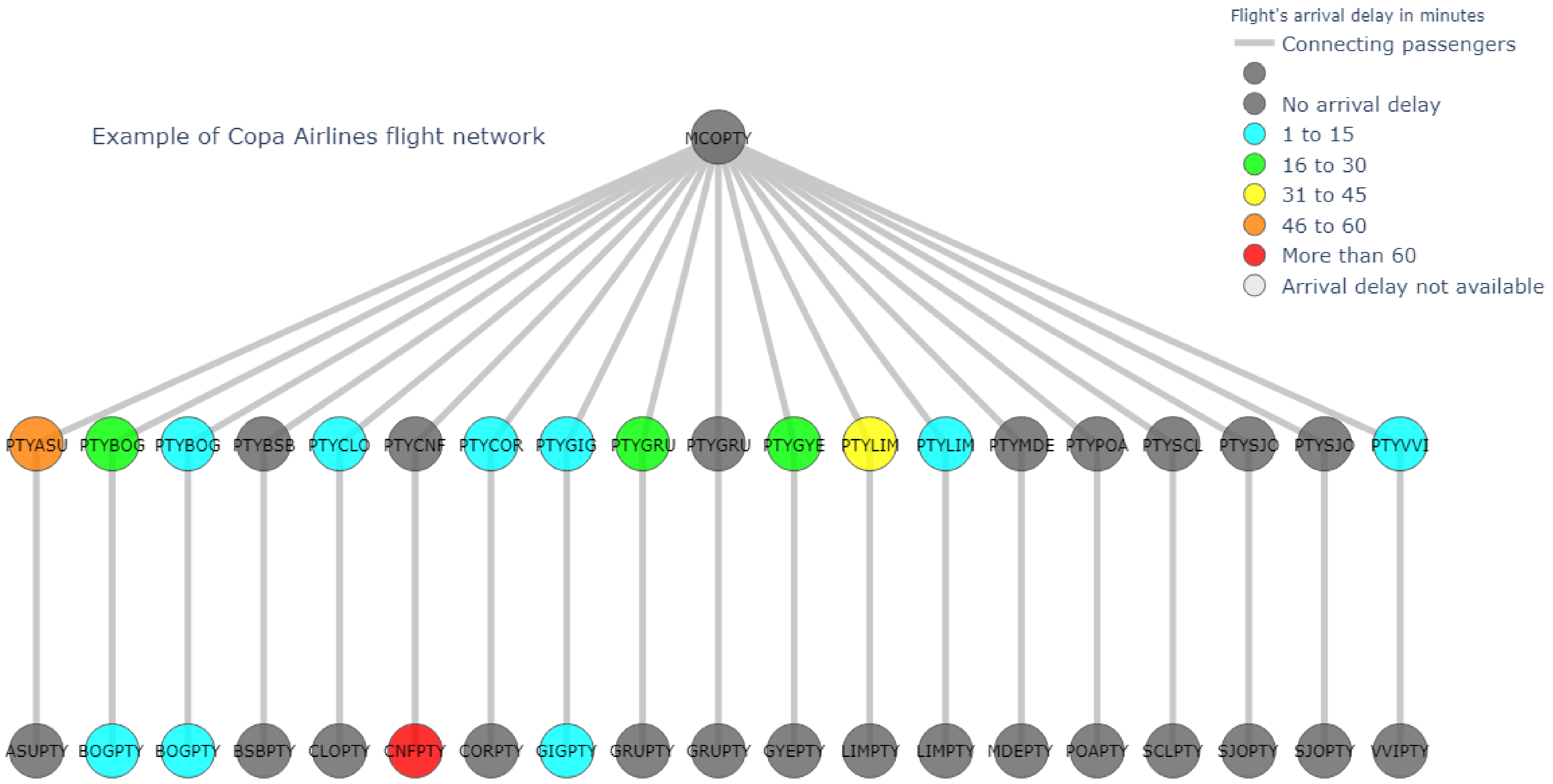


Data Visualization (continued)

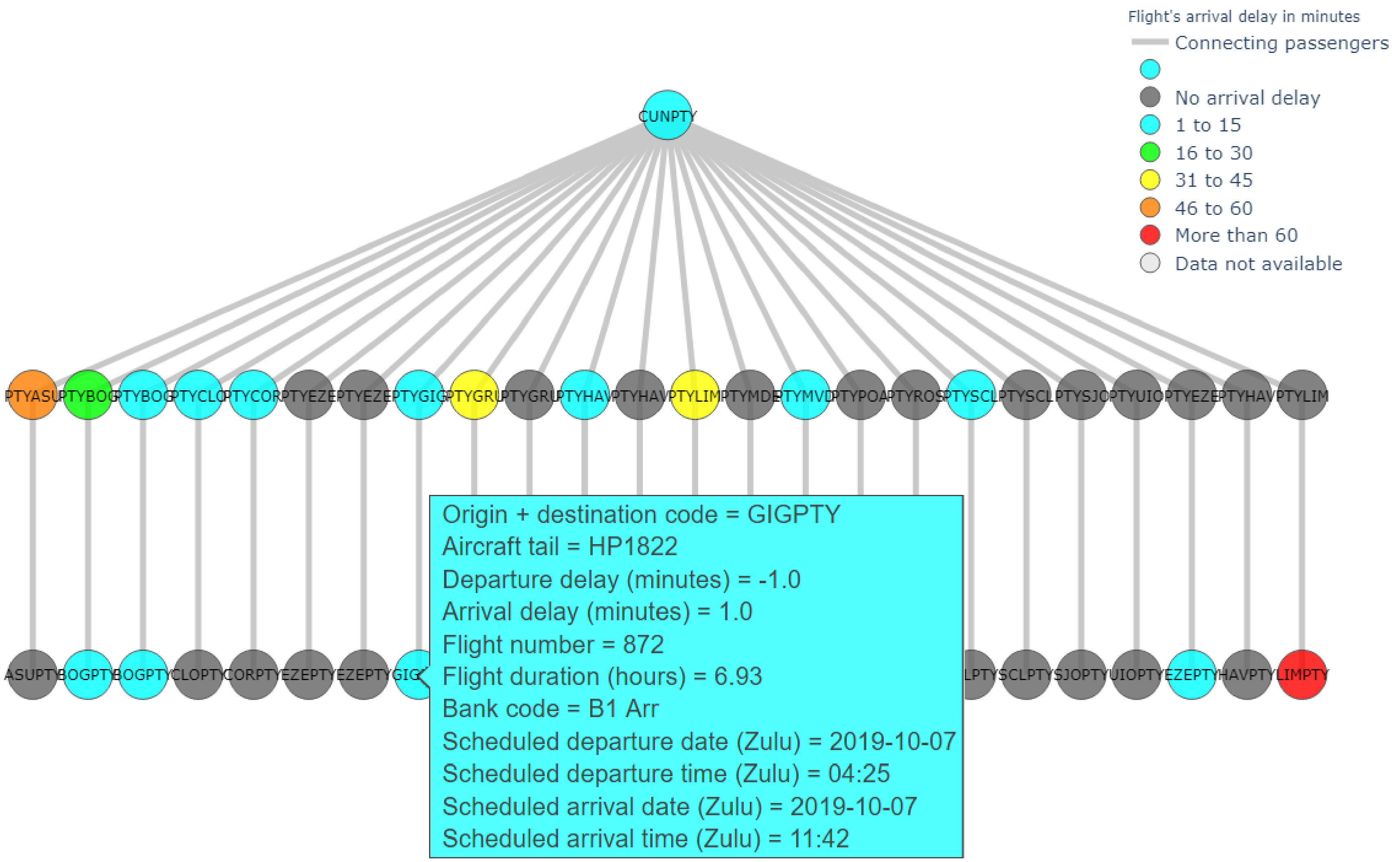
Tree showing one inbound flight, outbound flights fed by it, and return flights of same aircraft to PTY:

- Tree's structure has flight nodes arranged in three horizontal rows or levels:
  - Top: feeder (inbound) flight,
  - Middle: outbound flights getting passengers from feeder flight above,
  - Bottom: return (inbound) flights of the same aircraft (tail) as the outbound flights above.
- Passenger transfer direction is from top to bottom.
- Nodes colored according to flights' delay categories.

Data Visualization (continued)

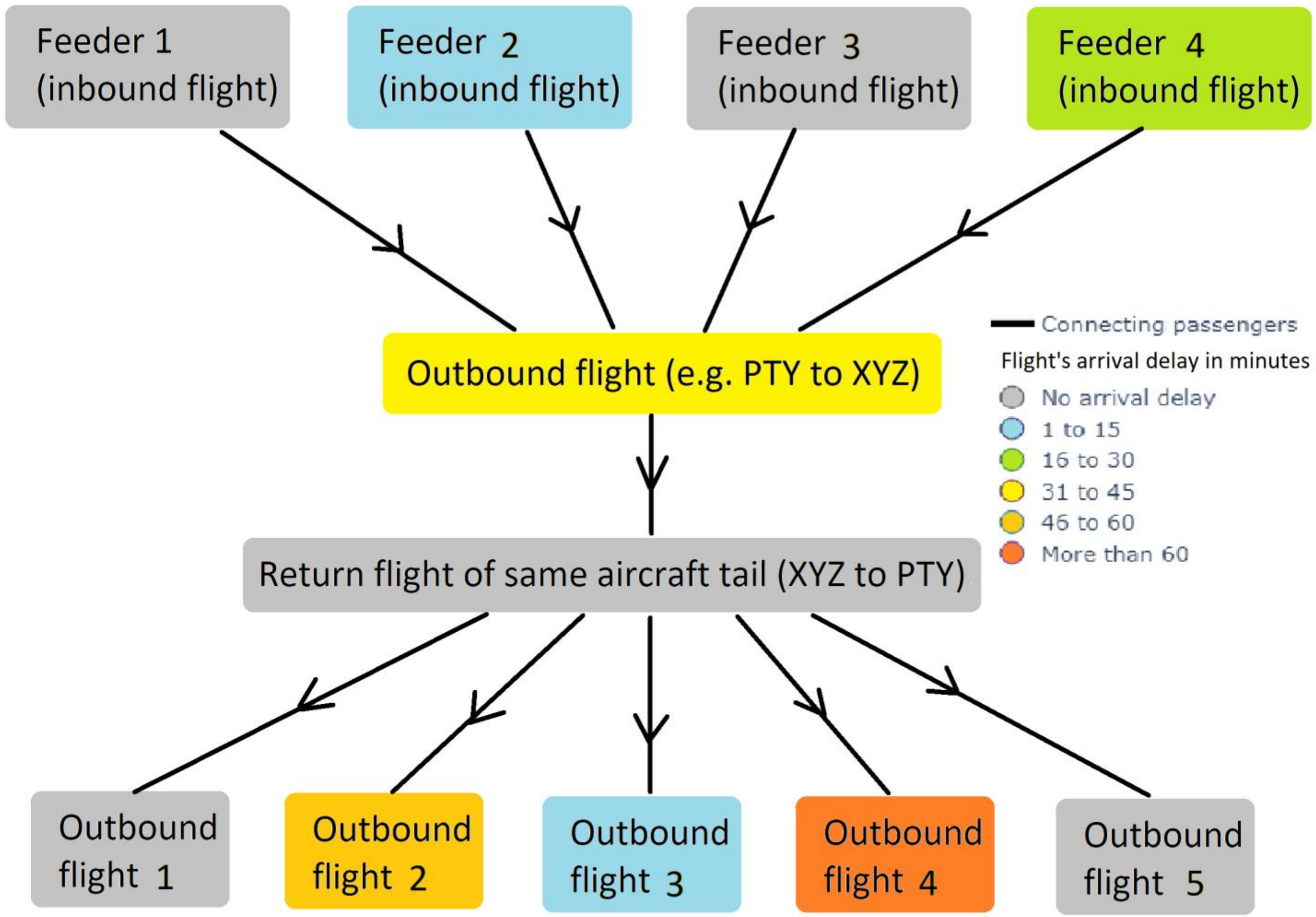


Data Visualization (continued)



- Tool-tips (hover information) for displaying important flight data, when cursor is moved over a flight node.

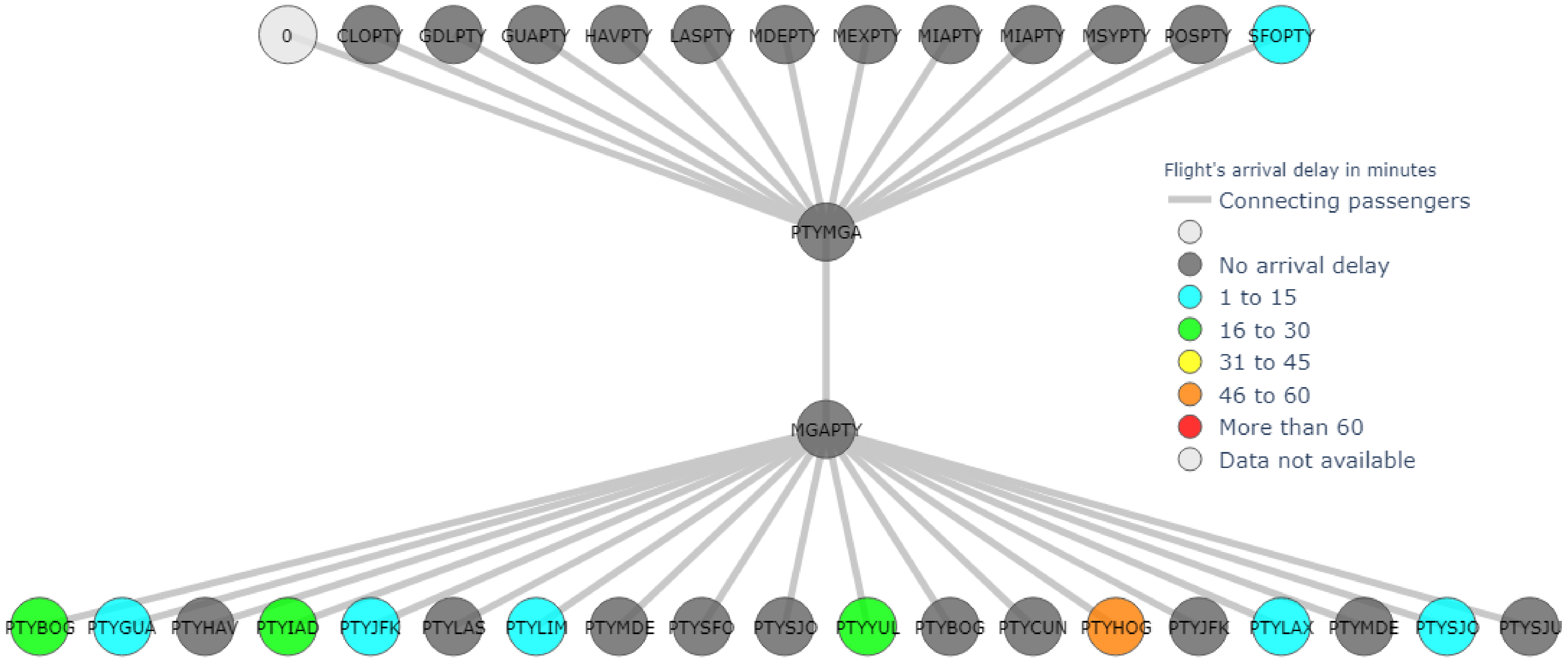
Data Visualization (continued)



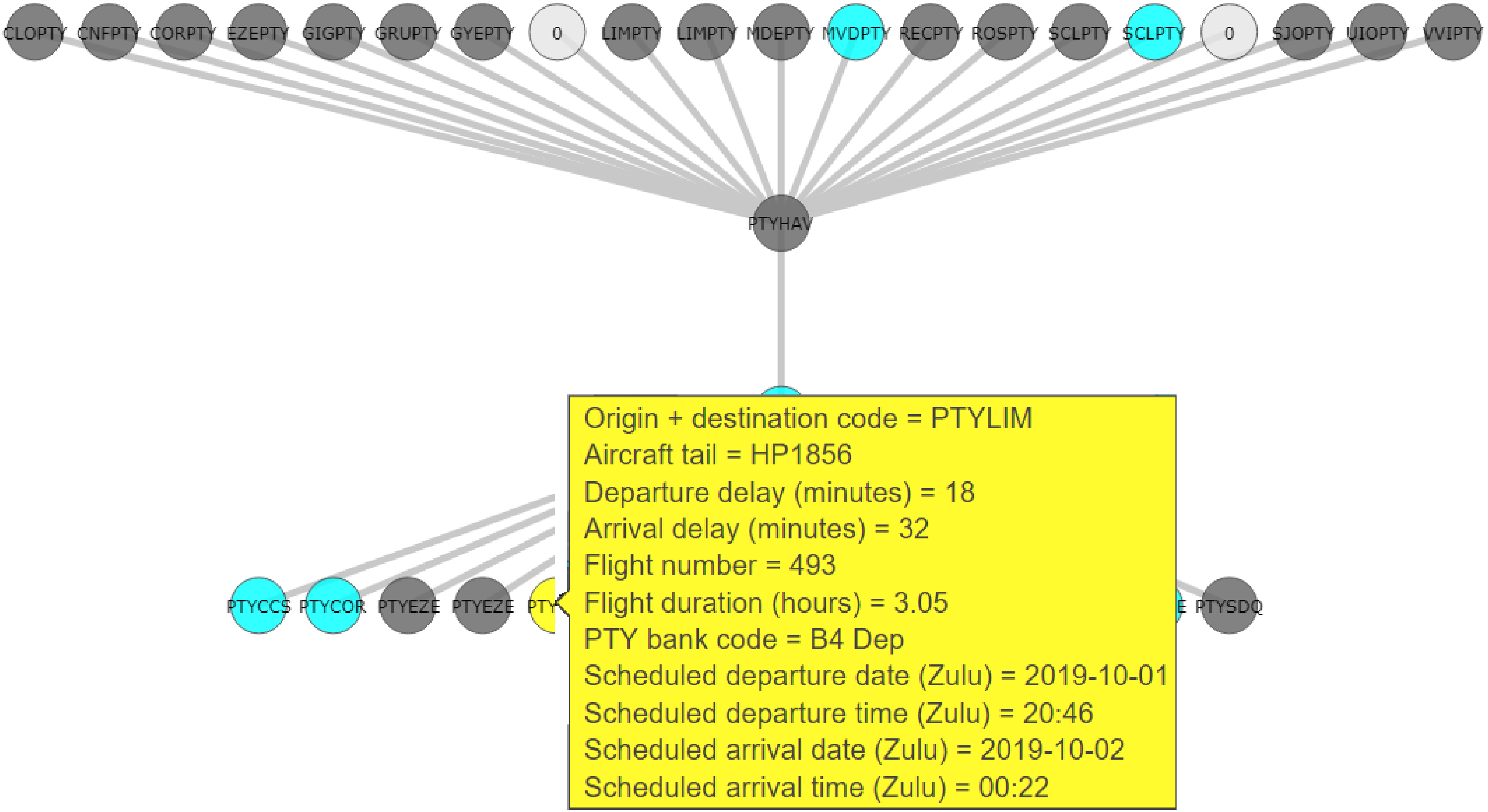
Displaying multiple feeders of a selected outbound flight:

- For an outbound flight, the delay depends on many factors, including all the feeder flights.
- Four horizonal rows or levels of flight nodes:
  - First level (top): all the feeder flights for the chosen outbound flight.
  - Second level: chosen outbound flight.
  - Third level: return (inbound) flight of the same aircraft as outbound flight above.
  - Fourth level (bottom): outbound flights fed by the return (inbound) flight above.
- Passenger transfer from top to bottom.
- Flight nodes colored according to their delay category.

Data Visualization (continued)

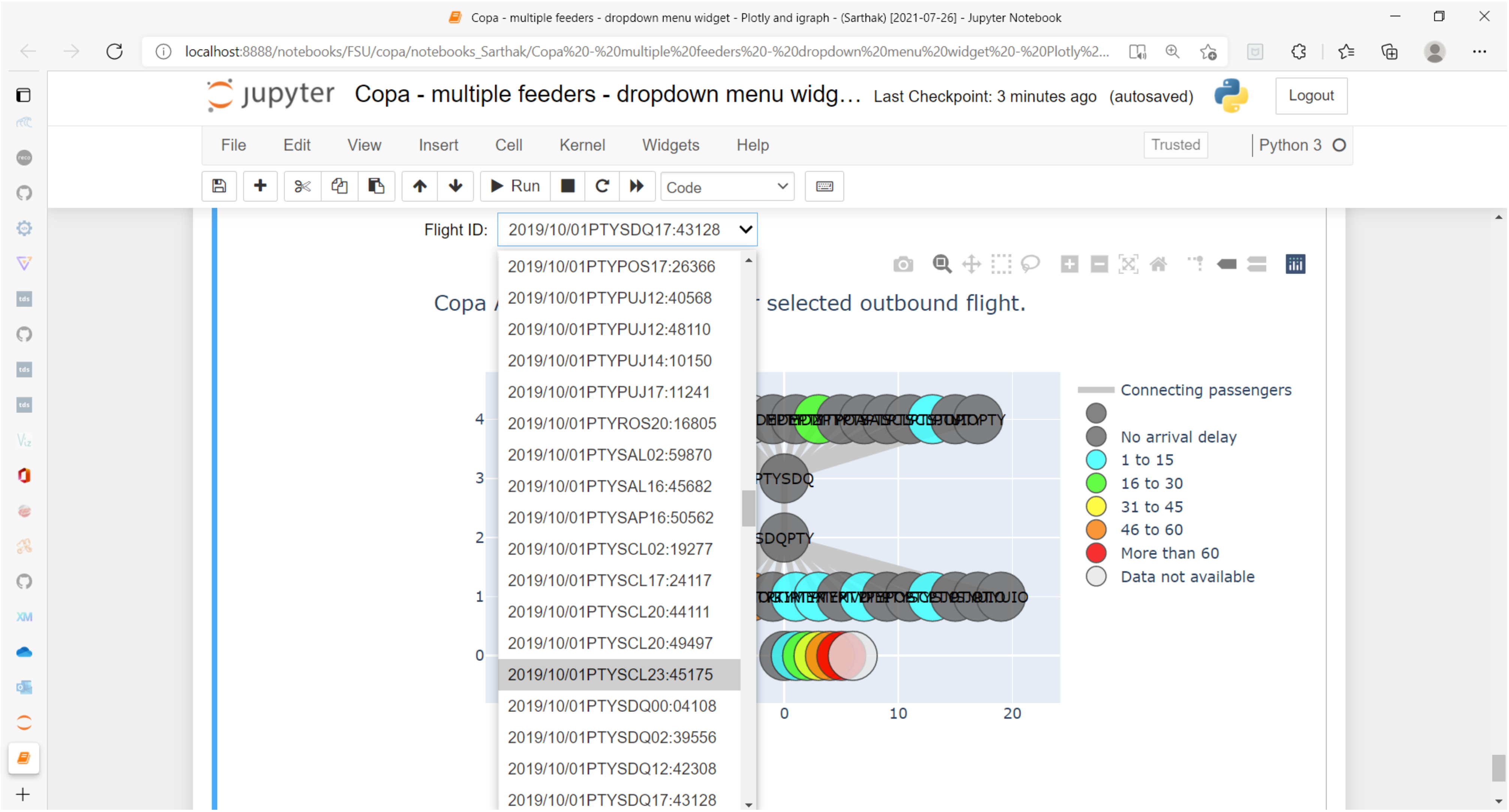


Data Visualization (continued)



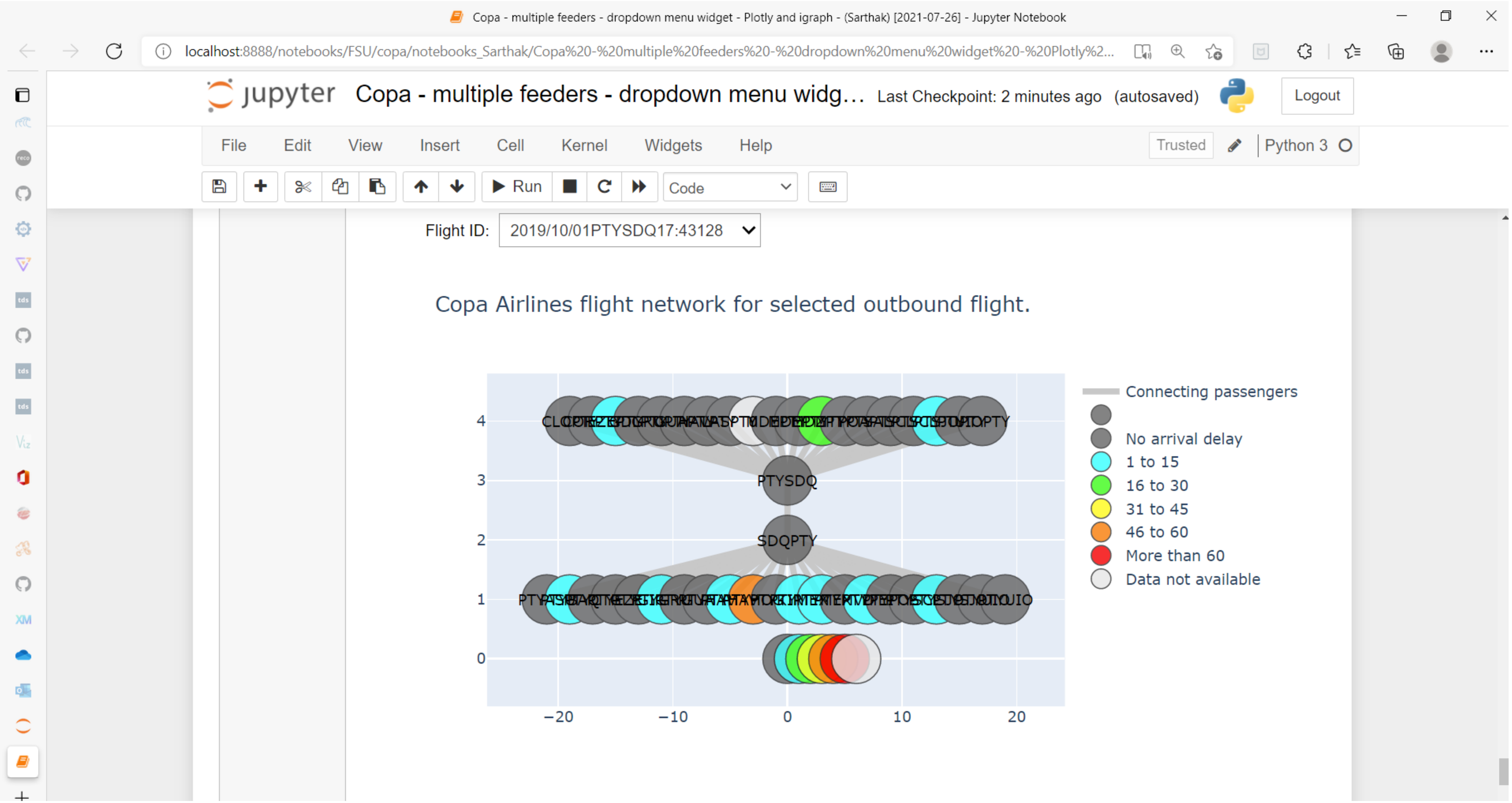
Tool-tips (hover information) for displaying important flight data, when cursor is moved over a flight node.

Data Visualization (continued)

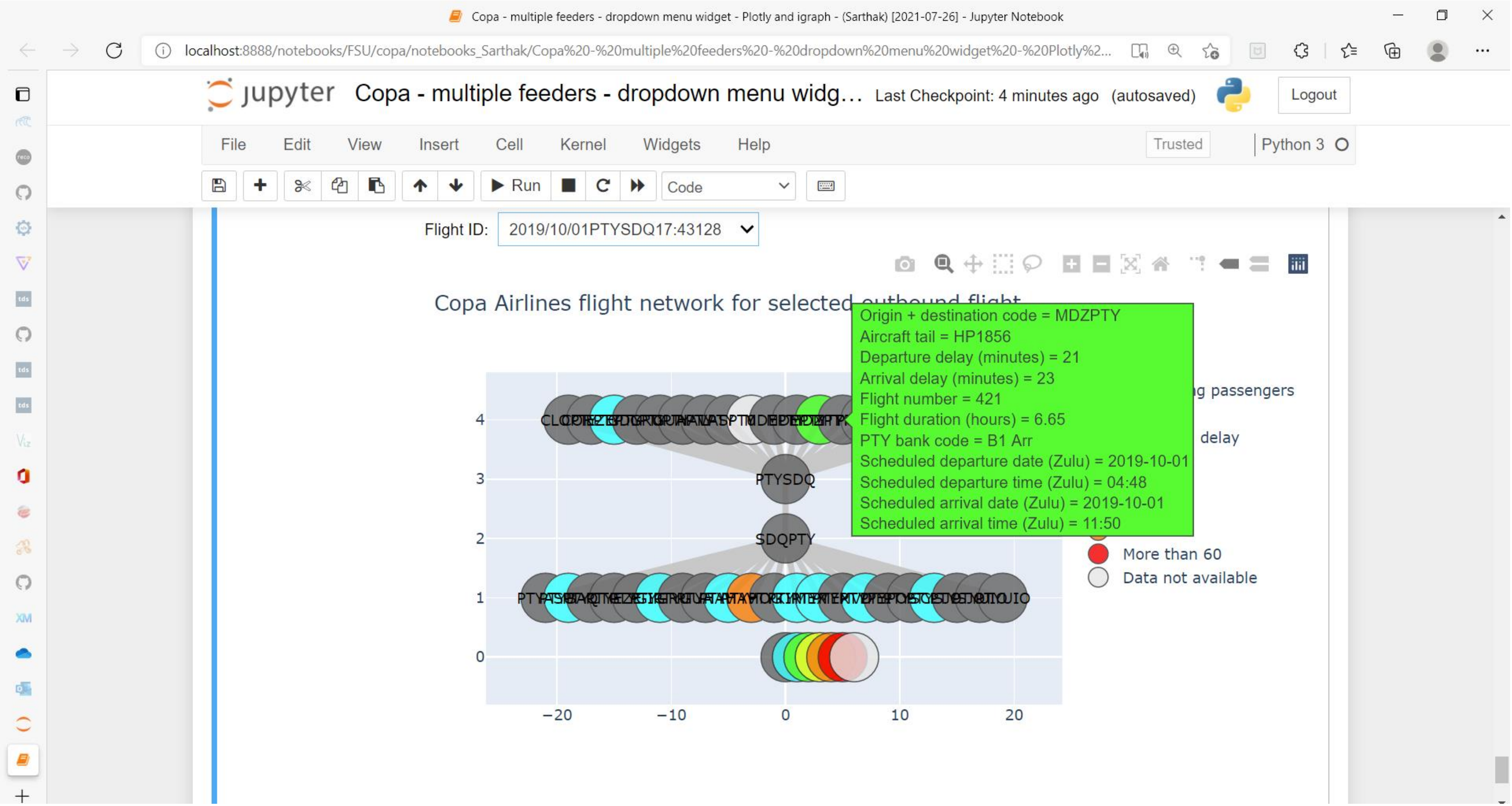


Screen-shot showing drop-down menu (in Jupyter Notebook) for selecting outbound flight of second level.

Data Visualization (continued)



Data Visualization (continued)



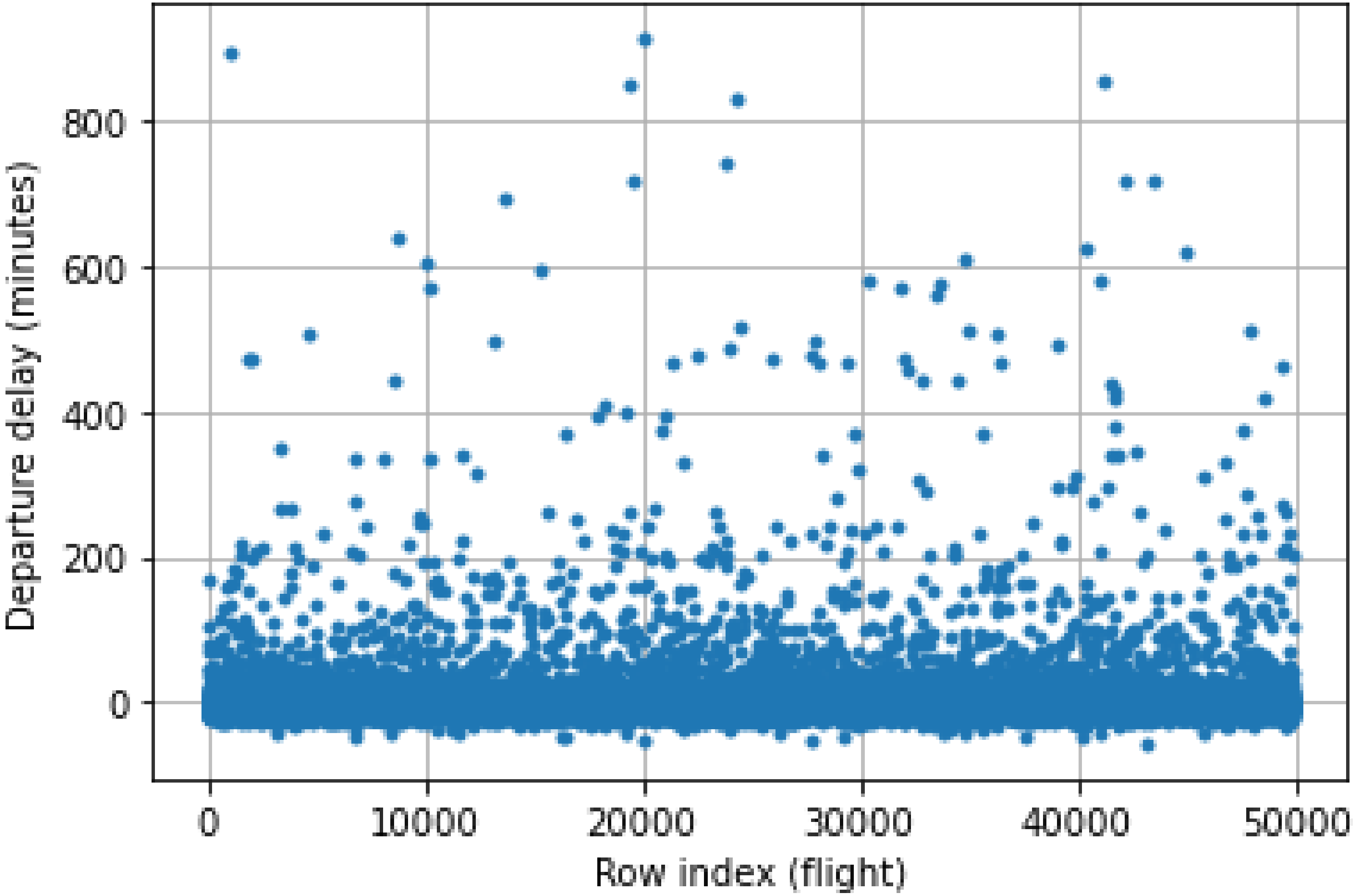
Screen-shot showing tool-tip (hover information) in Jupyter Notebook.

Flight Delay Prediction

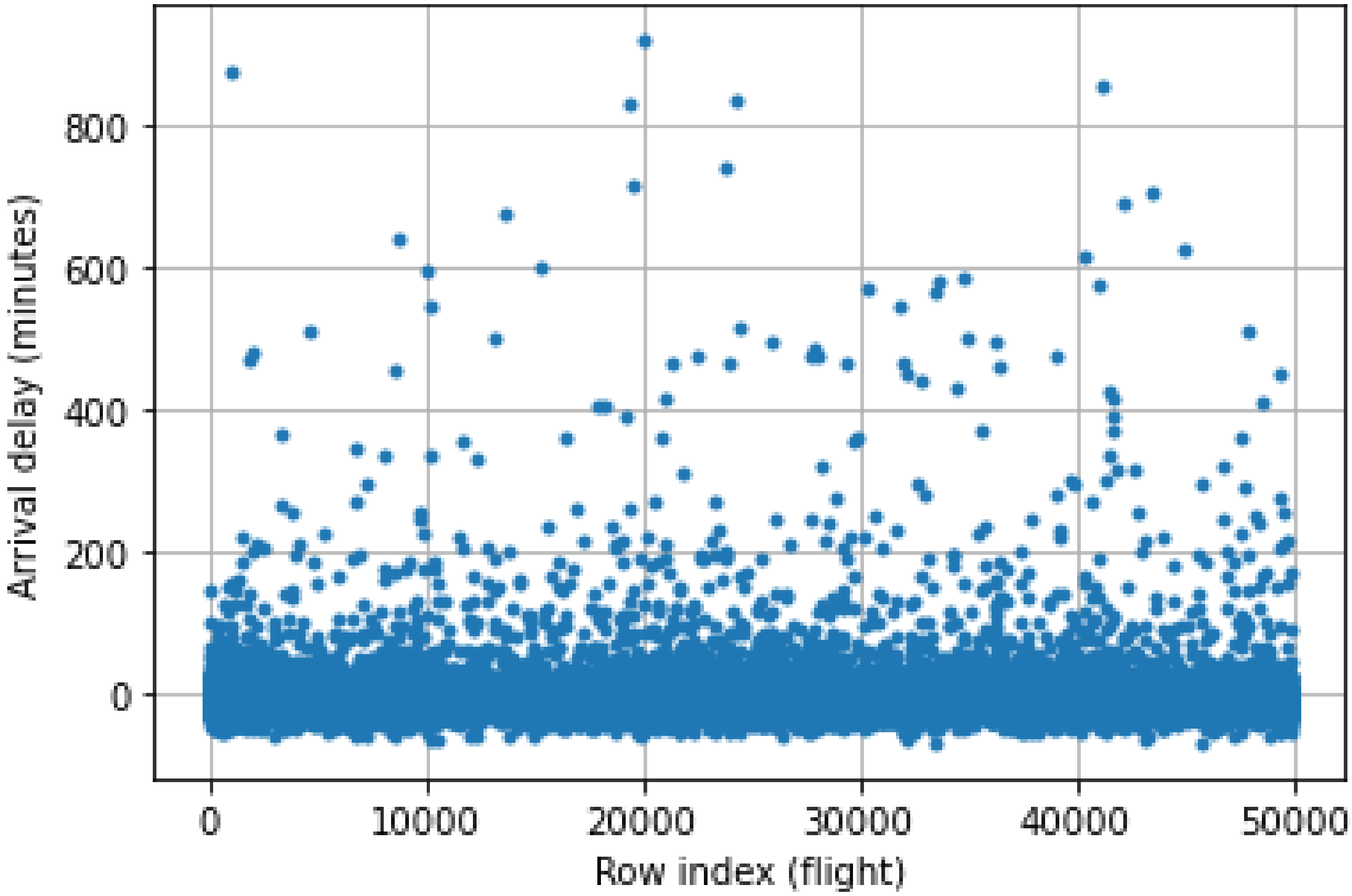
Flight Delay Prediction

Data exploration:

- Scatter plot of **departure delays** of Copa Airlines flights.

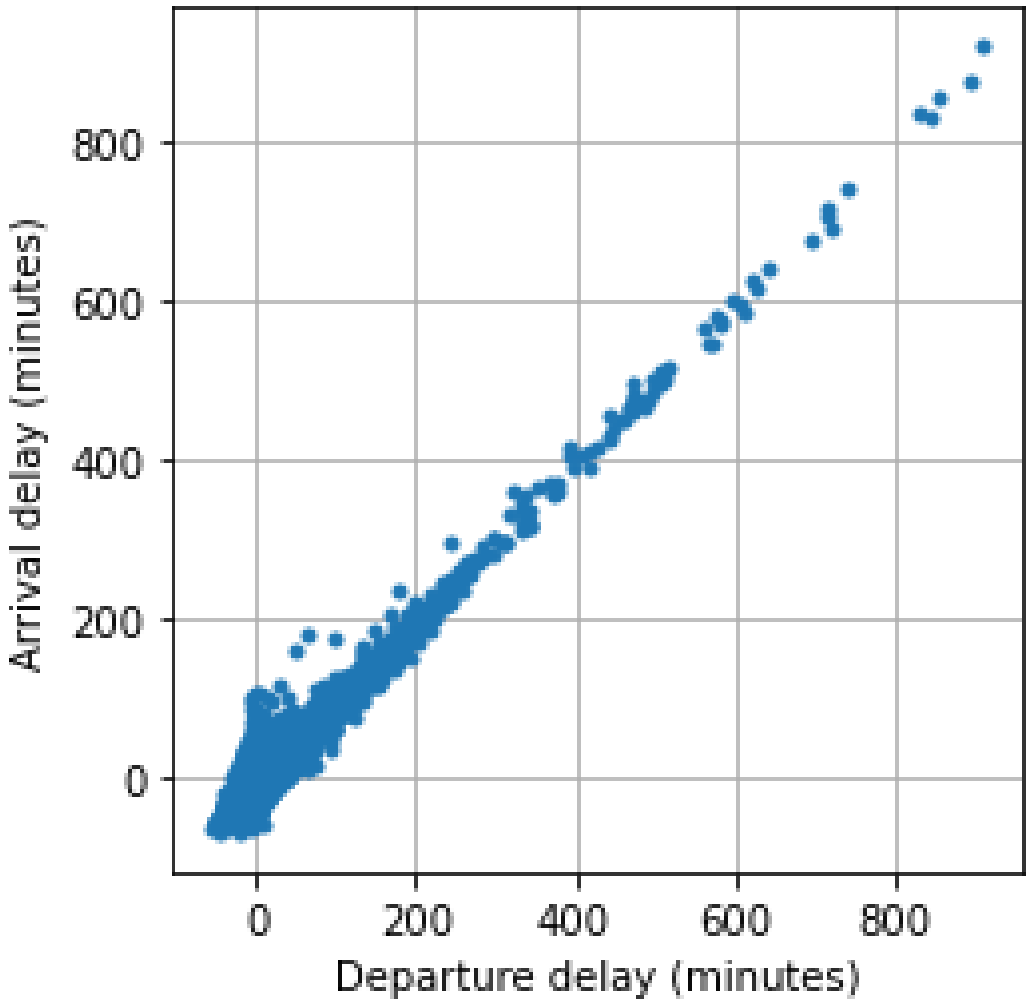


Flight Delay Prediction (continued)



- Scatter plot of **arrival delays** of Copa Airlines flights.
- Most flights are concentrated near zero delay, for both departure and arrival.

Flight Delay Prediction (continued)



- High positive correlation (about 0.9234) between departure delay and arrival delay.
- Expected, as flights that depart late by *n* minutes are likely to arrive late by a similar number of minutes.

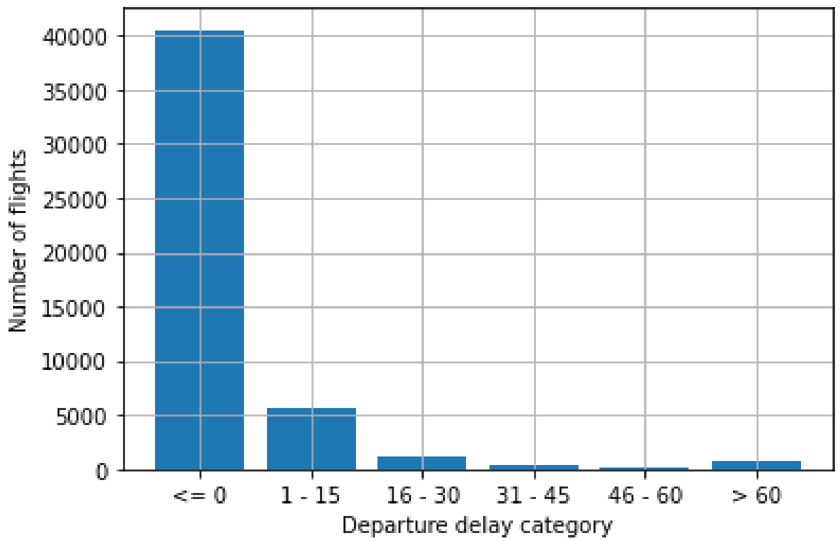
Flight Delay Prediction (continued)

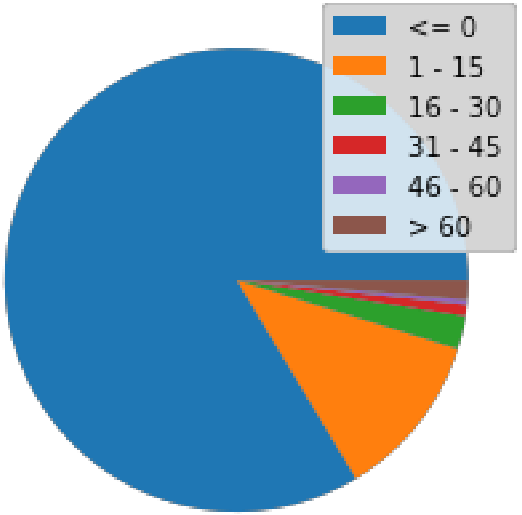
- We can try to predict:
  - number of minutes that the flight is delayed by (regression problem), or
  - delay category in, say, 15-minute intervals (classification problem).
- Distribution of Copa Airlines flights in different delay categories of 15-minute intervals:

Delay category (minutes)	**Fraction of flights in this departure delay category**	**Fraction of flights in this arrival delay category**
≤ 0	83.6 %	75.2 %
1 to 15	11.5 %	18.5 %
16 to 30	2.3 %	3.7 %
31 to 45	0.8 %	0.9 %
46 to 60	0.3 %	0.4 %
≥ 61	1.3 %	1.3 %

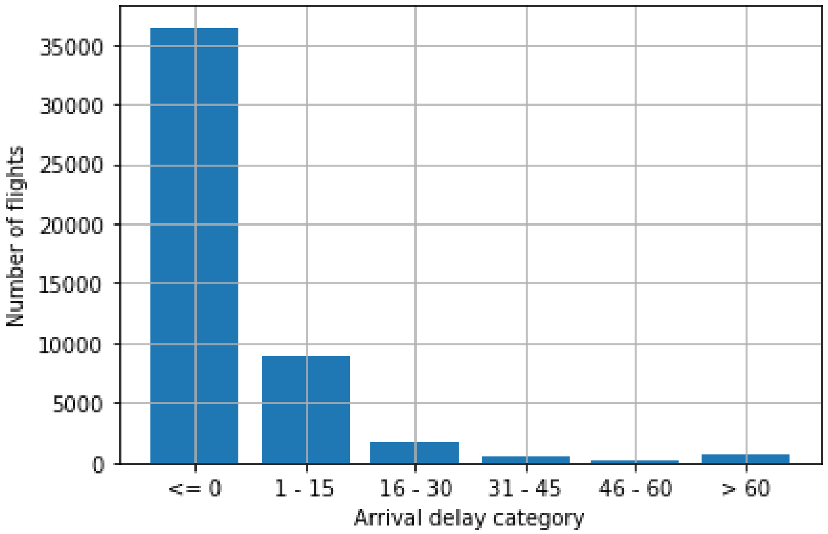
Flight Delay Prediction (continued)

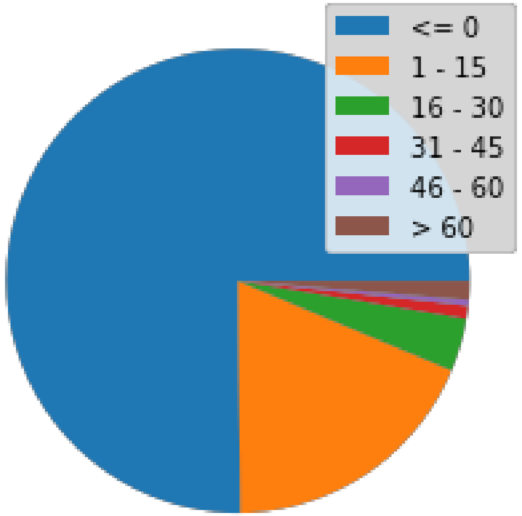
- Departure:





- Arrival:





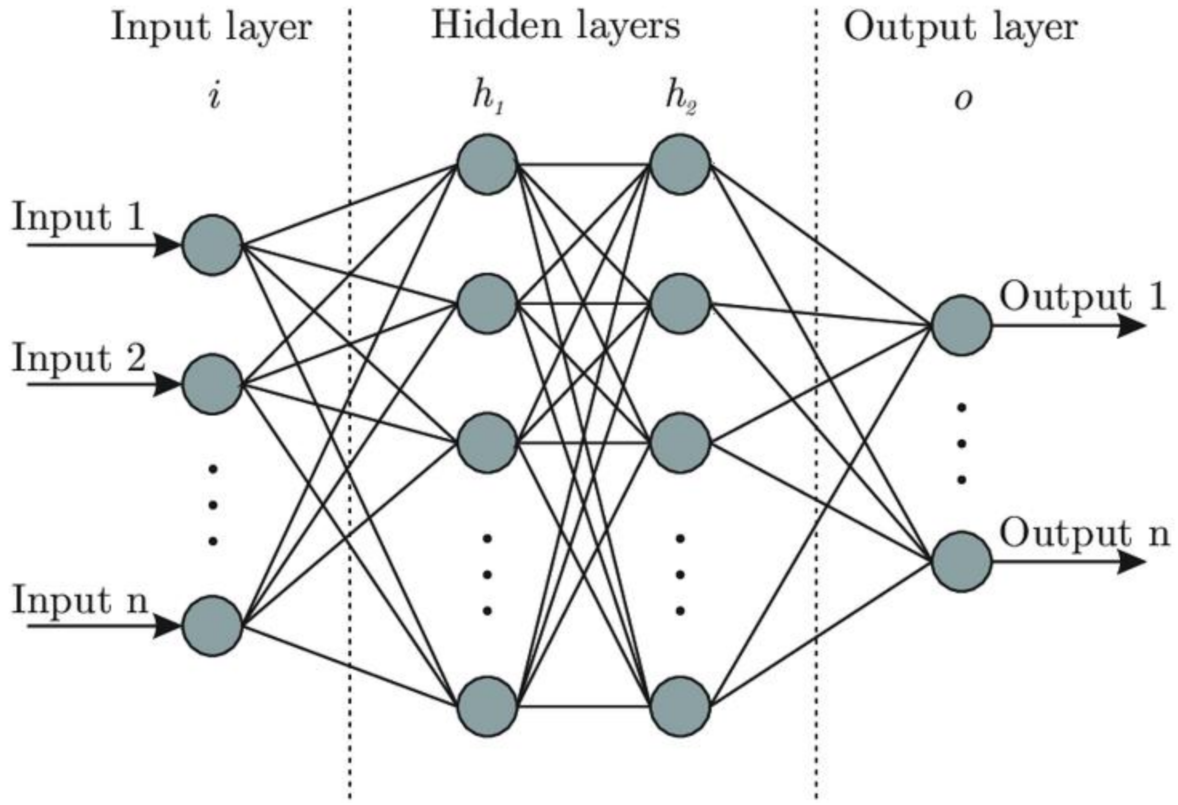
Flight Delay Prediction (continued)

- Simplest flight delay prediction: average departure delay and arrival delay for different routes. This considers variation in space, but not in time.
- For better accuracy: average departure or arrival delay for each route, for different times (e.g., hour of the day, day of the week, month of the year, etc.) this considers variation in both space and time.
- For even better accuracy, we can try to use Machine Learning for predicting flight delays.
- ML model can be trained on past flight data.

Flight Delay Prediction (continued)

- First ML model used: multi-layer perceptron (simple artificial neural network that is fully connected).
- Three hidden layers. Each having 100 artificial neurons.
- Synapses are modeled using linear combinations of weight functions, and activation functions use ReLU (rectified linear unit).
- Number of input points is number of features selected $(n_i = 28)$.
- Number of output points is number of delay categories $(n_o = 6)$.
- Selected features include scheduled times and dates of departure and arrival, origin and destination airports, aircraft tail numbers, flight numbers, available and scheduled rotation times (duration between two consecutive flights using the same aircraft), etc.
- Loss function: cross-entropy loss.
- Trained period: ten epochs, in each type of analysis (predicting categories of departure delays and arrival delays).
- Training and testing are implemented as functions that are called in each epoch.
- Batch size: 100 flights is used.

Flight Delay Prediction (continued)



- Result: “no delay” category is predicted for all flights.
- Reason: imbalanced data set, with majority of flights having “no delay” label.
- Accuracy may seem good initially (equal to the percentage of flights that do not have delay: about 84% for departure, and 75% for arrival).
- Improvement is needed.
- Classification model that always predicts a single class is not much use, even if the predicted class is seen in majority of labels.
- Next suggestion: using **Graph Neural Network** to predict flight delay.

Conclusions and Future Work

Conclusions and Future Work

Conclusions:

- Computational tools to visualize part of Copa Airline’s flight network are developed.
- We are also attempting to use Machine Learning to predict flight delays in departure and arrival.
- Flight network visualization and flight delay prediction can be valuable tools to help in the flight network management and flight scheduling, aiding flight planners in deciding how to allocate their limited resources, and anticipating future delays or other challenges
- This analysis may also be extended to other types of networks, e.g., road traffic networks, data transfer networks, etc.

Conclusions and Future Work (continued)

Future work:

- Using Artificial Intelligence to automate some or all aspects of the flight planning / scheduling process, to have better flight management and better overall passenger experience.
- As of now, if an outbound Copa flight's inbound feeder flights have delayed arrivals at PTY airport, then Copa Airlines staff manually decide whether to postpone the departure of that outbound flight (to prevent the delayed inbound passengers from missing that outbound flight).
- Conflicting interests like cost reduction, on-time performance, and customer satisfaction may have to be balanced.
- It may help to automate these decisions in a data-driven and AI-enabled way.

Thank you!