The Route Network Development Problem based on QSI Models

Assia Kamal Idrissi, Arnaud Malapert, Rémi Jolin

The Route Network Development Problem based on QSI Models

Assia Kamal Idrissi[1], Arnaud Malapert[2] and Rémi Jolin[1]
[1]*Milanamos, 1047 route des Dolines, Sophia Antipolis, France*
[2]*Université Côte d'Azur, CNRS, I3S, France*
{*assia.elafouani, remi.jolin*}*@milanamos.com, arnaud.malapert@unice.fr*

Keywords: Airline Schedule Design, Route Network Development, Forecasting demand, Quality of Service Index, Time-independent Model, Graph Database.

Abstract: The growth of air passenger needs has forced airlines to improve their quality of service. Airlines have to choose flight schedules by considering demand, passengers preferences and competitors. The problem of allocating a new flight involves the route network development, and consists to determine a set of (Origin-Destination) pairs to serve and then choose flight schedules with respect to the Quality of Service Index (QSI) model. In this PhD project, we work with a software tool developed by the company *Milanamos* that helps airline managers to make decisions about destinations to serve. As a starting point, we define the flight radius problem related to this software. It is a sub-problem of the route network development problem and aims to optimize the visualization of the pertinent network by showing only interesting airports regarding QSI model. In this paper, we present the problem of allocating a new flight and formulate the flight radius problem as a problem of finding maximal sub-graph. Our objective is to locate in the network what routes represent business opportunities and are attractive regarding competition so it can be visualized. We construct the graph from *Milanamos* the database using the time-independent approach and store it in `Neo4j` a graph database. We describe the process of generating and storing the graph in `Neo4j` and sum up by outlining the expected outcome.

1 RESEARCH PROBLEM

The growth of air passenger needs has forced airlines to improve their quality of service. At best, the airlines should offer flights that match the expectations of their passengers. In order to capture a large flow of the passengers, the airlines should focus on airline schedule design which is one of the important components and evokes complex decisions. It takes the airline passenger demand, airport and aircraft characteristics and then generates a selection of flight legs as outputs those maximize the airline company profit subject to resource constraints (aircraft and airport capacity, maximal working hours, minimal ground time,...). A flight leg or segment is defined with three attributes: Origin-Destination (OD) pair (an OD pair is a couple of airports), arrival/departure time and aircraft type (Hall, 2012). Airline schedule design aims to answer the following questions:

- Where to fly?
- How frequently to fly?
- When to fly?
- How much capacity to provide on each flight?
- What are the competitive choices of flight schedules?

It's a process with making decisions at different stages about opening new routes or adding new flights which need a demand forecasting that is not always based on historical data when airlines decide to include a new destination. Therefore, demand is calculated and the best decision is chosen to maximize their profit. These decisions are very important for an airline, quality of service and prices influence the airline's ability to attract travelers. However, we deal with a multi-objective problem. The airline company has to take into consideration the passengers choice. In fact, it may consider criteria such as travel costs, travel time and also the type of flight. For example, a businessman may try to optimize his travel time, a student wants to minimize his costs and a visitor may wish to avoid connecting flight. So, at the end, users have different preferences over the criteria and type of flights is one of these criteria. There are three different kinds of flights: non-stop flight, direct flight and connecting flight. A *non-stop flight* is a single flight with no intermediate stops. It is the preferred choice for most passengers. In the absence of such flights, passengers must take either a direct flight or

Idrissi, A., Malapert, A. and Jolin, R.
The Route Network Development Problem based on QSI Models.
In *Doctoral Consortium (DCORES 2017)*, pages 3-11

a connecting flight. A *direct flight* or *through flight* is operated by the same aircraft and includes at least one intermediate stop; passengers stay on board during the trip. Note that the flight number remains the same throughout the trip. A *connecting flight* requires an aircraft change for passengers in a *hub*. Thus, the trip includes at least two different planes with two different flight numbers. A *Trip* is a sequence of flight legs taken by passengers to complete a journey.

The problem of allocating a new flight concerns the first two questions of airline schedule design, that is, determining a set of OD pairs and then choosing the arrival and departure times for an aircraft, given certain constraints, that minimize costs and maximize the profit of an airline company.

Adding a new flight leg to the current network is complicated and involves several decisions:

- Scheduling decisions must be made according to all flight legs connected with the new addition flight leg. They must decide which flight leg to add after considering competitors.
- Measuring Route 'Profitability': determines economic profitability of opening a new flight, if it involves a new destination. The other costs must be considered including the additional cost of the airport and also calculate the prices in order not to lose their passengers but to capture a new demand. That depends on the existence of current routes that could be connecting flight as well as expected future competition.

In this work, we aim to determine a set of (OD) pairs for allocating new flights. This problem evokes the selection of routes to be flown but some operational and economic considerations must be taken to optimize an airline network. A *route* is a sequence of flights with unique flight numbers that begins at the origin airport and ends at the destination airport. Therefore, a forecasting demand is required to estimate passenger demand for each route and then determine the expected cost and finally compute flight time and revenue. Forecasting demand is the key element while an airline is planning to add a new flight. An airline needs to estimate the total number of passengers who are willing to take this route, especially if the route is already operated by other airlines. It has to supply enough seats to satisfy the demand. On the one hand, we deal with retrieving set of origin-destination pairs. On the other hand, we look for the best itineraries based on three principal criteria of ***QSI model***s (*Quality of Service Index*) which let airlines project potential market share impact on each decision. *QSI model* is a market share model that is used to estimate the probability that a traveler selects a specific itinerary connecting an airport pair (Jacobs et al., 2012). Two **KPIs** (*Key Performance Indicator*) are considered in potential market share impact: *number of passengers* and *revenue* for the airline company, that is, adding a new flight by allocating capacities to maximize the revenue.

1.1 State of the Art

The air transportation industry has evolved rapidly over the last years. Route network development, schedule design, fleet assignment, aircraft maintenance routing and crew scheduling that represent the five facets of the air transportation optimization problem (Rebetanety, 2006):

Route Network Development: deciding which set of origin-destination pairs to serve. The network design problem consists of determining where to fly (Belobaba et al., 2015).

Schedule design : defining the frequency of each flight. Scheduling determines where and when the airline will fly.

Fleet Assignment: specifying the type and the size of aircraft serving each flight in a given schedule (Rebetanety, 2006).

Aircraft Routing: determining feasible aircraft routes, sequences of flight legs flown by an aircraft type under maintenance and time constraints (Jacobs et al., 2012).

Crew Scheduling: assigning crews to the flights.

Airlines have the choice to create new routes or increase/decrease the frequency of exiting routes with respect to operational and economic constraints. The latter does not require the route network development since the route already exists. Note that creating a new route requires a lot of investment (Carmona Benitez, 2012). Relevant literature exists for airline scheduling and routing. Most researches deal with minimization of the airline cost. (Dobson and Lederer, 1993) studied profit maximization with respect to quality of service and modeled flight schedules including company costs and consumer choice. The objective was to maximize profit against fixed schedules and prices for other airlines, demand was calculated for each route as a function of the service quality of all routes to attract passengers. They used a heuristic algorithm to calculate optimal schedules and prices with two classes of customers business and non-business and solved a sample problem with 12-period and 5-city which gives 120 flights and 20 (OD). The competition between airlines has increased. In order to deal with competition, airlines have to increase their market share. In this regard, QSI criteria are integrated in our model to get potential market share impact for an airline. Figure 1 shows

the criteria used in the QSI model:

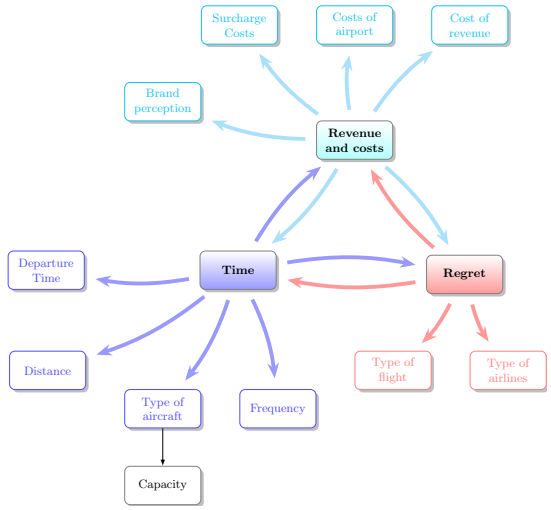

Figure 1: Criteria of QSI models.

The three principal criteria are highly correlated: ***Revenue and costs*** consists to fix an attractive price to capture more traffic. The price must be determined in comparison with other flights in respect of the competition. ***Time*** is the elapsed flying time and finally ***Regret*** is the regret compared to the optimal duration and cost. Each of these criteria is related to other factors that affect the estimation of potential market share: *Brand perception* is a factor included in the category *revenue and cost*s because the airline market share of total forecast demand for the new route depends on the existence of current and expected future competition. In addition, the Brand perception of travel costs criteria fixes the price for routes against competing airlines. Surcharge costs include maintenance costs, fuel and others stuff. When the route involves a new destination, additional costs of the airport must be considered. Cost of revenue concerns marketing and advertising. Moreover, time criteria concerns both airline managers who must decide which flight to schedule and passengers. Regarding revenue and costs, the focus is solely on the airline company. In fact, airline costs are driven by fleet and flight schedule. Finally, the regret criterion focuses on avoiding regrets that may result from making a non-optimal decision of time or price when choosing a low-cost airline or a connecting flight. The passengers do not have the same regret of the same route. Regarding competition, departure time can be determined and then frequency. In fact, passengers have preferences for routes. These preferences depend on the QSI criteria typically the departure time desired. The type of flight is either a principal criterion. In fact, airlines may choose to serve the route with non-stop flights rather than going through their hub especially, when they see a high demand for travel on this route.

The route network development is considered as the initial problem addressed by the airlines, our work position in this research field. Airlines choose what routes to serve and what prices to fix so that the passengers have the quality and the safety services, the first issue to consider is forecasting demand for potential new routes. Three common methods for forecasting air travel: trends, gravity models, and simulation (Swan, 2008). Since no single model guarantee accuracy, the most commonly used in literature is the gravity model. (Sivrikaya, 2013) studied this model to estimate the domestic air travel demand for any city pair, two levels of forecasts were considered: 1. Microscopic model; airport specific or city pair specific data. 2. Macroscopic model: region or country. The aim of the study is to analyze the determining factors in air travel demand, it is a semi-log linear model based on geographic, demographic and socioeconomic variables such as population, GDP, distance, travel Time. In addition, (Marwaha and Kokkolaras, 2015) used the gravity model to estimate OD demand figures for Canada but in function of routes distances.

Modeling a flight network is similar to a railway network which is smaller but more complicated than road networks (junctions are nodes, streets are edges) due to different operations constraints which rely on some kind of periodic timetable. *Timetable* is a listing of times at which events are intended to take place and are the basis behind any flight models. The problem can be solved by modeling the network as a simple graph where edge weights represent travel times on the corresponding connection. In order to model the flight network and thus obtain a graph, there are many approaches cited in the literature for route planning. The time-expanded model includes the time dependencies of the timetable in the graph where each node represents an event of the timetable and edge connects two consecutive events (Kirchler, 2013). This approach allows modeling the time-dependent information with more flexible modeling of additional constraints. Therefore, this approach yields a huge graph. The condensed model is a time-independent model where edges correspond to the aggregate of all available connections between two nodes. While the first approach allows a more flexible model but constructs a big graph, the latter uses only smaller inputs. A key point is that the output graph contains a single node per station rather than multiple nodes per station. Thus, instead of applying routing algorithms to the whole graph, we opt to choose the condensed model in such a way that a routing algorithm does not nec-

essarily explore some useless parts of the graph.
Once the graph is generated, routing problems are addressed by finding itineraries that are satisfying time constraints. The routing problem is modeled in literature in different ways. As an example, it was modeled to find an optimal flight path that avoids geographical obstacles (Bast et al., 2015). Another work was dealing with routing problems and researches were interested to find the shortest path in the flight networks during a time period with respect to several criteria (transfer, time...). As the shortest path algorithm, (Delling et al., 2009) used Dijkstra algorithm to solve the earliest arrival problem. It retrieves the quickest connections according to multiple criteria. Although, the railway model can be adapted to flight timetables, they developed a new model that takes into account check in and check out time and transfer time when it's a connecting flight; it was due to the fact that the problem of instantaneous transfers has an even greater impact on realism flights than on railway queries. These researches studied the routing problem related to passengers.

1.2 Our Contribution

In this paper, we aim to solve the problem of allocating a new flight. It is represented as the route network development problem. We work with *Milanamos*, a startup company specialized in air transportation. *Milanamos* has developed a decision tool for airline managers to analyze and simulate a new market. Our problem derived from this application and targets new destinations for a given (OD). We proposed this subproblem (Flight radius problem) which is related to the route network development problem. It helps to enhance the visualization of the application by showing only interesting destinations. Therefore, it can be implemented in the short term within the existing application and helps airline managers to make decision about where to fly. The data are stored in a `NoSQL` database. The first challenge is to generate the graph from a missing and erroneous data since the real-world data are generally incorrect inputted. Hence, a data pre-processing step is required to filter our data. The result of the selection of routes will be implemented to optimize the *Milanamos* application (See section 3.3). We use the time independent approach to model the flight network. The selection of routes is based on QSI criteria to be competitive against other airlines serving same (OD) pair. Thus, the condensed graph constructed includes these criteria. Once it's done, we would like to apply the routing algorithms to enumerate itineraries under certain constraints.

This paper is organized as follows. Section 2 describes objectives of this project. In Section 3, we explain the methodology and include formal definitions of timetables, present the formalization of the problem discussed above and describe the database information used in our research as well as the difficulties encountered when industry data are missing. We conclude with the current research of our problem.

2 OUTLINE OF OBJECTIVES

The long-term goal of this PhD project is the optimization of multimodal networks. Our first step is to enhance the visualization in `PlanetOptim`. Hence, for an airline managers, what is the relevant subnetwork related to a given flight? What are the passengers origins and destinations? This represents a preliminary step before studying the allocation of new flights, that is determining a set of (OD) pairs and the arrival and departure times with respect to the QSI criteria. The project is organized into five major steps. The first two ones have already been performed and their outcome is described in this document:

- Generation of the condensed graph and store it in a graph database;
- Formulation of the flight radius problem for enhancing the visualization;
- Solving the flight radius problem in `PlanetOptim`;
- Modeling and solving of the route network development problem with respect to QSI criteria;
- Integration other transportation modes in a multimodal network.

3 METHODOLOGY

This section presents the formalization of the flight radius problem, describes the flight database, and explain the construction of the condensed graph based on the flight database. Network design and routing problems often rely on graph theory. Therefore, we first recall basic definitions of graphs.

3.1 Preliminaries

A *graph* G is a tuple $G = (V,E)$ consisting of a finite set V of nodes or vertices and a set $E \subseteq V \times V$ of edges which are ordered pairs (u,v) if the graph is directed. The node u is called the *tail* of the edge, and v is called the *head*. Each edge $(u,v) \in E$ has an associated non-negative weight $w(u,v)$. In a directed graph,

the edges point from one node to another. For instance, airline networks are weighted directed graphs where the weights represent the prices or the duration of the flight. A direct flight from one city to another does not necessarily imply that there is also a direct return flight. A *sub-graph* $G' = (V', E')$ of a graph G where V' is a subset of V and E' is a subset of E. A *path* is a sequence of nodes $\{v_1, v_2, ..., v_k\}$ such that for each $1 \leq i < k$ the condition $(v_i, v_{i+1}) \in E$ holds. If additionally $v_1 = v_k$, then the path is a *cycle*. The length of a path is the sum of its edge weights along the path and is denoted by:

$$\delta(P) := \sum_{i=1}^{k-1} w(v_i, v_{i+1}).$$

A path in G is called *elementary* if no vertex occurs more than once. A graph G is *strongly connected* if there exists a path joining any two vertices. A transportation network should be a strongly connected graph.

3.2 Problem Formalization

The essence for each flight model is a timetable from which we construct the condensed graph. A *flight timetable* is defined by a 4-tuple $(\mathcal{C}, \mathcal{A}, \mathcal{F}, \mathcal{T})$ where $\mathcal{A}$ a set of airports, $\mathcal{F}$ is a set of flights, $\mathcal{T}$ is the periodicity of the timetable and $\mathcal{C}$ is a set of elementary connections. An *elementary connection* $c \in \mathcal{C}$ is a 5-tuple $c = (f, o, d, t_s, t_e)$ which represents *flight* $f \in \mathcal{F}$ departing from airport $o \in \mathcal{A}$ at $t_s < \mathcal{T}$ and arriving at airport $d \in \mathcal{A}$ at time $t_e < \mathcal{T}$. Concretely, an elementary connection corresponds to an event in a timetable. Let $cap(c)$ denote the capacity, let $pax(c)$ denote the number of passengers, and let $r(c)$ denote the total revenue. Let $t(c) = t_e - t_s$ be the flight duration associated with the elementary connection.
A *passenger trip* $(c_1, c_2, \ldots, c_{n-1}, c_n)$ is a sequence of elementary connections, with the origin of an elementary connection the same as the destination of its predecessor in the sequence, and the elapsed time between two successive connections at least as great as the minimum connecting time:

$$o(c_{i+1}) = d(c_i) \wedge t_e(c_i) + MCT(d) \leq t_s(c_{i+1}) \quad \forall 1 \leq i \leq n-1$$

Where *MCT* is the minimum connecting time at the destination airport d.
The condensed graph is a time-independent representation of the flight network. Nodes represent airports meanwhile the presence of an arc indicates that there exists at least one elementary connection between the two airports. Each arc is constructed by aggregating all elementary connections between each pair of airports. Let $\mathcal{C}_{od} = \{c \in \mathcal{C} \mid o = o(c) \wedge d = d(c)\}$ be the set of elementary connections between two airports o & d. The following labels are associated with the arc (o, d) :

- $F_{od} = |\mathcal{C}_{od}|$ is the number of elementary connection between o and d;
- $C_{od} = \sum_{c \in \mathcal{C}_{od}} cap(c)$ is the total capacity in terms of the number of passengers;
- $P_{od} = \sum_{c \in \mathcal{C}_{od}} pax(c)$ is the total number of passengers;
- $R_{od} = \sum_{c \in \mathcal{C}_{od}} r(c)$ is the total revenue;
- $\bar{R}_{od} = \min_{c \in \mathcal{C}_{od}} \frac{r(c)}{pax(c)}$ is the minimum revenue per passenger;
- $T_{od} = \min_{c \in \mathcal{C}_{od}} t(c)$ is the minimum flight duration;
- D_{od} is the distance between the two airports.

Frequency, capacity and number of passengers are the target market that determines for airlines what routes to operate. We choose the sum aggregation for these criteria since it indicates the importance of the route so the airline decides to increase the frequency or open new route. Concretely, a *passenger trip* is a unique path p in the time-expanded graph whereas many passengers trips are associated with the same path in the condensed graph. The existence of a path in the condensed graph is a necessary (but not sufficient) condition to the existence of a passenger trip. In addition, a path between o & d in the condensed graph gives a lower bound on the cost and duration of a passenger trip that goes along the same airports.

We aim to determine the set of (OD) pairs that would be interesting for an airline manager. The problem is to identify flights and routes that represent business opportunities and are attractive regarding competition. But the flight network is so large that it can't be visualized. So, we aim to display only the relevant airports with respect to the edge (o, d). It means that there exists a route connecting these airports passing through the arc (o, d) subject to time and cost constraints. A naive algorithm could enumerate all paths passing through the arc, but the number of paths can grow exponentially. With this aim, let R be a Boolean regret function defined on paths of the condensed graph G. The problem consists in finding a maximal sub-graph such that each node or arc supports a path accepted by the regret function.

Hence, the problem is formulated as follows:

Input: a graph $G = (V, E)$, the arc (o, d), the regret function R

Output: a subset $E' \subseteq E$ such that $G' = (V', E')$ is a sub-graph of G and that each node supports a path accepted by the regret function.

The function is true if there exists a path between v_1 &

v_n that passes by (o,d) and is shorter than the shortest path between these nodes with a constant factor and false otherwise. If such path exists then all vertices of this one are added to the sub-graph. This function depends on the shortest path between v_1 and v_n in terms of duration or cost and it's defined for each criterion: cost and time (duration). We fix a lower bound for each criterion with a minimum stopover desired. Let's o & d represent two airports in the graph G where the node o is the tail of the edge (o,d) and d the head. The problem consists in finding an interesting path for a given vertex o_1 regarding the edge (o,d). That is finding at least one path from node o_1 to node d_1 that passes through the edge (o,d) such that is shorter than the shortest path p^* between o_1 and d_1 plus a constant factor K. See Figure 2: $\exists$ a path $p = \{o_1, o_2 ... o_k, o, d, d_1\}$ where:

$$\delta(p) \leq \delta^*(p^*) + K \quad (1)$$

Then $o_i \in V'$, $\forall o_i \in P \setminus \{o,d\}$ since {o,d} are already in V'.

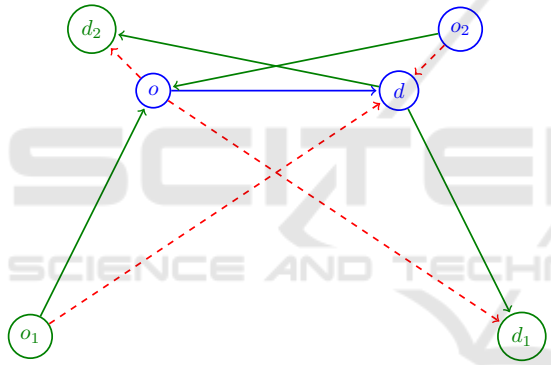

Figure 2: The flight radius problem.

Figure 2 explains the flight radius problem. The blue edge is the (o,d) connection and $\{o_1, d_1, o_2, d_2\}$ represent the set of candidate vertices meanwhile the red dashed edge is the shortest path. The nodes o_2 and d_2 are not going since there exists a shortest path between o and d_2 in the graph (respectively from o_2 to d).

Concretely, such a path that satisfies the regret function exists if and only if exists a path reaching o from o_1 or departing from d to d_1 acceptable to the function $R(p)$.

Proof. Let $p = \{o_1, \ldots, o, d, d_1\}$ be the path from o_1 to d_1 that passes through the edge (o,d) and verify inequality 1. Then: $\delta(p) \leq \delta(p*) + K$
$\Rightarrow \delta(\{o_1, \ldots, o, d, d_1\}) \leq \delta(p*) + K$
$\Rightarrow \delta(\{o_1, \ldots, o, d\} \cup (d, d_1)) \leq \delta(\{o1, d_1\}) + K$
$\Rightarrow \delta(\{o_1, \ldots, o, d\}) + w(d, d_1) \leq \delta(\{o_1, d_1\}) + K$
However, the shortest path satisfies the triangle inequality property:
$\delta(\{o1, d_1\}) \leq \delta(\{o_1, d\}) + w(d, d_1)$
Thus:
$\delta(\{o_1, ..., o, d\}) \leq \delta(\{o_1, d\}) + K$

In this paper, we start by modeling the flight radius problem in the additive case and aim to study also the multiplicative case which is more complicated.

3.3 Flight Database

We first present the software tool `PlanetOptim` and the graph database proposed besides `MongoDB` and conclude by describing the process of generating and storing the condensed graph in `Neo4j`.

3.3.1 Description of `PlanetOptim`

Our inputs derived from `PlanetOptim` software developed by the firm *Milanamos*. It is a decision-making tool that helps airline companies to analyze market, scheduling flights and forecasting demand to maximize revenue. `PlanetOptim` is composed of three principal modules: Analysis market, Flight simulators and analysis of the hub and routes. We are interested in the second module which is the flight simulator, it consists of simulating flight in function of supply and demand. This one assists the user in evaluating and displaying for a specific flight departure time the connections from 45 until 360 minutes after the Minimum Connecting Time (Milanamos, 2016). Especially when a company wants to add a new flight, it looks for the best time of departure time in order to maximize the connections and thus the number of passengers traveling in that flight. `PlanetOptim` is based on a `NoSQL` flight database named *Optimode*. *Milanamos* uses `MongoDB` for this database to store both structured and unstructured data without schema constraints and thus no option join. In `MongoDB`, we talk about collections and documents rather than tables and rows in relational databases. We are interested in these collections while generating the condensed graph, data are monthly:

- Capacity: Includes data about equipment, origin, destination, airlines. All data about frequency, seat per operation, capacity.
- O&D: It contains all information about passengers itinerary: origin, destination, connecting points, duration, number of passengers, revenue...
- Schedule: Schedule is composed of individual flights between two airports. We distinguish between the two types of schedule: Rotations and routing. A routing is a set of aircraft routes. However, Rotations are routing which begin and end at the same airport. (Barnhart et al., 2003) Schedule collection contains information related to fleets, (OD) pairs, flight numbers...

- Segment: Contains all data of only flight legs: origin, destination, number of passengers, revenue, distance, type of aircraft.

In the air transportation management, the O&D market is defined by a passenger's point of entry and exit from the airline system and it's important to the airlines because it let's them know how many passengers travel between the two cities during a certain time period. However, Segment market information gathered for a specific route operated by an aircraft from a point of origin and a destination when it's a non-stop flight (Milanamos, 2016). We are based on these collections from which we extract information to construct the condensed model. Basically, we are interested in the segment collection since it provides all information for each segment typically the number of passengers traveled from origin to a destination rather than O&D collection which only gives the number passengers for the whole route and not for each sequence or segment of the route.

In optimode, data are collected monthly so it is worthwhile to keep this frequency rather than aggregate to high frequency (e.g. yearly). Note that this make most sense for a monthly result when the original data covers a whole number of months: in particular aggregating a monthly data to yearly starting in February does not give a conventional yearly data. Moreover, it provides a better analysis of evolution traffic per month and then gives a more accurate result. Besides, database does not use graph structure and stores data in disconnected way.

That needs for compact structure to regroup data, store and visualize the graph.

3.3.2 Graph Database

The graph was implemented using the `Neo4j` graph database. It is one of the popular graph databases stands for Cypher query language. It is used in many use cases, typically network routing.
`Neo4j` Graph Database follows the property graph model to store and manage its data. It has the following characteristics (Robinson et al., 2013): Represents data in nodes, relationships and properties. Both nodes and relationships contain properties. A relationship connects a pair of nodes, it has a direction, type, a start node and an end node.

We use `Neo4j` as another alternative that represents a data structure to store the condensed graph. It is an open source project with more utilities besides `NoSQL` database (`MongoDB`). `Neo4j` represents a graph structure that regroups data contained in collections and helps to visualize the graph and also the result of queries. As an example: See what happens if we allocate a new flight to the network which it's realized by adding a new edge to the graph. It makes a lot of sense to store it there since relationships describe if there is at least flight between an OD pair in contrast with `MongoDB` which requires updating all related collections. Moreover, we can do our query easily without any join now that it's provided free by the graph. This graph database response perfectly to our needs since it performed well on the graph traversal. (Holzschuher and Peinl, 2014). With `Neo4j`, we are able to implement algorithms and then store them as a stored procedure to use it in Cypher. It is easy to handle it by the user in `PlanetOptim`.
We are working in a business context. Aviation data are highly connected and grows day over days and `Neo4j` performs well and handles this. Actually, we are studying different models to store our data in `Neo4j`:

- Store all information in a single relationship between an (OD) pair;
- Store monthly data per relationship;
- Construct relationship per criterion between each pair of nodes.

The latter two models generate a huge graph since we duplicate relationships per month but allows to quickly access information. We are benchmarking response time of the following questions (and many others):

- What is the path between an OD pair with the minimum transfer?
- What is the average capacity since 2015?
- What is set of O-D for a period 2016 with capacity greater than a certain value?
- If I want to fly from Paris to New York. Checking for direct flights or check for flights with a stopover at some airport.

The second model is more flexible for several reasons. Firstly, data are monthly collected so it makes sense to create relationship per period which is a month of the year. Secondly, adding a period is realized by adding a relationship. However, it consists to fetch and get the relation and then set the properties in the first models even for the deletion is the same thing. Thirdly, the existence of a 'null' value does not prevent the calculation of an average or a sum and finally response time is so fast.

The graph in `Neo4j` is represented as an adjacency list. To test the existence of a relationship between a couple of nodes, the time response is not constant in contrary to adjacency matrix.

3.3.3 Extraction Process

We get into some issues while extracting data as a part of collecting real world data. We estimate that at least 10% of data are erroneous. Besides, missing data of

certain airports such as: distance, region...and about some flights such as: Departure time, arrival time, duration.

The graph was generated based on three collections: Segment, Capacity and Schedule. The data are extracted as follows: firstly, we get the information about revenue, number of passengers and distance for each month and (OD) pair by aggregated the monthly data then a first join is set to obtain the frequency and the capacity corresponding to this month and (OD). Finally, a second join between Segment and Schedule collection to fetch the duration. Another issue encountered is the distance. Both tables Schedule and Segment contains this information. The first one is collected from booking service. However, the second is calculated by our formula. After a comparison, we conclude that we don't have the same distance for routes matched between collection. The problem is due to erroneous geographic coordinates of airports, the distance is recalculated using the correct coordinates. The process of extracting data is as follows:

- Step 1: Aggregate the frequency and seat per operation for each (OD) pair in capacity collection using the aggregation functions cited in 3.2;
- Step 2: Aggregate duration of schedule collection;
- Step 3: Aggregate the revenue and the number of passengers in segment collection and then use the function join to get the attributes aggregated in step 1 & 2 for that (OD) pair.

Note that we use python to do join since this option is absent in `NoSQL` `MongoDB`. Python works perfectly with `MongoDB` via the pymongo api. The following figure 3 describes the process of generating condensed graph:

Since we opt to monthly frequency, we aggregate data per (origin, destination, month of the year): (O,D,Y_M). After the process of extracting data, we store the graph in `Neo4j` using the driver py2neo. Nodes represent airports, relationship per period (a month of year) and properties regroup the labels of the condensed graph. The graph was generated over the last year (2015) and has 11,668 nodes and 608,812 relations.

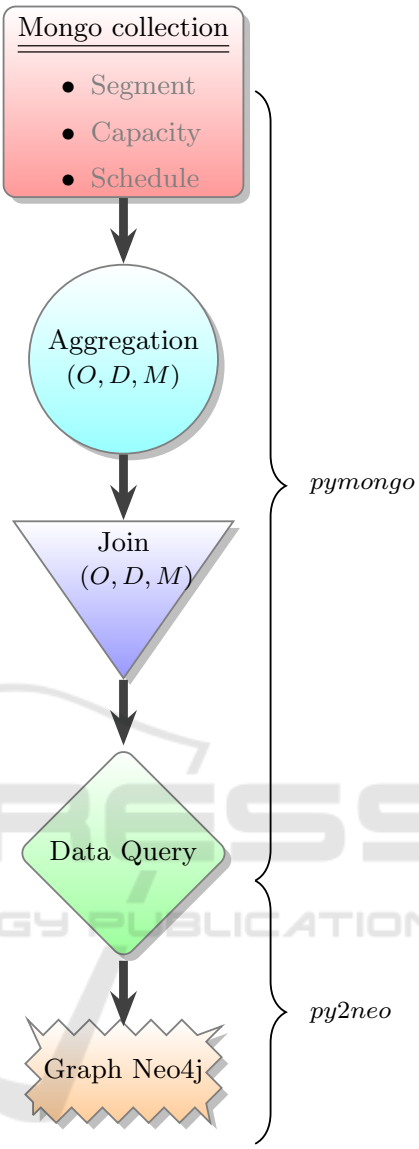

Figure 3: The process of constructing the graph.

4 STAGE OF THE RESEARCH

The overall airline scheduling design process involves hierarchical steps starting with the route network development and ending with crew assignment. Route network development is especially important with respect to timing and costs for airlines. In this paper, we focus on allocating a new flight problem that consists of determining a set of OD pairs and then choosing the arrival and departure times for an aircraft. We presented a sub-problem of this problem (Flight radius problem) which helps airline managers to visualize the interesting sub-network of such a flight based on QSI models. The result can be implemented in the short term within the existing application in order to improve the visualization of flight network and thus help airline managers to make routing decision about set of (OD) pairs to serve and also choose schedules that maximize revenue. We formulated the problem as finding a maximal sub-graph such that each edge supports a valid path. Such path depends on regret function that model the passengers preferences regarding the cost and the duration and then showed how to construct graph from a real data which it's stored in a

NoSQL database and presented an alternative to store the graph.

Our research aims to solve the problem of allocating a new flight which is a sub-problem of route network development problem based on QSI methodology that is most used by airlines in trip choice.

Regarding future work, it would be interesting to choose a shortest path algorithm to solve the flight radius problem and as mentioned in section 3 'Outline of Objectives', the next step in this research is to solve the allocating a new flight since the first two of the five objectives outlined have already been accomplished. We may extend our recent work on the flight route planning to multimodal route planning.

5 EXPECTED OUTCOME

The final outcome of this research is to help airline managers to make rational decisions by improving the application PlanetOptim. We aim to define a new problem of modeling expert decision-making in air traffic and then solve it by using our expertise in graph theory; apply algorithms that have already studied in the literature but also propose new one adaptable to graph database. Moreover, our objective is to study the complexity of these algorithms in the case of the graph databases that are little studied in academic area especially that they don't stand for adjacency matrix.

REFERENCES

Barnhart, C., Belobaba, P., and Odoni, A. R. (2003). Applications of operations research in the air transport industry. *Transportation science*, 37(4):368–391.

Bast, H., Delling, D., Goldberg, A. V., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., and Werneck, R. F. (2015). Route planning in transportation networks. *CoRR*, abs/1504.05140.

Belobaba, P., Odoni, A., and Barnhart, C. (2015). *The global airline industry*. John Wiley & Sons.

Carmona Benitez, R. (2012). *The Design of a Large Scale Airline Network*. TU Delft, Delft University of Technology.

Delling, D., Pajor, T., and Wagner, D. (2009). Engineering time-expanded graphs for faster timetable information. In *Robust and Online Large-Scale Optimization*, pages 182–206. Springer.

Dobson, G. and Lederer, P. J. (1993). Airline scheduling and routing in a hub-and-spoke system. *Transportation Science*, 27(3):281–297.

Hall, R. (2012). *Handbook of transportation science*, volume 23. Springer Science & Business Media.

Holzschuher, F. and Peinl, R. (2014). Performance optimization for querying social network data. In *EDBT/ICDT Workshops*, pages 232–239.

Jacobs, T. L., Garrow, L. A., Lohatepanont, M., Koppelman, F. S., Coldren, G. M., and Purnomo, H. (2012). Airline planning and schedule development. In *Quantitative Problem Solving Methods in the Airline Industry*, pages 35–99. Springer.

Kirchler, D. (2013). *Efficient routing on multi-modal transportation networks*. PhD thesis, Ecole Polytechnique X.

Marwaha, G. and Kokkolaras, M. (2015). System-of-systems approach to air transportation design using nested optimization and direct search. *Structural and Multidisciplinary Optimization*, 51(4):885–901.

Milanamos (2016). *User Manual of PlanetOptim*.

Rebetanety, A. (2006). *Airline schedule planning integrated flight schedule design and product line design. University Karlsruhe (TH)*. PhD thesis, PhD thesis, 2006. Available at http://www. iks. kit. edu/fileadmin/User/calmet/stdip/dip-rabentanety. pdf. Accessed 2013 January 30.

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O'Reilly Media, Inc.

Sivrikaya, O. (2013). Demand forecasting for domestic air transportation in turkey. *The Open Transportation Journal*, 7(1):20–26.

Swan, W. (2008). Forecasting air travel with open skies. In *joint EWCKOTI Conference*.