Semantic Web 0 (2018) 1
IOS Press

Generating Public Transport Data based on Population Distributions for RDF Benchmarking

Ruben Taelman, Pieter Colpaert, Erik Mannens and Ruben Verborgh
imec – Ghent University – IDLab, Technologiepark-Zwijnaarde 15, B-9052 Ghent, Belgium
E-mail: ruben.taelman@ugent.be

Abstract.
When benchmarking RDF data management systems such as public transport route planners, system evaluation needs to happen under various realistic circumstances, which requires a wide range of datasets with different properties. Real-world datasets are almost ideal, as they offer these realistic circumstances, but they are often hard to obtain and inflexible for testing. For these reasons, synthetic dataset generators are typically preferred over real-world datasets due to their intrinsic flexibility. Unfortunately, many synthetic dataset that are generated within benchmarks are insufficiently realistic, raising questions about the generalizability of benchmark results to real-world scenarios. In order to benchmark geospatial and temporal RDF data management systems such as route planners with sufficient external validity and depth, we designed PODIGG, a highly configurable generation algorithm for synthetic public transport datasets with realistic geospatial and temporal characteristics comparable to those of their real-world variants. The algorithm is inspired by real-world public transit network design and scheduling methodologies. This article discusses the design and implementation of PODIGG and validates the properties of its generated datasets. Our findings show that the generator achieves a sufficient level of realism, based on the existing coherence metric and new metrics we introduce specifically for the public transport domain. Thereby, PODIGG provides a flexible foundation for benchmarking RDF data management systems with geospatial and temporal data.

Keywords: Public Transport, Dataset Generator, Benchmarking, RDF, Linked Data

1. Introduction

The Resource Description Framework (RDF) and Linked Data technologies enable distributed use and management of semantic data models [11,6]. Datasets with an interoperable domain model can be stored and queried by different data owners in different ways. In order to discover the strengths and weaknesses of different storage and querying possibilities, data-driven benchmarks with different sizes of datasets and varying characteristics can be used.

Regardless of whether existing data-driven benchmarks use real or synthetic datasets, the *external validity* of their results can be too limited, which makes a generalization to other datasets difficult. Real datasets, on the one hand, are often only scarcely available for testing, and only cover very specific scenarios, such that not all aspects of systems can be assessed. Synthetic datasets, on the other hand, are typically generated by *mimicking algorithms* [7,21,29,30], which are not always sufficiently realistic [15]. Features that are relevant for real-world datasets may not be tested. As such, conclusions drawn from existing benchmarks do not always apply to the envisioned real-world scenarios. One way to get the best of both worlds is to design mimicking algorithms that generate realistic synthetic datasets.

The *public transport* domain provides data with both geospatial and temporal properties, which makes this

an especially interesting source of data for benchmarking. Its representation as Linked Data is valuable because i) of the many shared entities, such as stops, routes and trips, across different existing datasets on the Web. ii) These entities can be distributed over different datasets and iii) benefit from interlinking for the improvement of discoverability. Synthetic public transport datasets are particularly important and needed in cases where public transport route planning algorithms are evaluated. The Linked Connections framework [10] and Connection Scan Algorithm [14] are examples of such public transport route planning systems. Because of the limited availability of real-world datasets with desired properties, these systems were evaluated with only a very low number of datasets, respectively one and three datasets. A synthetic public transport dataset generator would make it easier for researchers to include a higher number of realistic datasets with various properties in their evaluations, which would be beneficial to the discovery of new insights from the evaluations. Network size, network sparsity and temporal range are examples of such properties, and different combinations of them may not always be available in real datasets, which motivates the need for generating synthetic, but realistic datasets with these properties.

Not only are public transport datasets useful for benchmarking route planning systems, they are also highly useful for benchmarking geospatial [22,5] and temporal RDF systems [3,23] due to the intrinsic geospatial and temporal properties of public transport datasets. While synthetic dataset generators already exist in the geospatial and temporal domain [17,24], no systems exist yet that focus on realism, and specifically look into the generation of public transport datasets. As such, the main topic that we address in this work, is solving the need for realistic public transport datasets with geospatial and temporal characteristics, so that they can be used to benchmark RDF data management and route planning systems. More specifically, we introduce a mimicking algorithm for generating realistic public transport data, which is the main contribution of this work.

We observed a significant correlation between transport networks and the population distributions of their geographical areas, which is why population distributions are the driving factor within our algorithm. The cause of this correlation is obvious, considering transport networks are frequently used to transport people, but other – possibly independent – factors exist that influence transport networks as well, like certain points of interest such as tourist attractions and shopping areas. Our algorithm is subdivided into five sequential steps, inspired by existing methodologies from the domains of public transit planning [20] as a means to improve the realism of the algorithm's output data. These steps include the creation of a geospatial region, the placement of stops, edges and routes, and the scheduling of trips. We provide an implementation of this algorithm, with different parameters to configure the algorithm. Finally, we confirm the realism of datasets that are generated by this algorithm using the existing generic structuredness metric [15] and new metrics that we introduce, which are specific to the public transport domain. The notable difference of this work compared to other synthetic dataset generators is that our generation algorithm specializes in generating public transit networks, while other generators either focus on other domains, or aim to be more general-purpose. Furthermore, our algorithm is based on population distributions and existing methodologies from public transit network design.

In the next section, we introduce the related work on dataset generation, followed by the background on public transit network design, and transit feed formats in Section 3. In Section 4, we introduce the main research question and hypothesis of this work. Next, our algorithm is presented in Section 5, followed by its implementation in Section 6. In Section 7, we present the evaluation of our implementation, followed by a discussion and conclusion in Section 8 and Section 9.

2. Related Work

In this section, we present the related work on spatiotemporal and RDF dataset generation,

Spatiotemporal database systems store instances that are described using an identifier, a spatial location and a timestamp. In order to evaluate spatiotemporal indexing and querying techniques with datasets, automatic means exist to generate such datasets with predictable characteristics [26].

Brinkhoff [8] argues that moving objects tend to follow a predefined network. Using this and other statements, he introduces a spatiotemporal dataset generator. Such a network can be anything over which certain objects can move, ranging from railway networks to air traffic connections. The proposed parameter-based generator restricts the existence of the spatiotemporal objects to a predefined time period $[t_{\text{min}}, t_{\text{max}})$. It is assumed that each edge in the network has a maximum allowed speed and capacity over which objects can move at a certain speed. The eventual speed of each object is defined by the maximum speed of its class, the

maximum allowed speed of the edge, and the congestion of the edge based on its capacity. Furthermore, external events that can impact the movement of the objects, such as weather conditions, are represented as temporal grids over the network, which apply a *decreasing factor* on the maximum speed of the objects in certain areas. The existence of each object that is generated starts at a certain timestamp, which is determined by a certain function, and *dies* when it arrives at its destination. The starting node of an object can be chosen based on three approaches:

dataspace-oriented approaches Selecting the nearest node to a position picked from a two-dimensional distribution function that maps positions to nodes.

region-based approaches Improvement of the dataspace oriented approach where the data space is represented as a collection of cells, each having a certain chance of being the place of a starting node.

network-based approaches Selection of a network node based on a one-dimensional distribution function that assigns a chance to each node.

Determining the destination node using one of these approaches leads to non-satisfying results. Instead, the destination is derived from the preferred length of a route. Each route is determined as the fastest path to a destination, weighed by the external events. Finally, the results are reported as either textual output, insertion into a database or a figure of the generated objects. Compared to our work, this approach assumes a predefined network, while our algorithm also includes the generation of the network. For our work, we reuse the concepts of object speed and region-based node selection with relation to population distributions.

In order to improve the testability of Information Discovery Systems, a generic synthetic dataset generator [25] was developed that is able to generate synthetic data based on declarative graph definitions. This graph is based on objects, attributes and relationships between them. The authors propose to generate new instances, such as people, based on a set of dependency rules. They introduce three types of dependencies for the generation of instances:

independent Attribute values that are independent of other instances and attributes.

intra-record (horizontal) dependencies Attribute values depending on other values of the same instance.

inter-record (vertical) dependencies Relationships between different instances.

Their engine is able to accept such dependencies as part of a semantic graph definition, and iteratively create new instances to form a synthetic dataset. This tool however outputs non-RDF CSV files, which makes it impossible to directly use this system for the generation of public transport datasets in RDF using existing ontologies. For our public transport use case, individual entities such as stops, stations and connections would be possible to generate up to a certain level using this declarative tool. However, due to the underlying relation to population distributions and specific restrictions for resembling real datasets, declarative definitions are too limited.

The need for benchmarking RDF data management systems is illustrated by the existence of the Linked Data Benchmark Council [2] and the HOBBIT[1] H2020 EU project for benchmarking of Big Linked Data. RDF benchmarks are typically based on certain datasets that are used as input to the tested systems. Many of these datasets are not always very closely related to real datasets [15], which may result in conclusions drawn from benchmarking results that do not translate to system behaviours in realistic settings.

Duan et al. [15] argue that the realism of an RDF dataset can be measured by comparing the *structuredness* of that dataset with a realistic equivalent. The authors show that real-world datasets are typically less structured than their synthetic counterparts, which can results in significantly different benchmarking results, since this level of structuredness can have an impact on how certain data is stored in RDF data management systems. This is because these systems may behave differently on datasets with different levels of structuredness, as they can have certain optimizations for some cases. In order to measure this structuredness, the authors introduce the *coherence* metric of a dataset D with a type system $\mathcal{T}$ that can be calculated as follows:

$$\mathrm{CH}(\mathcal{T}, D) = \sum_{\forall T \in \mathcal{T}} \mathrm{WT}(\mathrm{CV}(T, D)) * \mathrm{CV}(T, D) \quad (1)$$

The type system $\mathcal{T}$ contains all the RDF types that are present in a dataset. $\mathrm{CV}(T, D)$ represents the *coverage* of a type T in a dataset D, and is calculated as the fraction of type instances that set a value for all its properties. The factor $\mathrm{WT}(\mathrm{CV}(T, D))$ is used to weight this sum, so that the coherence is always a value between 0 and 1, with 1 representing a perfect structuredness. A maximal coherence means that all instances in the dataset have values for all possible properties in the type system,

[1] http://project-hobbit.eu/

which is for example the case in relational databases without optional values. Based on this metric, the authors introduce a generic method for creating variants of real datasets with different sizes while maintaining a similar structuredness.The authors describe a method to calculate the coverage value of this dataset, which has been implemented as a procedure in the Virtuoso RDF store [30]. As the goal of our work is to generate *realistic* RDF public transport datasets, we will use this metric to compare the realism of generated datasets with real datasets. As this high-level metric is used to define *realism* over any kind of RDF dataset, we will introduce new metrics to validate the realism for specifically the case of public transport datasets.

3. Public Transit Background

In this section, we present background on public transit planning that is essential to this work. We discuss existing public transit network planning methodologies and formats for exchanging transit feeds.

3.1. Public Transit Planning

The domain of public transit planning entails the design of public transit networks, rostering of crews, and all the required steps inbetween. The goal is to maximize the quality of service for passengers while minimizing the costs for the operator. Given a public demand and a topological area, this planning process aims to obtain routes, timetables and vehicle and crew assignment. A survey about 69 existing public transit planning approaches shows that these processes are typically subdivided into five sequential steps [20]:

1. **route design**, the placement of transit routes over an existing network.
2. **frequencies setting**, the temporal instantiation of routes based on the available vehicles and estimated demand.
3. **timetabling**, the calculation of arrival and departure times at each stop based on estimated demand.
4. **vehicle scheduling**, vehicle assignment to trips.
5. **crew scheduling and rostering**, the assignment of drivers and additional crew to trips.

In this paper, we only consider the first three steps for our mimicking algorithm, which lead to all the required information that is of importance to passengers in a public transit schedule. We present the three steps from this survey in more detail hereafter.

The first step, route design, requires the topology of an area and public demand as input. This topology describes the network in an area, which contains possible stops and edges between these stops. Public demand is typically represented as *origin-destination* (OD) matrices, which contain the number of passengers willing to go from origin stops to destination stops. Given this input, routes are designed based on the following objectives [20]:

area coverage The percentage of public demand that can be served.

route and trip directness A metric that indicates how much the actual trips from passengers deviate from the shortest path.

demand satisfaction How many stops are close enough to all origin and destination points.

total route length The total distance of all routes, which is typically minimized by operators.

operator-specific objectives Any other constraints the operator has, for example the shape of the network.

historical background Existing routes may influence the new design.

The next step is the setting of frequencies, which is based on the routes from the previous step, public demand and vehicle availability. The main objectives in this step are based on the following metrics [20]:

demand satisfaction How many stops are serviced frequently enough to avoid overcrowding and long waiting times.

number of line runs How many times each line is serviced – a trade-off between the operator's aim for minimization and the public demand for maximization.

waiting time bounds Regulation may put restrictions on minimum and maximum waiting times between line runs.

historical background Existing frequencies may influence the new design.

The last important step for this work is timetabling, which takes the output from the previous steps as input, together with the public demand. The objectives for this step are the following:

demand satisfaction Total travel time for passengers should be minimized.

transfer coordination Transfers from one line to another at a certain stop should be taken into account during stop waiting times, including how many passengers are expected to transfer.

fleet size The total amount of available vehicles and their usage will influence the timetabling possibilities.

historical background Existing timetables may influence the new design.

3.2. Transit Feed Formats

The de-facto standard for public transport time schedules is the General Transit Feed Specification (GTFS)[2]. GTFS is an exchange format for transit feeds, using a series of CSV files contained in a ZIP file. The specification uses the following terminology to define the rules for a public transit system:

Stop is a geospatial location where vehicles stop and passengers can get on or off, such as platform 3 in the train station of Brussels.

Stop time indicates a scheduled arrival and departure time at a certain stop.

Route is a time-independent collection of stops, describing the sequence of stops a certain vehicle follows in a certain public transit line. For example the train route from Brussels to Ghent.

Trip is a collection of stops with their respective stop times, such as the route from Brussels to Ghent at a certain time.

The ZIP file is put online by a public transit operator, to be downloaded by route planning [13] software. Two models are commonly used to then extract these rules into a graph [28]. In a *time-expanded model*, a large graph is modeled with arrivals and departures as nodes and edges connect departures and arrivals together. The weights on these edges are constant. In a *time-dependent model*, a smaller graph is modeled in which vertices are physical stops and edges are transit connections between them. The weights on these edges change as a function of time. In both models, Dijkstra and Dijkstra-based algorithms can be used to calculate routes.

In contrast to these two models, the *Connection Scan Algorithm* [14] takes an ordered array representation of *connections* as input. A connection is the actual departure time at a stop and an arrival at the next stop. These connections can be given a URI, and described using RDF, using the Linked Connections [10] ontology. For this base algorithm and its derivatives, a connection object is the smallest building block of a transit schedule.

In our work, generated public transport networks and time schedules can be serialized to both the GTFS format, and RDF datasets using the Linked Connections ontology.

4. Research Question

In order to generate public transport networks and schedules, we start from the hypothesis that both are correlated with the population distribution within the same area. More populated areas are expected to have more nearby and more frequent access to public transport, corresponding to the recurring demand satisfaction objective in public transit planning [20]. When we calculate the correlation between the distribution of stops in an area and its population distribution, we discover a positive correlation[3] of 0.439 for Belgium and 0.459 for the Netherlands, thereby validating our hypothesis with a confidence of 99%. Because of the continuous population variable and the binary variable indicating whether or not there is a stop, the correlation is calculated using the point-biserial correlation coefficient[4]. For the calculation of these correlations, we ignored the population value outliers. Following this conclusion, our mimicking algorithm will use such population distributions as input, and derive public transport networks and trip instances.

The main objective of a mimicking algorithm is to create *realistic* data, so that it can be used to by benchmarks to evaluate systems under realistic circumstances. We will measure dataset realism in high-level by comparing the levels of structuredness of real-world datasets and their synthetic variants using the *coherence metric* introduced by Duan et al. [15]. Furthermore, we will measure the realism of different characteristics within public transport datasets, such as the location of stops, density of the network of stops, length of routes or the frequency of connections. We will quantify these aspects by measuring the distance of each aspect between real and synthetic datasets. These dataset characteristics will be linked with potential evaluation metrics within RDF data management systems, and tasks to evaluate them. This generic coherence metric together with domain-specific metrics will provide a way to evaluate dataset realism.

Based on this, we introduce the following research question for this work: “Can population distribution data be used to generate realistic synthetic public transport

[2] https://developers.google.com/transit/gtfs/

[3] p-values in both cases < 0.00001

[4] https://github.com/PoDiGG/podigg-evaluate/blob/master/stats/correlation.r

networks and scheduling?" We provide an answer to this question by first introducing an algorithm for generating public transport networks and their scheduling based on population distributions in Section 5. After that, we validate the realism of datasets that were generated using an implementation of this algorithm in Section 7.

5. Method

In order to formulate an answer to our research question, we designed a mimicking algorithm that generates realistic synthetic public transit feeds. We based it on techniques from the domains of public transit planning, spatiotemporal and RDF dataset generation. We reuse the route design, frequencies setting and timetabling steps from the domain public transit planning, but prepend this with a network generation phase.

Figure 1 shows the model of the generated public transit feeds, with connections being the primary data element. We consider different properties in this model based on the independent, intra-record or inter-record dependency rules [25], as discussed in Section 2. The arrival time in a connection can be represented as a fully intra-record dependency, because it depends on the time it departed and the stops it goes between. The departure time in a connection is both an intra-record and inter-record dependency, because it depends on the stop at which it departs, but also on the arrival time of the connection before it in the trip. Furthermore, the delay value can be seen as an inter-record dependency, because it is influenced by the delay value of the previous connection in the trip. Finally, the geospatial location of a stop depends on the location of its parent station, so this is also an inter-record dependency. All other unmentioned properties are independent.

In order to generate data based on these dependency rules, our algorithm is subdivided in five steps:

1. **Region**: Creation of a two-dimensional area of cells annotated with population density information.
2. **Stops**: Placement of stops in the area.
3. **Edges**: Connecting stops using edges.
4. **Routes**: Generation of routes between stops by combining edges.
5. **Trips**: Scheduling of timely trips over routes by instantiating connections.

These steps are not fully sequential, since stop generation is partially executed before and after edge generation. The first three steps are required to generate a network, step 4 corresponds to the route design step in public transit planning and step 5 corresponds to both the frequencies setting and timetabling. These steps are explained in the following subsections.

5.1. Region

In order to create networks, we sample geographic regions in which such networks exist as two-dimensional matrices. The resolution is defined as a configurable number of cells per square of one latitude by one longitude. Network edges are then represented as links between these cells. Because our algorithm is population distribution-based, each cell contains a population density. These values can either be based on real population information from countries, or this can be generated based on certain statistical distributions. For the remainder of this paper, we will reuse the population distribution from Belgium as a running example, as illustrated in Figure 2.

5.2. Stops

Stop generation is divided into two steps. First, stops are placed based on population values, then the edge generation step is initiated after which the second phase of stop generation is executed where additional stops are created based on the generated edges.

Population-based For the initial placement of stops, our algorithm only takes a population distribution as input. The algorithm iteratively selects random cells in the two-dimensional area, and tags those cells as stops. To make it region-based [8], the selection uses a weighted Zipf-like-distribution, where cells with high population values have a higher chance of being picked than cells with lower values. The shape of this Zipf curve can be scaled to allow for different stop distributions to be configured. Furthermore, a minimum distance between stops can be configured, to avoid situations where all stops are placed in highly population areas.

Edge-based Another stop generation phase exists after the edge generation because real transit networks typically show line artifacts for stop placement. Figure 3a shows the actual train stops in Belgium, which clearly shows line structures. Stop placement after the first generation phase results can be seen in Figure 3b, which does not show these line structures. After the second stop generation phase, these line structures become more apparent as can be seen in Figure 3c. In this

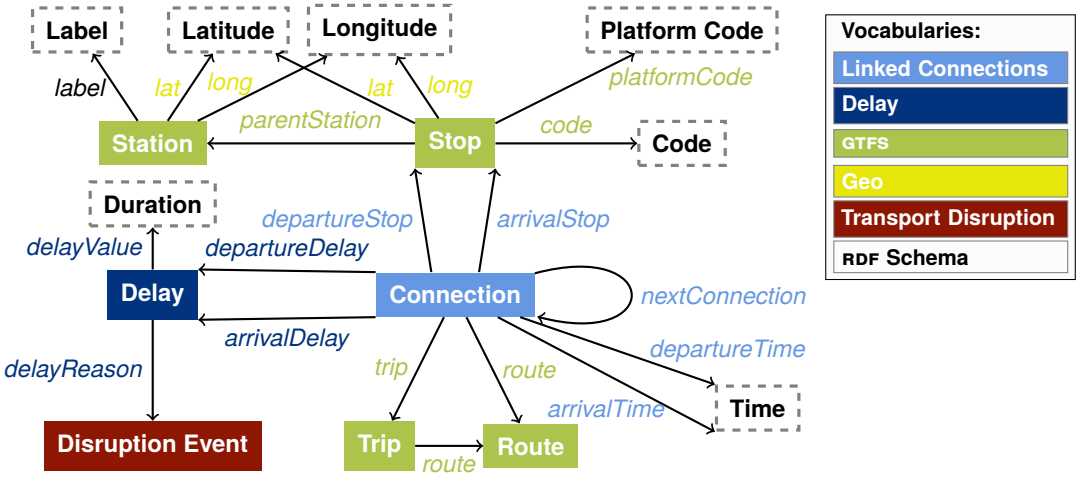

Figure 1. The resources (rectangle), literals (dashed rectangle) and properties (arrows) used to model the generated public transport data. Node and text colors indicate vocabularies.

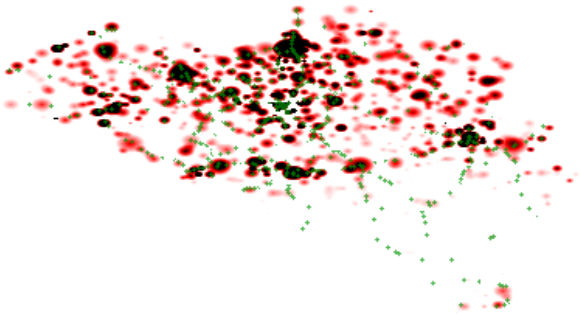

Figure 2. Heatmap of the population distribution in Belgium, which is illustrated for each cell as a scale going from white (low), to red (medium) and black (high). The actual placement of train stops are indicated as green points.

Fig. a: Real stops with line structures

Fig. b: Synthetic stops after the first stop generation phase without line structures.

Fig. c: Synthetic stops after the second stop generation phase with line structures.

Figure 3. Placement of train stops in Belgium, each dot represents one stop.

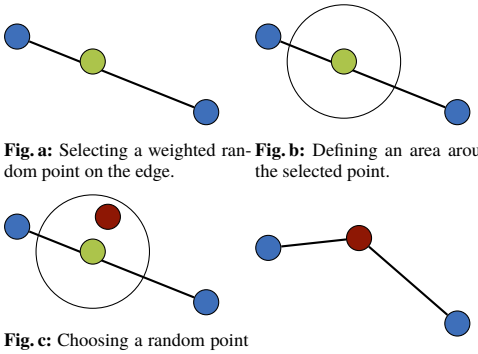

Fig. a: Selecting a weighted random point on the edge.

Fig. b: Defining an area around the selected point.

Fig. c: Choosing a random point within the area, weighted by population value.

Fig. d: Modify edges so that the path includes this new point.

Figure 4. Illustration of the second phase of stop generation where edges are modified to include sufficiently populated areas in paths.

second stop generation phase, edges are modified so that sufficiently populated areas will be included in paths formed by edges, as illustrated by Figure 4. Random edges will iteratively be selected, weighted by the edge length measured as Euclidian distance[5]. On each edge, a random cell is selected weighed by the population value in the cell. Next, a weighed random point in a certain area around this point is selected. This selected point is marked as a stop, the original edge is removed and two new edges are added, marking the path between the two original edge nodes and the newly selected node.

[5] The Euclidian distance based on geographical coordinates is always used to calculate distances in this work.

5.3. Edges

The next phase in public transit network generation connects the stops that were generated in the previous phase with edges. In order to simulate real transit network structures, we split up this generation phase into three sequential steps. In the first step, clusters of nearby stops are formed, to lay the foundation for short-distance routes. Next, these local clusters are connected with each other, to be able to form long-distance routes. Finally, a cleanup step is in place to avoid abnormal edge structures in the network.

Short-distance The formation of clusters with nearby stations is done using agglomerative hierarchical clustering. Initially, each stop is part of a seperate cluster, where each cluster always maintains its centroid. The clustering step will iteratively try to merge two clusters with their centroid distance below a certain threshold. This threshold will increase for each iteration, until a maximum value is reached. The maximum distance value indicates the maximum inter-stop distance for forming local clusters. When merging two clusters, an edge is added between the closest stations from the respective clusters. The center location of the new cluster is also recalculated before the next iteration.

Long-distance At this stage, we have several clusters of nearby stops. Because all stops need to be reachable from all stops, these separate clusters also need to be connected. This problem is related to the domain of route planning over public transit networks, in which networks can be decomposed into smaller clusters of nearby stations to improve the efficiency of route planning. Each cluster contains one or more *border stations* [4], which are the only points through which routes can be formed between different clusters. We reuse this concept of border stations, by iteratively picking a random cluster, identifying its closest cluster based on the minimal possible stop distance, and connecting their border stations using a new edge. After that, the two clusters are merged. The iteration will halt when all clusters are merged and there is only one connected graph.

Cleanup The final cleanup step will make sure that the number of stops that are connected by only one edge are reduced. In real train networks, the majority of stations are connected with at least more than one other stations. The two earlier generation steps however generate a significant number of *loose stops*, which are connected

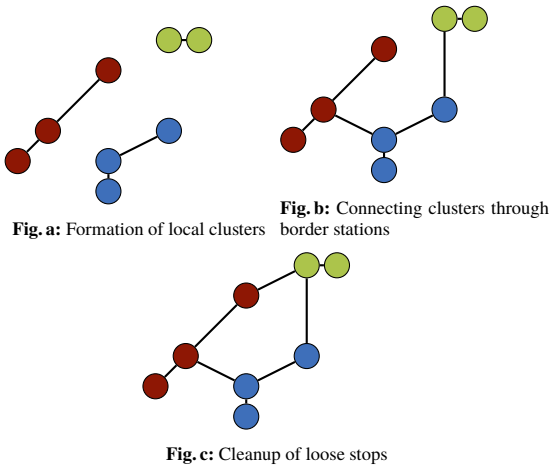

Fig. a: Formation of local clusters

Fig. b: Connecting clusters through border stations

Fig. c: Cleanup of loose stops

Figure 5. Example of the different steps in the edges generation algorithm

with only a single other stop with a direct edge. In this step, these loose stops are identified, and an attempt is made to connect them to other nearby stops as shown in Algorithm 1. For each loose stop, this is done by first identifying the direction of the single edge of the loose stop on line 8. This direction is scaled by the radius in which to look for stops, and defines the stepsize for the loop the starts on line 10. This loop starts from the loose stop and iteratively moves the search position in the defined direction, until it finds a random stop in the radius, or the search distance exceeds the average distance of between the stops in the neighbourhood of this loose stop. This random stop from line 12 can be determined by finding all stations that have a distance to the search point that is below the radius, and picking a random stop from this collection. If such a stop is found, an edge is added from our loose stop to this stop.

Figure 5 shows an example of these three steps. After this phase, a network with stops and edges is available, and the actual transit planning can commence.

Generator Objectives The main guaranteed objective of the edge generator is that the stops form a single connected transit network graph. This is to ensure that all stops in the network can be reached from any other stop using at least one path through the network.

5.4. Routes

Given a network of stops and edges, this phase generates routes over the network. This is done by creating short and long distance routes in two sequential steps.

Function `RemoveLooseStops`(S, E, N, θ, ϑ)
  Input: Set of stops S;
  Set of edges E between the stops from S;
  Maximum number N of closest stations to consider;
  Maximum average distance θ around a stop to be considered a loose station;
  Radius ϑ in which to look for stops.
foreach $s \in S$ *with degree of 1 w.r.t.* E **do**
  s_x = x coordinate of s; s_y = y coordinate of s;
  Λ = N closest stations to s in S excluding s;
  λ = closest station to s in S excluding s; λ_x = x coordinate of λ; λ_y = y coordinate of λ;
  δ = average distance between each pair of stops in Λ;
  if $\delta <= \theta$ *and* Λ *not empty* **then**
    $\phi_x = (s_x - \lambda_x) * \vartheta$; $\phi_y = (s_y - \lambda_y) * \vartheta$;
    $o_x = s_x$; $o_y = s_y$;
    while *distance between* o *and* $s < \delta$ **do**
      o_x += ϕ_x; o_y += ϕ_y;
      s' = random station around o with radius $\delta * \vartheta$;
      if s' *exists*, add edge between s and s' to E and continue next for-loop iteration;

Algorithm 1: Reduce the number of loose stops by adding additional edges.

Short-distance The goal of the first step is to create short routes where vehicles deliver each passed stop. This step makes sure that all edges are used in at least one route, this ensures that each stop can at least be reached from each other stop with one or more transfers to another line. The algorithm does this by first determining a subset of the largest stops in the network, based on the population value. The shortest path from each large stop to each other large stop through the network is determined. if this shortest path is shorter than a predetermined value in terms of the number of edges, then this path is stored as a route, in which all passed stops are considered as actual stops in the route. For each edge that has not yet been passed after this, a route is created by iteratively adding unpassed edges to the route that are connected to the edge until an edge is found that has already been passed.

Long-distance In the next step, longer routes are created, where the transport vehicle not necessarily halts at each passed stop. This is done by iteratively picking two stops from the list of largest stops using the network-based method [8] with each stop having an equal chance to be selected. A heuristical shortest path algorithm is used to determine a route between these stops. This algorithm searches for edges in the geographical direction of the target stop. This is done to limit the complexity of finding long paths through potentially large networks. A random amount of the largest stops on the path are selected, where the amount is a value between a minimum and maximum preconfigured route length. This iteration ends when a predetermined number of routes are generated.

Generator Objectives This algorithm takes into account the objectives of route design [20], as discussed in Section 2. More specifically, by first focusing on the largest stops, a minimal level of *area coverage* and *demand satisfaction* is achieved, because the largest stops correspond to highly populated areas, which therefore satisfies at least a large part of the population. By determining the shortest path between these largest stops, the *route and trip directness* between these stops is optimal. Finally, by not instantiating all possible routes over the network, the *total route length* is limited to a reasonable level.

5.5. Trips

A time-agnostic transit network with routes has been generated in the previous steps. In this final phase, we temporally instantiate routes by first determining starting times for trips, after which the following stop times can be calculated based on route distances. Instead of generating explicit timetables, as is done in typical transit scheduling methodologies, we create fictional rides of vehicles. In order to achieve realistic trip times, we approximate real trip time distributions, with the possibility to encounter delays.

As mentioned before in Section 2, each consecutive pair of start and stop time in a trip over an edge corre-

sponds to a connection. A connection can therefore be represented as a pair of timestamps, a link to the edge representing the departure and arrival stop, a link to the trip it is part of, and its index within this trip.

Trip Starting Times The trips generator iteratively creates new connections until a predefined number is reached. For each connection, a random route is selected with a larger chance of picking a long route. Next, a random start time of the connection is determined. This is done by first picking a random day within a certain range. After that, a random hour of the day is determined using a preconfigured distribution. This distribution is derived from the public logs of iRail[6], a route planning API in Belgium [9]. A seperate hourly distribution is used for weekdays and weekends, which is chosen depending on the random day that was determined.

Stop Times Once the route and the starting time have been determined, different stop times across the trip can be calculated. For this, we take into account the following factors:

- Maximum vehicle speed ω, preconfigured constant.
- Vehicle acceleration ς, preconfigured constant.
- Connection distance δ, Euclidian distance between stops in network.
- Stop size σ, derived from population value.

For each connection in the trip, the time it takes for a vehicle to move between the two stops over a certain distance is calculated using the formula in Equation 4. Equation 2 calculates the required time to reach maximum speed and Equation 3 calculates the required distance to reach maximum speed. This formula simulates the vehicle speeding up until its maximum speed, and slowing down again until it reaches its destination. When the distance is too short, the vehicle will not reach its maximum speed, and just speeds up as long as possible until is has to slow down again to stop in time.

$$T_\omega = \omega/\varsigma \tag{2}$$

$$\delta_\omega = T_\omega^2 \cdot \varsigma \tag{3}$$

$$\text{duration} = \begin{cases} 2T_\omega + (\delta - 2\delta_\omega)/\omega & \text{if } \delta_\omega < \delta/2 \\ \sqrt{2\delta/\varsigma} & \text{otherwise} \end{cases} \tag{4}$$

Not only the connection duration, but also the waiting times of the vehicle at each stop are important for determining the stop times. These are calculated as a constant minimum waiting time together with a waiting time that increases for larger stop sizes σ, this increase is determined by a predefined growth factor.

Delays Finally, each connection in the trip will have a certain chance to encounter a delay. When a delay is applicable, a delay value is randomly chosen within a certain range. Next to this, also a cause of the delay is determined from a preconfigured list. These causes are based on the Traffic Element Events from the Transport Disruption ontology[7], which contains a number of events that are not planned by the network operator such as strikes, bad weather or animal collisions. Different types of delays can have a different impact factor of the delay value, for instance, simple delays caused by rush hour would have a lower impact factor than a major train defect. Delays are carried over to next connections in the trip, with again a chance of encountering additional delay. Furthermore, these delay values can also be reduced when carried over to the next connection by a certain predetermined factor, which simulates the attempt to reduce delays by letting vehicles drive faster.

Generator Objectives For trip generation, we take into account several objectives from the setting of frequencies and timetabling from transit planning [20]. By instantiating more long distance routes, we aim to increase *demand satisfaction* as much as possible, because these routes deliver busy and populated areas, and the goal is to deliver these more frequently. Furthermore, by taking into account realistic time distributions for trip instantiation, we also adhere to this objective. Secondly, by ensuring waiting times at each stop that are longer for larger stations, the *transfer coordination* objective is taken into account to some extent.

6. Implementation

In this section, we discuss the implementation details of PODIGG, based on the generator algorithm introduced in Section 5. PODIGG is split up into two parts: the main PODIGG generator, which outputs GTFS data, and PODIGG-LC, which depends on the main generator to output RDF data. Serialization in RDF using existing ontologies,

[6] https://hello.irail.be

[7] https://transportdisruption.github.io/

such as the GTFS[8] and Linked Connections[9] ontologies, allows this inherently linked data to be used within RDF data management systems, where it can for instance be used for benchmarking purposes. Providing output in GTFS will allow this data to be used directly within all systems that are able to handle transit feeds, such as route planning systems. The two generator parts will be explained hereafter, followed by a section on how the generator can be configured using various parameters.

6.1. PODIGG

The main requirement of our system is the ability to generate realistic public transport datasets using the mimicking algorithm that was introduced in Section 5. This means that given a population distribution of a certain region, the system must be able to design a network of routes, and determine timely trips over this network.

PODIGG is implemented to achieve this goal. It is written in JavaScript using Node.js, and is available under an open license on GitHub[10]. In order to make installation and usage more convenient, PODIGG is available as a Node module on the NPM[11] package manager and as a Docker image on Docker Hub[12] to easily run on any platform. Every sub-generator that was explained in Section 5, is implemented as a separate module. This makes PODIGG highly modifiable and composable, because different implementations of sub-generators can easily be added and removed. Furthermore, this flexible composition makes it possible to use real data instead of certain sub-generators. This can be useful for instance when a certain public transport network is already available, and only the trips and connections need to be generated.

We designed PODIGG to be highly configurable to adjust the characteristics of the generated output across different levels, and to define a certain *seed* parameter for producing deterministic output.

All sub-generators store generated data in-memory, using list-based data structures directly corresponding to the GTFS format. This makes GTFS serialization a simple and efficient process. Table 1 shows the GTFS files that are generated by the different PODIGG modules. This table does not contain references to the region and edges generator, because they are only used internally as prerequisites to the later steps. All required files are created to have a valid GTFS dataset. Next to that, the optional file for exceptional service dates is created. Furthermore, `delays.txt` is created, which is not part of the GTFS specification. It is an extension we provide in order to serialize delay information about each connection in a trip. These delays are represented in a CSV file containing columns for referring to a connection in a trip, and contains delay values in milliseconds and a certain reason per connection arrival and departure, as shown in Listing 1.

File	Generator
`agency.txt`	*Constant*
`stops.txt`	Stops
`routes.txt`	Routes
`trips.txt`	Trips
`stop_times.txt`	Trips
`calendar.txt`	Trips
`calendar_dates.txt`	Trips
`delays.txt`	Trips

Table 1

The GTFS files that are written by PODIGG, with their corresponding sub-generators that are responsible for generating the required data. The files in bold refer to files that are required by the GTFS specification.

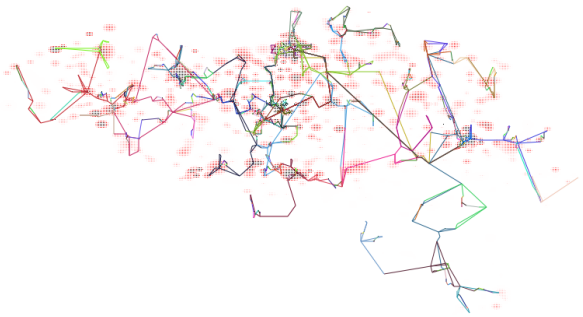

Figure 6. Visualization of a generated public transport network based on Belgium's population distribution. Each route has a different color, and dark route colors indicate more frequent trips over them than light colors. The population distribution is illustrated for each cell as a scale going from white (low), to red (medium) and black (high). Full image: `https://linkedsoftwaredependencies.org/raw/podigg/gen.png`

In order to easily observe the network structure in the generated datasets, PODIGG will always produce a figure accompanying the GTFS dataset. Figure 6 shows an example of such a visualization.

Because the generation of large datasets can take a long time depending on the used parameters, PODIGG has a logging mechanism, which provides continuous

[8] `http://vocab.gtfs.org/terms`
[9] `http://semweb.mmlab.be/ns/linkedconnections`
[10] `https://github.com/PoDiGG/podigg`
[11] `https://www.npmjs.com/package/podigg`
[12] `https://hub.docker.com/r/podigg/podigg/`

```
trip_id,stop_sequence,delay_departure,delay_arrival,delay_departure_reason         ,delay_arrival_reason
100_4  ,0             ,0              ,1405754      ,                              ,td:RepairWork
100_6  ,0             ,0              ,1751671      ,                              ,td:BrokenDownTrain
100_6  ,1             ,1751671        ,1553820      ,td:BrokenDownTrain            ,td:BrokenDownTrain
100_7  ,0             ,2782295        ,0            ,td:TreeAndVegetationCuttingWork,
```

Listing 1: Sample of a `delays.txt` file in a GTFS dataset

feedback to the user about the current status and progress of the generator.

Finally, PODIGG provides the option to derive realistic public transit queries over the generated network, aimed at testing the load of route planning systems. This is done by iteratively selecting two random stops weighed by their size and choosing a random starting time based on the same time distribution as discussed in Section 5.5. This is serialized to a JSON format[13] that was introduced for benchmarking the Linked Connections route planner [10].

6.2. PODIGG-LC

PODIGG-LC is an extension of PODIGG, that outputs data in Turtle/RDF using the ontologies shown in Figure 1. It is also implemented in JavaScript using Node.js, and available under an open license on GitHub[14]. PODIGG-LC is also available as a Node module on NPM[15] and as a Docker image on Docker Hub[16]. For this, we extended the GTFS2LC tool[17] that is able to convert GTFS datasets to RDF using the Linked Connections and GTFS ontologies. The original tool serializes a minimal subset of the GTFS data, aimed at being used for Linked Connections route planning over connections. Our extension also serializes trip, station and route instances, with their relevant interlinking. Furthermore, our GTFS extension for representing delays is also supported, and is serialized using a new Linked Connections Delay ontology[18] that we created.

6.3. Configuration

PODIGG accepts a wide range of parameters that can be used to configure properties of the different sub-generators. Table 2 shows an overview of the parameters, grouped by each sub-generator. PODIGG and PODIGG-LC accept these parameters[19] either in a JSON configuration file or via environment variables. Both PODIGG and PODIGG-LC produce deterministic output for identical sets of parameters, so that datasets can easily be reproduced given the configuration. The `seed` parameter can be used to introduce pseudo-randomness into the output.

7. Evaluation

In this section, we discuss our evaluation of PODIGG. We first evaluate the realism of the generated datasets using a constant seed by comparing its coherence to real datasets, followed by a more detailed realism evaluation of each sub-generator using distance functions. Finally, we provide an indicative efficiency and scalability evaluation of the generator and discuss practical dataset sizes. All scripts that were used for the following evaluation can be found on GitHub[20]. Our experiments were executed on a 64-bit Ubuntu 14.04 machine with 128 GB of memory and a 24-core 2.40 GHz CPU.

7.1. Coherence

Metric In order to determine how closely synthetic RDF datasets resemble their real-world variants in terms of *structuredness*, the coherence metric [15] can be used. In RDF dataset generation, the goal is to reach a level of structuredness similar to real datasets. As mentioned before in Section 2, many synthetic datasets have a level of structuredness that is higher than their real-world counterparts. Therefore, our coherence evaluation should indicate that our generator is not subject to the same problem. We have implemented a command-line tool[21] to calculate the coherence value for any given input dataset.

[13] https://github.com/linkedconnections/benchmark-belgianrail#transit-schedules
[14] https://github.com/PoDiGG/podigg-lc
[15] https://www.npmjs.com/package/podigg-lc
[16] https://hub.docker.com/r/podigg/podigg-lc/
[17] https://github.com/PoDiGG/gtfs2lc
[18] http://semweb.datasciencelab.be/ns/linked-connections-delay/
[19] https://github.com/PoDiGG/podigg#parameters
[20] https://github.com/PoDiGG/podigg-evaluate
[21] https://github.com/PoDiGG/graph-coherence

	Name	Default Value	Description
	seed	1	The random seed
Region	region_generator	isolated	Name of a region generator. (isolated, noisy or region)
	lat_offset	0	Value to add with all generated latitudes
	lon_offset	0	Value to add with all generated longitudes
	cells_per_latlon	100	How many cells go in 1 latitude/longitude
Stops	stops	600	How many stops should be generated
	min_station_size	0.01	Minimum cell population value for a stop to form
	max_station_size	30	Maximum cell population value for a stop to form
	start_stop_choice_power	4	Power for selecting large population cells as stops
	min_interstop_distance	1	Minimum distance between stops in number of cells
	factor_stops_post_edges	0.66	Factor of stops to generate after edges
	edge_choice_power	2	Power for selecting longer edges to generate stops on
	stop_around_edge_choice_power	4	Power for selecting large population cells around edges
	stop_around_edge_radius	2	Radius in number of cells around an edge to select points
Edges	max_intracluster_distance	100	Maximum distance between stops in one cluster
	max_intracluster_distance_growthfactor	0.1	Power for clustering with more distant stops
	post_cluster_max_intracluster_distancefactor	1.5	Power for connecting a stop with multiple stops
	loosestations_neighbourcount	3	Neighbours around a loose station that should define its area
	loosestations_max_range_factor	0.3	Maximum loose station range relative to the total region size
	loosestations_max_iterations	10	Maximum iteration numberto try to connect one loose station
	loosestations_search_radius_factor	0.5	Loose station neighbourhood size factor
Routes	routes	1000	The number of routes to generate
	largest_stations_fraction	0.05	The fraction of stops to form routes between
	penalize_station_size_area	10	The area in which stop sizes should be penalized
	max_route_length	10	Maximum number of edges for a route in the macro-step
	min_route_length	4	Minimum number of edges for a route in the macro-step
Connections	time_initial	0	The initial timestamp (ms)
	time_final	24 * 3600000	The final timestamp (ms)
	connections	30000	Number of connections to generate
	stop_wait_min	60000	Minimum waiting time per stop
	stop_wait_size_factor	60000	Waiting time to add multiplied by station size
	route_choice_power	2	Power for selecting longer routes for connections
	vehicle_max_speed	160	Maximum speed of a vehicle in km/h
	vehicle_speedup	1000	Vehicle speedup in km/(h^2)
	hourly_weekday_distribution	... [1]	Hourly connection chances for weekdays
	hourly_weekend_distribution	... [1]	Hourly connection chances for weekend days
	delay_chance	0	Chance for a connection delay
	delay_max	3600000	Maximum delay
	delay_choice_power	1	Power for selecting larger delays
	delay_reasons	... [2]	Default reasons and chances for delays
	delay_reduction_duration_fraction	0.1	Maximum part of connection duration to subtract for delays
Queryset	start_stop_choice_power	4	Power for selecting large starting stations
	query_count	100	The number of queries to generate
	time_initial	0	The initial timestamp
	time_final	24 * 3600000	The final timestamp
	max_time_before_departure	3600000	The maximum time until a query should be started
	hourly_weekday_distribution	... [1]	Chance for each hour to have a connection on a weekday
	hourly_weekend_distribution	... [1]	Chance for each hour to have a connection on a weekend day

Table 2

Configuration parameters for the different sub-generators. Time values are represented in milliseconds. [1] Time distributions are based on public route planning logs [9]. [2] Default delays are based on the Transport Disruption ontology (`https://transportdisruption.github.io/`).

Results When measuring the coherence of the Belgian railway, buses and Dutch railway datasets, we discover high values for both the real-world datasets and the synthetic datasets, as can be seen in Table 3. These nearly maximal values indicate that there is a very high level of structuredness in these real-world datasets. Most instances have all the possible values, unlike most typical RDF datasets, which have values around or below 0.6 [15]. That is because of the very specialized nature of this dataset, and the fact that they originate from GTFS datasets that have the characteristics of relational databases. Only a very limited number of classes and predicates are used, where almost all instances have the same set of attributes. In fact, these very high coherence values for real-world datasets simplify the process of synthetic dataset generation, as less attention needs to be given to factors that lead to lower levels of structuredness, such as optional attributes for instances. When generating synthetic datasets using PODIGG with the same number of stops, routes and connections for the three gold standards, we measure very similar coherence values, with differences ranging from 0.08% to 1.64%. This confirms that PODIGG is able to create datasets with the same high level of structuredness to real datasets of these types, as it inherits the relational database characteristics from its GTFS-centric mimicking algorithm.

7.2. Distance to Gold Standards

While the coherence metric is useful to compare the level of structuredness between datasets, it does not give any detailed information about how *real* synthetic datasets are in terms of their *distance* to the real datasets. In this case, we are working with public transit feeds with a known structure, so we can look at the different datasets aspects in more detail. More specifically, we start from real geographical areas with their population distributions, and consider the distance functions between stops, edges, routes and trips for the synthetic and gold standard datasets. In order to check the applicability of PODIGG to different transport types and geographical areas, we compare with the gold standard data of the Belgian railway, the Belgian buses and the Dutch railway. The scripts that were used to derive these gold standards from real-world data can be found on GitHub[22].

[22] https://github.com/PoDiGG/population-density-generator

In order to construct distance functions for the different generator elements, we consider several helper functions. The function in Equation 5 is used to determine the closest element in a set of elements B to a given element a, given a distance function f. The function in Equation 6 calculates the distance between all elements in A and all elements in B, given a distance function f. The computational complexity of χ is $O(\|B\| \cdot \kappa(f))$, where $\kappa(f)$ is the cost for one distance calculation for f. The complexity of Δ then becomes $O(\|A\| \cdot \|B\| \cdot \kappa(f))$.

$$\chi(a, B, f) \coloneqq \arg\min_{b \in B} f(a, b) \tag{5}$$

$$\Delta(A, B, f) \coloneqq \frac{\sum_{a \in A} f(a, \chi(a, B, f)) + \sum_{b \in B} f(b, \chi(b, A, f))}{\|A\| + \|B\|} \tag{6}$$

Stops Distance For measuring the distance between two sets of stops S_1 and S_2, we introduce the distance function from Equation 7. This measures the distance between every possible pair of stops using the Euclidian distance function d. Assuming a constant execution time for $\kappa(d)$, the computational complexity for Δ_s is $O(\|S_1\| \cdot \|S_2\|)$.

$$\Delta_s(S_1, S_2) \coloneqq \Delta(S_1, S_2, d) \tag{7}$$

Edges Distance In order to measure the distance between two sets of edges E_1 and E_2, we use the distance function from Equation 8, which measures the distance between all pairs of edges using the distance function d_e. This distance function d_e, which is introduced in Equation 9, measures the Euclidian distance between the start and endpoints of each edge, and between the different edges, weighed by the length of the edges. The constant 1 in Equation 9 is to ensure that the distance between two edges that have an equal length, but exist at a different position, is not necessarily zero. The computational cost of d_e can be considered as a constant, so the complexity of Δ_e becomes $O(\|E_1\| \cdot \|E_2\|)$.

$$\Delta_e(E_1, E_2) \coloneqq \Delta(E_1, E_2, d_e) \tag{8}$$

	Belgian railway	**Belgian buses**	**Dutch railway**
Real	0.9845	0.9969	0.9862
Synthetic	0.9879	0.9805	0.9870
Difference	0.0034	0.0164	0.0008

Table 3

Coherence values for three gold standards compared to the values for equivalent synthetic variants.

$$\begin{aligned} d_\mathrm{e}(e_1, e_2) \coloneqq &\min(d(e_1^\mathrm{from}, e_2^\mathrm{from}) + d(e_1^\mathrm{to}, e_2^\mathrm{to}), \\ &d(e_1^\mathrm{from}, e_2^\mathrm{to}) + d(e_1^\mathrm{to}, e_2^\mathrm{from})) \\ &\cdot (d(e_1^\mathrm{from}, e_1^\mathrm{to}) - d(e_2^\mathrm{from}, e_2^\mathrm{to}) + 1) \end{aligned} \tag{9}$$

Routes Distance Similarly, the distance between two sets of routes R_1 and R_2 is measured in Equation 10 by applying Δ for the distance function d_r. Equation 11 introduces this distance function d_r between two routes, which is calculated by considering the edges in each route and measuring the distance between those two sets using the distance function Δ_e from Equation 8. By considering the maximum amount of edges per route as e_max, the complexity of d_r becomes $O(e_\mathrm{max}^2)$ This leads to a complexity of $O(\|R_1\| \cdot \|R_2\| \cdot e_\mathrm{max}^2)$ for Δ_r.

$$\Delta_\mathrm{r}(R_1, R_2) \coloneqq \Delta(R_1, R_2, d_\mathrm{r}) \tag{10}$$

$$d_\mathrm{r}(r_1, r_2) \coloneqq \Delta_\mathrm{e}(r_1^\mathrm{edges}, r_2^\mathrm{edges}) \tag{11}$$

Connections Distance Finally, we measure the distance between two sets of connections C_1 and C_2 using the function from Equation 12. The distance between two connections is measured using the function from Equation 13, which is done by considering their respective temporal distance weighed by a constant ϵ [23], and their geospatial distance using the edge distance function d_e. The complexity of time calculation in d_c can be considered being constant, which makes it overall complexity $O(e_\mathrm{max}^2)$. For Δ_c, this leads to a complexity of $O(\|C_1\| \cdot \|C_2\| \cdot e_\mathrm{max}^2)$.

$$\Delta_\mathrm{c}(C_1, C_2) \coloneqq \Delta(C_1, C_2, d_\mathrm{c}) \tag{12}$$

[23] When serializing time in milliseconds, we set ϵ to 60000.

$$\begin{aligned} d_\mathrm{c}(c_1, c_2) \coloneqq &((c_1^\mathrm{departureTime} - c_2^\mathrm{departureTime}) \\ &+ (c_1^\mathrm{arrivalTime} - c_2^\mathrm{arrivalTime})/\epsilon) \\ &+ d_\mathrm{e}(c_1, c_2) \end{aligned} \tag{13}$$

Computability When using the introduced functions for calculating the distance between stops, edges, routes or connections, execution times can become long for a large number of elements because of their large complexity. When applying these distance functions for realistic numbers of stops, edges, routes and connections, several optimizations should be done in order to calculate these distances in a reasonable time. A major contributor for these high complexities is χ for finding the closest element from a set of elements to a given element, as introduced in Equation 5. In practice, we only observed extreme execution times for the respective distance between routes and connections. For routes, we implemented an optimization, with the same worst-case complexity, that indexes routes based on their geospatial position, and performs radial search around each route when the closest one from a set of other routes should be found. For connections, we consider the linear time dimension when performing binary search for finding the closest connection from a set of elements.

Metrics In order to measure the realism of each generator phase, we introduce a *realism* factor ρ for each phase. These values are calculated by measuring the distance from randomly generated elements to the gold standard, divided by the distance from the actually generated elements to the gold standard, as shown below for respectively stops, edges, routes and connections. We consider these randomly generated elements having the lowest possible level of realism, so we use these as a weighting factor in our realism values.

$$\rho_\mathrm{s}(S_\mathrm{rand}, S_\mathrm{gen}, S_\mathrm{gs}) \coloneqq \Delta_\mathrm{s}(S_\mathrm{rand}, S_\mathrm{gs})/\Delta_\mathrm{s}(S_\mathrm{gen}, S_\mathrm{gs})$$

$$\rho_\mathrm{e}(E_\mathrm{rand}, E_\mathrm{gen}, E_\mathrm{gs}) \coloneqq \Delta_\mathrm{e}(E_\mathrm{rand}, E_\mathrm{gs})/\Delta_\mathrm{e}(E_\mathrm{gen}, E_\mathrm{gs})$$

$$\rho_\mathrm{r}(R_\mathrm{rand}, R_\mathrm{gen}, R_\mathrm{gs}) \coloneqq \Delta_\mathrm{r}(R_\mathrm{rand}, R_\mathrm{gs})/\Delta_\mathrm{r}(R_\mathrm{gen}, R_\mathrm{gs})$$

$$\rho_\mathrm{c}(C_\mathrm{rand}, C_\mathrm{gen}, C_\mathrm{gs}) \coloneqq \Delta_\mathrm{c}(C_\mathrm{rand}, C_\mathrm{gs})/\Delta_\mathrm{c}(C_\mathrm{gen}, C_\mathrm{gs})$$

Results We measured these realism values with gold standards for the Belgian railway, the Belgian buses and the Dutch railway. In each case, we used an optimal set of parameters[24] to achieve the most realistic generated output. Table 4 shows the realism values for the three cases, which are visualized in Figures 7 to 10. Each value is larger than 1, showing that the generator at least produces data that is closer to the gold standard, and is therefore more realistic. The realism for edges is in each case very large, showing that our algorithm produces edges that are very similar to actual the edge placement in public transport networks according to our distance function. Next, the realism of stops is lower, but still sufficiently high to consider it as realistic. Finally, the values for routes and connections show that these sub-generators produce output that is closer to the gold standard than the random function according to our distance function. Routes achieve the best level of realism for the Belgian railway case. For this same case, the connections are however only slightly closer to the gold standard than random placement, while for the other cases the realism is more significant. All of these realism values show that PODIGG is able to produce realistic data for different regions and different transport types.

7.3. Performance

Metrics While performance is not the main focus of this work, we provide an indicative performance evaluation in this section in order to discover the bottlenecks and limitations of our current implementation that could be further investigated and resolved in future work. We measure the impact of different parameters on the execution times of the generator. The three main parameters for increasing the output dataset size are the number of stops, routes and connections. Because the number of edges is implicitly derived from the number of stops in order to reach a connected network, this can not be configured directly. In this section, we start from a set of parameters that produces realistic output data that is similar to the Belgian railway case. We let the value for each of these parameters increase to see the evolution of the execution times and memory usage.

[24] `https://github.com/PoDiGG/podigg-evaluate/blob/master/bin/evaluate.js`

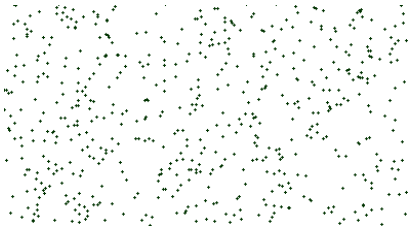

Fig. a: Random

Fig. b: Generated

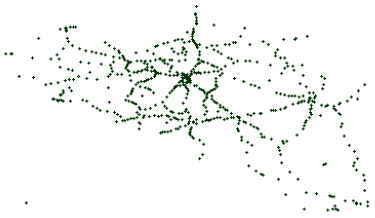

Fig. c: Gold standard

Figure 7. Stops for the Belgian railway case.

Results Figure 11 shows a linear increase in execution times when increasing the routes or connections. The execution times for stops do however increase much faster, which is caused by the higher complexity of networks that are formed for many stops. The used algorithms for producing this network graph proves to be the main bottleneck when generating large networks. Networks with a limited size can however be generated quickly, for any number of routes and connections. The memory usage results from Figure 12 also show a linear increase, but now the increase for routes and connections is higher than for the stops parameter. These figures show that stops generation is a more CPU intensive process than routes and connections generation. These last two are able to make better usage of the available memory for speeding up the process.

7.4. Dataset size

An important aspect of dataset generation is its ability to output various dataset sizes. In PODIGG, different options are available for tweaking these sizes. Increasing the time range parameter within the generator increases

	Belgian railway	**Belgian buses**	**Dutch railway**
Stops	5.5490	297.0888	4.0017
Edges	147.4209	1633.4693	318.4131
Routes	2.2420	1.6094	1.3095
Connections	1.0451	1.5006	1.3017

Table 4

Realism values for the three gold standards in case of the different sub-generators, respectively calculated for the stops ρ_s, edges ρ_e, routes ρ_r and connections ρ_c.

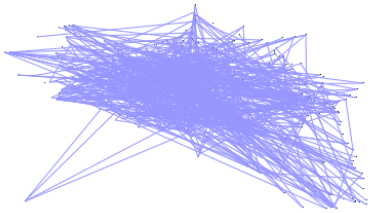

Fig. a: Random

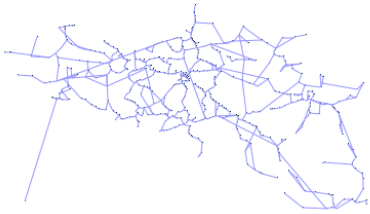

Fig. b: Generated

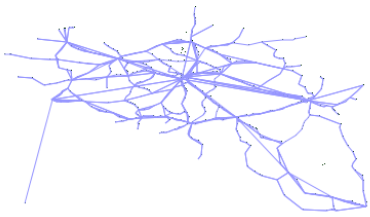

Fig. c: Gold standard

Figure 8. Edges for the Belgian railway case.

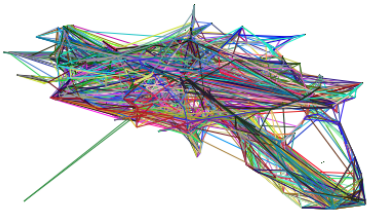

Fig. a: Random

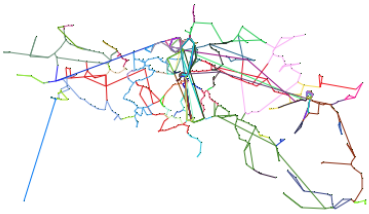

Fig. b: Generated

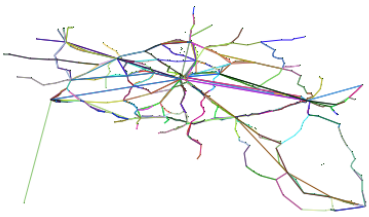

Fig. c: Gold standard

Figure 9. Routes for the Belgian railway case.

the number of connections while the number of stops and routes will remain the same. When enlarging the geographical area over the same period of time, the opposite is true. As a rule of thumb, based on the number of triples per connection, stops and routes, the total number of generated triples per dataset is approximately $7 \cdot \#connections + 6 \cdot \#stops + \#routes$. For the Belgian railway case, containing 30,011 connections over a period of 9 months, with 583 stops and 362 routes, this would theoretically result in 213,937 triples. In practice, we reach 235,700 triples when running with these parameters, which is slightly higher because of the other triples that are not taken into account for this simplified formula, such as the ones for trips, stations and delays.

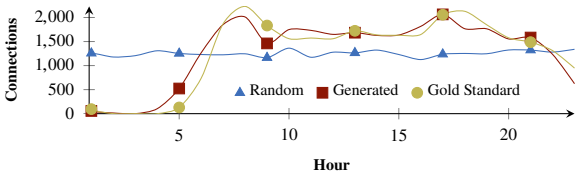

Figure 10. Connections per hour for the Belgian railway case.

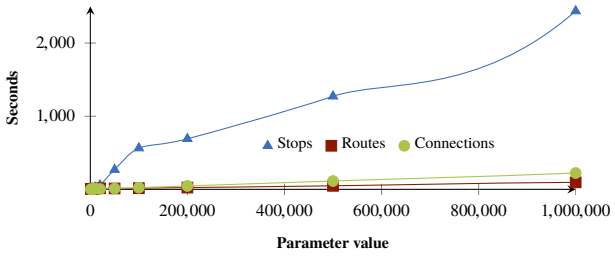

Figure 11. Execution times when increasing the number of stops, routes or connections.

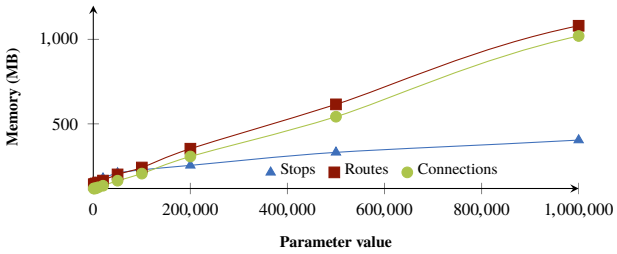

Figure 12. Memory usage when increasing the number of stops, routes or connections.

8. Discussion

In this section, we discuss the main characteristics, the usage within benchmarks and the limitations of this work. Finally, we mention several PODIGG use cases.

8.1. Characteristics

Our main research question on how to generate realistic synthetic public transport networks has been answered by the introduction of the mimicking algorithm from Section 5, based on commonly used practises in transit network design. This is based on the accepted hypothesis that the population distribution of an area is correlated with its transport network design and scheduling. We measured the realism of the generated datasets using the coherence metric in Section 7.1 and more fine-grained realism metrics for different public transport aspects in Section 7.2.

PODIGG, our implementation of the algorithm, accepts a wide range of parameters to configure the mimicking algorithm. PODIGG and PODIGG-LC are able to output the mimicked data respectively as GTFS and RDF datasets, together with a visualization of the generated transit network. Our system can be used without requiring any extensive setup or advanced programming skills, as it consists of simple command lines tools that can be invoked with a number of optional parameters to configure the generator. Our system is proven to be generalizable to other transport types, as we evaluated PODIGG for the bus and train transport type, and the Belgium and Netherlands geospatial regions in Section 7.2.

8.2. Usage within Benchmarks

A synthetic dataset generator, which is one of the main contributions of this work, forms an essential aspect of benchmarks for (RDF) data management systems [2,19]. Prescribing a concrete benchmark that includes the evaluation of tasks is out of scope. However, to provide a guideline on how our dataset generator can be used as part of a benchmark, we relate the primary elements of public transport datasets to *choke points* in data management systems, i.e., key technical challenges in these system. Below, we list choke points related to *storage* and *querying* within data management systems and route planning systems. For each choke point, we introduce example tasks to evaluate them in the context of public transport datasets. The querying choke points are inspired by the choke points identified by Petzka et. al. [27] for faceted browsing.

1. Storage of entities.
   (a) Storage of stops, stations, connections, routes, trips and delays.
2. Storage of links between entities.
   (a) Storage of stops per station.
   (b) Storage of connections for stops.
   (c) Storage of the next connection for each connection.
   (d) Storage of connections per trip.
   (e) Storage of trips per route.
   (f) Storage of a connection per delay.
3. Storage of literals.
   (a) Storage of latitude, longitude, platform code and code of stops.
   (b) Storage of latitude, longitude and label of stations.
   (c) Storage of delay durations.
   (d) Storage of the start and end time of connections.
4. Storage of sequences.
   (a) Storage of sequences of connections.
5. Find instances by property value.
   (a) Find latitude, longitude, platform code or code by stop.
   (b) Find station by stop.
   (c) Find country by station.

   (d) Find latitude, longitude, or label by station.
   (e) Find delay by connection.
   (f) Find next connection by connection.
   (g) Find trip by connection.
   (h) Find route by connection.
   (i) Find route by trip.
6. Find instances by inverse property value.
   (a) *Inverse of examples above.*
7. Find instances by a combination of properties values.
   (a) Find stops by geospatial location.
   (b) Find stations by geospatial location.
8. Find instances for a certain property path with a certain value.
   (a) Find the delay value of the connection after a given connection.
   (b) Find the delay values of all connections after a given connection.
9. Find instances by inverse property path with a certain value.
   (a) Find stops that are part of a certain trip that passes by the stop at the given geospatial location.
10. Find instances by class, including subclasses.
   (a) Find delays of a certain class.
11. Find instances with a numerical value within a certain interval.
   (a) Find stops by latitude or longitude range.
   (b) Find stations by latitude or longitude range.
   (c) Find delays with durations within a certain range.
12. Find instances with a combination of numerical values within a certain interval.
   (a) Find stops by geospatial range.
   (b) Find stations by geospatial range.
13. Find instances with a numerical interval by a certain value for a certain property path.
   (a) Find connections that pass by stops in a given geospatial range.
14. Find instances with a numerical interval by a certain value.
   (a) Find connections that occur at a certain time.
15. Find instances with a numerical interval by a certain value for a certain property path.
   (a) Find trips that occur at a certain time.
16. Find instances with a numerical interval by a certain interval.
   (a) Find connections that occur during a certain time interval.
17. Find instances with a numerical interval by a certain interval for a certain property path.
   (a) Find trips that occur during a certain time interval.
18. Find instances with numerical intervals by intervals with property paths.
   (a) Find connections that occur during a certain time interval with stations that have stops in a given geospatial range.
   (b) Find trips that occur during a certain time interval with stops in a given geospatial range.
   (c) Plan a route that gets me from stop A to stop B starting at a certain time.

This list of choke points and tasks can be used as a basis for benchmarking spatiotemporal data management systems using public transport datasets. For example, SPARQL queries can be developed based on these tasks and executed by systems using a public transport dataset. For the benchmarking with these tasks, it is essential that the used datasets are realistic, as discussed in Section 7.2. Otherwise, certain choke points may not resemble the real world. For example, if an unrealistic dataset would contain only a single trip that goes over all stops, then finding a route between two given stops could be unrealistically simple.

8.3. Limitations and Future Work

In this section, we discuss the limitations of the current mimicking algorithm and its implementation, together with further research opportunities.

Memory Usage The sequential steps in the presented mimicking algorithm require persistence of the intermediary data that is generated in each step. Currently, PODIGG is implemented in such a way that all data is kept in-memory for the duration of the generation, until it is serialized. When large datasets need to be generated, this requires a larger amount of memory to be allocated to the generator. Especially for large amounts of routes or connections, where 100 million connections already require almost 10GB of memory to be allocated. While performance was not the primary concern in this work, in future work, improvements could be made in the future. A first possible solution would be to use a memory-mapped database for intermediary data, so that not all data must remain in memory at all times. An alternative solution would be to modify the mimicking

process to a streaming algorithm, so that only small parts of data need to be kept in memory for datasets of any size. Considering the complexity of transit networks, a pure streaming algorithm might not be feasible, because route design requires knowledge of the whole network. The generation of connections however, could be adapted so that it works as a streaming algorithm.

Realism We aimed to produce realistic transit feeds by reusing the methodologies learned in public transit planning. Our current evaluation compares generated output to real datasets, as no similar generators currently exist. When similar generation algorithms are introduced in the future, this evaluation can be extended to compare their levels of realism. Our results showed that all subgenerators, except for the trips generator, produced output with a high realism value. The trips are still closer to real data than a random generator, but this can be further improved in future work. This can be done by for instance taking into account network capacities [20] on certain edges when instantiating routes as trips, because we currently assume infinite edge capacities, which can result in a large amount of connections over an edge at the same time, which may not be realistic for certain networks. Alternatively, we could include other factors in the generation algorithm, such as the location of certain points of interest, such as shopping areas, schools and tourist spots. In the future, a study could be done to identify and measure the impact of certain points of interest on transit networks, which could be used as additional input to the generation algorithm to further improve the level of realism. Next to this, in order to improve transfer coordination [20], possible transfers between trips should be taken into account when generating stop times. Limiting the network capacity will also lead to natural congestion of networks [8], which should also be taken into account for improving the realism. Furthermore, the total vehicle fleet size [20] should be considered, because we currently assume an infinite number of available vehicles. It is more realistic to have a limited availability of vehicles in a network, with the last position of each vehicle being of importance when choosing the next trip for that vehicle.

Alternative Implementations An alternative way of implementing this generator would be to define declarative dependency rules for public transport networks, based on the work by Pengyue et. al. [25]. This would require a semantic extension to the engine so that is aware of the relevant ontologies and that it can serialize to one or more RDF formats. Alternatively, machine learning techniques could be used to automatically learn the structure and characteristics of real datasets and create similar realistic synthetic datasets [16], or to create variants of existing datasets [31]. The downside of machine learning techniques is however that it is typically more difficult to tweak parameters of automatically learned models when specific characteristics of the output need to be changed, when compared to a manually implemented algorithm. Sensitivity analysis could help to determine the impact of such parameters in order to understand the learned models better.

Streaming Extension Finally, the temporal aspect of public transport networks is useful for the domain of RDF stream processing [12]. Instead of producing single static datasets as output, PODIGG could be adapted to produce RDF streams of connections and delays, where information about stops and routes are part of the background knowledge. Such an extension can become part of a benchmark, such as CityBench [1] and LSBench [24], for assessing the performance of RDF stream processing systems with temporal and geospatial capabilities.

8.4. PODIGG In Use

PODIGG and PODIGG-LC have been developed for usage within the HOBBIT platform. This platform is being developed within the HOBBIT project and aims to provide an environment for benchmarking RDF systems for Big Linked Data. The platform provides several default dataset generators, including PODIGG, which can be used to benchmark systems.

PODIGG, and its generated datasets are being used in the ESWC Mighty Storage Challenge 2017 and 2018 [18]. The first task of this challenge consists of RDF data ingestion into triple stores, and querying over this data. Because of the temporal aspect of public transport data in the form of connections, PODIGG datasets are fragmented by connection departure time, and transformed to a data stream that can be inserted. In task 4 of this challenge, the efficiency of faceted browsing solutions is benchmarked [27]. In this work, a list of choke points are identified regarding SPARQL queries on triple stores, which includes points such as the selection of subclasses and property-path transitions. Because of the geographical property of public transport data, PODIGG datasets are being used for this benchmark.

Finally, PODIGG is being used for creating virtual transit networks of variable size for the purposes of benchmarking route planning frameworks, such as Linked Connections [10].

9. Conclusions

In this article, we introduced a mimicking algorithm for public transport data, based on steps that are used in real-world transit planning. Our method splits this process into several sub-generators and uses population distributions of an area as input. As part of this article, we introduced PODIGG, a reusable framework that accepts a wide range of parameters to configure the generation algorithm.

Results show that the structuredness of generated datasets are similar to real public transport datasets. Furthermore, we introduced several functions for measuring the realism of synthetic public transport datasets compared to a gold standard on several levels, based on distance functions. The realism was confirmed for different regions and transport types. Finally, the execution times and memory usages were measured when increasing the most important parameters, which showed a linear increase for each parameter, showing that the generator is able to scale to large dataset outputs.

The public transport mimicking algorithm we introduced, with PODIGG and PODIGG-LC as implementations, is essential for properly benchmarking the efficiency and performance of public transport route planning systems under a wide range of realistic, but synthetic circumstances. Flexible configuration allows datasets of any size to be created and various characteristics to be tweaked to achieve highly specialized datasets for testing specific use cases. In general, our dataset generator can be used for the benchmarking of geospatial and temporal RDF data management systems, and therefore lowers the barrier towards more efficient and performant systems.

Acknowledgements

We wish to thank Henning Petzka for his help with discovering issues and providing useful suggestions for the PODIGG implementation. The described research activities were funded by the H2020 project HOBBIT (#688227).

References

[1] M. I. Ali, F. Gao, and A. Mileo. CityBench: a configurable benchmark to evaluate RSP engines using smart city datasets. In *International Semantic Web Conference*, pages 374–389. Springer, 2015.

[2] R. Angles, P. Boncz, J. Larriba-Pey, I. Fundulaki, T. Neumann, O. Erling, P. Neubauer, N. Martinez-Bazan, V. Kotsev, and I. Toma. The Linked Data Benchmark Council: a graph and RDF industry benchmarking effort. *ACM SIGMOD Record*, 43(1):27–31, 2014.

[3] D. F. Barbieri, D. Braga, S. Ceri, E. Della Valle, and M. Grossniklaus. C-SPARQL: SPARQL for continuous querying. In *Proceedings of the 18th international conference on World wide web*, pages 1061–1062. ACM, 2009. doi: 10.1145/1526709.1526856.

[4] H. Bast, M. Hertel, and S. Storandt. Scalable transfer patterns.

[5] R. Battle and D. Kolas. Enabling the geospatial semantic web with parliament and GEOSPARQL. *Semantic Web*, 3(4):355–370, 2012. doi: 10.3233/SW-2012-0065.

[6] T. Berners-Lee. Linked Data, July 2006.

[7] C. Bizer and A. Schultz. The Berlin SPARQL benchmark. *International Journal on Semantic Web and Information Systems*, 5(2):1–24, 2009. doi: 10.4018/jswis.2009040101.

[8] T. Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002. doi: 10.1023/A:1015231126594.

[9] P. Colpaert, A. Chua, R. Verborgh, E. Mannens, R. Van de Walle, and A. Vande Moere. What public transit API logs tell us about travel flows. In *Proceedings of the 6th USEWOD Workshop on Usage Analysis and the Web of Data*, pages 873–878, Apr. 2016. doi: 10.1145/2872518.2891069.

[10] P. Colpaert, A. Llaves, R. Verborgh, O. Corcho, E. Mannens, and R. Van de Walle. Intermodal public transit routing using Linked Connections. In *Proceedings of the 14th International Semantic Web Conference: Posters and Demos*, 2015.

[11] R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1: Concepts and abstract syntax. Recommendation, W3C, Feb. 2014.

[12] E. Della Valle, S. Ceri, F. van Harmelen, and D. Fensel. It's a streaming world! Reasoning upon rapidly changing information. *Intelligent Systems, IEEE*, 24(6):83–89, Nov 2009. doi: 10.1109/MIS.2009.125.

[13] D. Delling, P. Sanders, D. Schultes, and D. Wagner. Engineering route planning algorithms. In *Algorithmics of large and complex networks*, pages 117–139. Springer, 2009. doi: 10.1007/978-3-642-02094-0_7.

[14] J. Dibbelt, T. Pajor, B. Strasser, and D. Wagner. Intriguingly simple and fast transit routing. In *Experimental Algorithms*, pages 43–54. Springer, 2013.

[15] S. Duan, A. Kementsietsidis, K. Srinivas, and O. Udrea. Apples and oranges: a comparison of RDF benchmarks and real RDF datasets. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 145–156. ACM, 2011. doi: 10.1145/2627692.2627697.

[16] J. Eno and C. W. Thompson. Generating synthetic data to match data mining patterns. *IEEE Internet Computing*, 12(3), 2008. doi: 10.1109/MIC.2008.55.

[17] G. Garbis, K. Kyzirakos, and M. Koubarakis. Geographica: A benchmark for geospatial RDF stores. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, editors, *Proceedings of the 12th International Semantic Web Conference*, pages 343–359. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-41338-4_22.

[18] K. Georgala, M. Spasić, M. Jovanovik, H. Petzka, M. Röder, and A.-C. N. Ngomo. MOCHA2017: The mighty storage challenge at ESWC 2017. In *Semantic Web Evaluation Challenge*, pages 3–15. Springer, 2017.

[19] J. Gray. *Benchmark handbook: for database and transaction processing systems*. Morgan Kaufmann Publishers Inc., 1992.

[20] V. Guihaire and J.-K. Hao. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10):1251–1273, 2008. doi: 10.1016/j.tra.2008.03.011.

[21] Y. Guo, Z. Pan, and J. Heflin. LUBM: A benchmark for OWL knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2):158–182, 2005. doi: 10.1016/j.websem.2005.06.005.

[22] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis. Strabon: a semantic geospatial DBMS. In *International Semantic Web Conference*, pages 295–311. Springer, 2012.

[23] D. Le-Phuoc, M. Dao-Tran, J. X. Parreira, and M. Hauswirth. A native and adaptive approach for unified processing of Linked Streams and Linked Data. In *International Semantic Web Conference*, pages 370–388. Springer, 2011.

[24] D. Le-Phuoc, M. Dao-Tran, M.-D. Pham, P. Boncz, T. Eiter, and M. Fink. Linked stream data processing engines: Facts and figures. In *International Semantic Web Conference*, pages 300–312. Springer, 2012.

[25] P. J. Lin, B. Samadi, A. Cipolone, D. R. Jeske, S. Cox, C. Rendon, D. Holt, and R. Xiao. Development of a synthetic data set generator for building and testing information discovery systems. In *Third International Conference on Information Technology: New Generations (ITNG'06)*, pages 707–712. IEEE, 2006. doi: 10.1109/ITNG.2006.51.

[26] M. A. Nascimento, D. Pfoser, and Y. Theodoridis. Synthetic and real spatiotemporal datasets. *IEEE Data Eng. Bull.*, 26(2):26–32, 2003.

[27] H. Petzka, C. Stadler, G. Katsimpras, B. Haarmann, and J. Lehmann. Benchmarking faceted browsing capabilities of triplestores. pages 128–135, 2017. doi: 10.1145/3132218.3132242.

[28] E. Pyrga, F. Schulz, D. Wagner, and C. Zaroliagis. Efficient models for timetable information in public transportation systems. *Journal of Experimental Algorithmics (JEA)*, 12:2–4, 2008. doi: 10.1145/1227161.1227166.

[29] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel. SP^{2}bench: a SPARQL performance benchmark. In *2009 IEEE 25th International Conference on Data Engineering*, pages 222–233. IEEE, 2009. doi: 10.1109/ICDE.2009.28.

[30] M. Spasić, M. Jovanovik, and A. Prat-Pérez. An RDF dataset generator for the social network benchmark with real-world coherence. In I. Fundulaki, A. Krithara, A.-C. Ngonga Ngomo, and V. Rentoumi, editors, *Proceedings of the Workshop on Benchmarking Linked Data*, volume 1700 of *CEUR Workshop Proceedings*, 2016.

[31] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–6. IEEE, 2016.