Studying public transit API query logs to get an indication of travel flows

Pieter Colpaert
pieter.colpaert@ugent.be

Alvin Chua
alvin.chua@asro.kuleuven.be

Ruben Verborgh
ruben.verborgh@ugent.be

Erik Mannens
erik.mannens@ugent.be

Rik Van de Walle
rik.vandewalle@ugent.be

Andrew Vande Moere
andrew.vandemoere
@asro.kuleuven.be

ABSTRACT

In the field of urban planning, researchers need an indication of how people move between cities. Yet, getting statistics of travel flows within public transit systems has proven to be troublesome. In order to get an indication of travel flows between cities in Belgium, we analyzed the query logs of the iRail API, a highly expressive route planning API for the Belgian railways. We were able to study ∼100k to 500k requests for each month between October 2012 and November 2015, which is between 0.56% and 1.66% of the amount of monthly passengers. Using data visualizations, we illustrate the commuting patterns in Belgium and confirm that Brussels, the capital, acts as a central hub. The Flemish region appears to be polycentric, while in the Walloon region, everything converges on Brussels. The findings correspond to the real travel demand, according to experts of the passenger federation Trein Tram Bus. We conclude that query logs of route planners are of high importance in getting an indication of travel flows. However, better travel intentions would be acquirable using dedicated HTTP POST requests.

1. INTRODUCTION

Today, the Belgian railway company (SNCB) takes manual samples of people getting on and off trains in stations. To date, this is still the SNCB's only source to determine people flows through their network[1]. Also within other transit systems the data is tedious to get, certainly when passengers do not have to check-in and check-out. Nonetheless, indications of how people move within a transit network is of uttermost importance for, amongst others, making demand-driven policy decisions concerning mobility.

The *iRail* project[2] started in 2008 to make the data of the Belgian railway accessible for developers. The project offers developers both a data dump for third party apps (since ∼ August 2015) and a route planning API (since ∼ October 2010). The query logs of the latter has been stored since November 2012 (requests before that period got lost due to server migrations).

Getting indications of people flows through public transit networks is a challenge. Yet, *can we not get an indication of these transit flows by studying the query logs of Web-services?*

In this paper, we give a small overview of related work to gather data for flow analysis. Next, we introduce the iRail query logs, which we study by the means of visualizations. Finally, we look at public transit data is being published, and how we would be able to get a better indication of travel flows than just mere query logs.

2. RELATED WORK

Flow analysis is a topic of theoretical interest and practical importance in various disciplines. Flow analysis is conventionally conducted to study spatial dynamics and identify routine patterns in the movement of people. For instance, interest in modelling traffic flows emerged from the strain placed on urban transportation systems during peak hours [10, 6]. Likewise, insight into routine travel patterns is crucial for the conceptualisation of functional urban areas [11, **?**], urban hierarchies [4] and other territorial structures.

Over the past decade, large datasets have become increasingly commonplace due to the proliferation of sensor networks and portable devices like smartphones. Termed "Big Data" due to the large volume of data records that emerge from real-time sensing[**?**], such datasets typically contain information of activities or processes linked to the space and time where they occur. In the domain of "Smart City" research, much has been accomplished with the use of "Big Data" to monitor human movement. Smart card data from public transport systems [10, 3], taxi journeys [6] as well as cellular call data [**?**] have provided planners with new opportunities to develop greater understanding of mobility patterns in urban environments [**?**]. Yet query logs were, to our knowledge, never used for flow analysis so far.

On highly expressive servers, such as SPARQL-endpoints, query logs can play an important role in fine-tuning the underlying datastores [2]. E.g., they determined that in 2011 in the context of the graph query language SPARQL, most of the queries are simple and include few triple patterns and joins. Furthermore, the graph patterns are usually star-shaped and despite triple pattern chains exist, they are generally short. Other, more cacheable interfaces more to the left on the Linked Data Fragments axis [15], also benefit from query logs in order to e.g., know who is using what clients to query the Web, yet the user intention is lost [14]. This was called "a blessing for privacy", yet adding the user query as an HTTP header was suggested.

Visualizations are crucial for making data analysis tangible, so

[1] *http://www.belgianrail.be/nl/corporate/in-de-kijker/comptages_voyageurs_2014.aspx*
[2] *https://hello.irail.be*

that results can be communicated and debated [9]. This is key for our research, since the questions presented are exploratory in nature [8] and can be addressed in many ways. Movement data has received substantial attention from the visualization community and a range of techniques has been developed to support flow analysis [1]. Flow visualizations can be broadly categorized into two groups. Representations that display the complete trajectory and Origin Destination (OD) type representations that take only the start and end locations into account. Since our data solely consists of OD information, we narrow our description to the latter group of visualization techniques. OD type representations are commonly employed when the origin and destination preside over intermediate locations on the path of movement.

The OD Matrix is most common visualization technique based on this type of representation. Rows and columns correspond to locations while cells are coloured to express the volume of flow. OD matrices can be reordered to emphasise connectivity between locations. While this feature is particularly useful for identifying trends and outliers in the data, the tabular display is challenging for lay users to interpret and less expressive of connections between locations than representations based on the node-link metaphor. In recent years, such representations have become increasingly prominent for displaying movement in science as well as popular culture. Locations are depicted as nodes and arrows are drawn between them to indicate movement in a certain direction, allowing for exact and individual relationships to be traced. This representation is detailed and straightforward to understand but becomes cluttered when multiple lines intersect. A number of solutions have been proposed for clutter reduction. Filtering is the most commonly practiced. The same goal can be achieved by clustering lines that have similar properties like common start, end or intermediate locations {**TODO**: Andrienko, 2007 -198}. Another solution is to transform the visual representation. Flow maps {**TODO**: Phan, 2005 -88} reduces clutter by rerouting the arrows to minimize intersection and obtain layouts that emphasize key geographic features such as roads, rivers or coastlines. {**TODO**: Alvin: Will expand abit more to describe why chord diagrams were chosen}

3. THE QUERY LOGS

The iRail API is an XML/JSON HTTP API: each request on *http://api.irail.be* returns a response that can be used directly in an end-user app. It exposes four features:

1. *Route Planning* (*http://api.irail.be/connections/{?from,to,date, time}*) plans a route from one station to another, taking into account a preferred start time.
2. *Next Departures* (*http://api.irail.be/liveboard/{?station,date, time}*) shows the next train departures in a certain station, useful for creating quick overviews.
3. *Trip Status* (*http://api.irail.be/vehicle/{?id,date}*) returns the stations this train passes by on the day of the request.
4. *List of all stations* (*http://api.irail.be/stations/*) gives an overview of all stations Belgian trains can arrive at.

This interface is open source and we were able to suggest changes to the system in which the query logs become open data[3]. As of the 17th of December 2015, iRail publishes the most recent 1,000 requests done on the API as open data, each second, at *http://api.irail.be/logs/*. In order to publish this data, we chose a newline-based format: each new line contains a new query object encoded in JSON. The JSON objects can be read one per one, so UNIX commands such as *grep* work well with this format. We chose this over CSV as we have more flexibility towards changing the object properties in the future. Next, we chose to add an HTTP header with a link to the query log context according to the JSON-LD specification[4] to turn it into Linked Data.

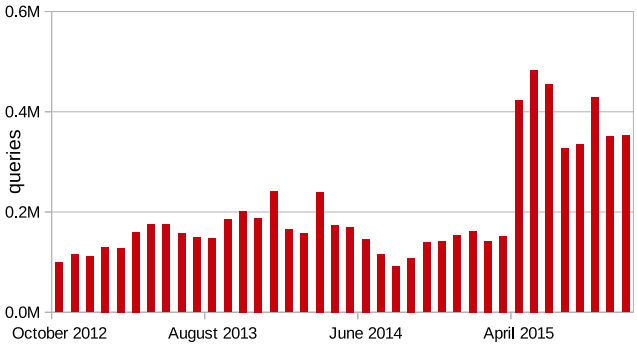


Figure 1: Amount of queries on the route planning part of the iRail API between October 2012 and December 2015. In April 2015 an official app of the Belgian railways was discontinued, which explains the sudden raise of iRail API queries.

Using this data, everyone can build their own historic query log dataset. The server costs for iRail to host this dataset stay limited. Furthermore, this way of publishing data enables research for real-time predictions of travel intentions. Privacy preservation is paramount [12] when dealing with query logs and, therefore, IP addresses have been removed to protect the privacy of our users. Additionally, the integrity of this dataset, without IP addresses, has been vetted by the Belgian Privacy Commission and permission has been granted for these query logs to be published as open data.

We are particularly interested in studying the potential application of *route planning* queries as an information resource for flow analysis due to the input it captures from human users. The other features are less relevant for this purpose as they are generally polled by digital signage providers or status boards, and thus do not reflect direct human demands. The dataset described in this paper is compiled from the *Apache access.log* files generated between October 2012 and December 2015, filtered by "connections" entries. We made a copy of this dataset available in CSV format can be downloaded at {**TODO**: *http://datawijs.be/apilog.tar.gz*}. Each row within this dataset contains:

1. A timestamp of when the request was executed,
2. The request path between two stations and their associated geographic coordinates,
3. The user-agent string cfr. RFC2616,
4. The operating system of the user-agent and
5. A flag stating if the user-agent is a bot.

There has been an increase in the number of queries made between 2012 and 2015. Figure 1 provides a break down of the numbers on a monthly basis. A distinctive increase in the number of queries can be observed in April 2015 when the official route planning service provided by of the Belgian rail company was discontinued[5], resulting in widespread adoption of alternative software applications built on top of the iRail API. Calculated with indicators provided by the Flemish regional government of Belgium[6], we infer a monthly average of 19.3, 19.4, 19,6 and 19.7 million passengers in 2012, 2013, 2014 and 2015 respectively. Correspondingly, the average

[3] *https://github.com/iRail/iRail/pull/138*

[4] *http://www.w3.org/TR/json-ld/*

[5] *https://hello.irail.be/2015/04/22/april-updates/*

[6] *http://www4.vlaanderen.be/sites/svr/Cijfers/Exceltabellen/mobiliteit/vervoersprestaties/personenvervoer/MOBIOPEN006.xls*

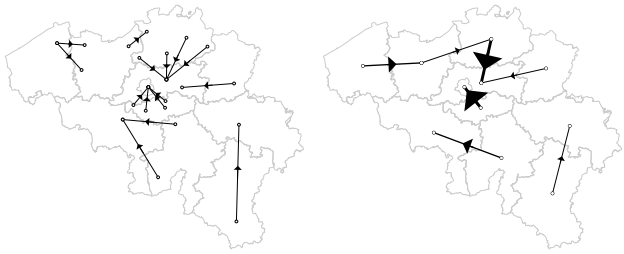
number of iRail queries per month amounts to 0.11, 0.17, 0.15 and 0.33 million respectively. Assuming that each request reciprocates with an intention to travel by rail, our data captures a respective 0.56%, 0.88%, 0.77% and 1.66% of the actual journeys that have occurred.

4. METHOD

Flow analysis involves the comparison of aggregated movement between distinct locations within a specific time frame. In most cases, the analytical objective is to identify trends that occur in journeys made between pairs of locations and allow for the sensitivities to be explored. Data visualization is typically employed to facilitate in this process thus the input data must be transformed into a format that allows for the quantity and directional flow of movement to be visually encoded. The data must undergo several stages of data processing to arrive at this outcome.

4.1 Data Processing

We were specifically interested in understanding how travel during the morning rush hour on weekdays differ from weekends. In this regard, the dataset was filtered to exclude data records created beyond the time range of 06:00 hours and 10:00 hours. Queries made by automated user-agents like search engine bots and data harvesters can be considered as valid traffic but may over represent connections between certain stations. Accordingly, such queries were excluded since they do not explicitly represent human travel behavior. Station names tend to contain be highly inconsistent since any text information can be parsed to the API. This may include invalid character encoding, spelling mistakes and the use of unstandardized abbreviations. We employed an open-source reconciliator (*https://github.com/irail/stations*) to reduce spelling variations and remove data records with missing information. Through this process, station names were also linked to their corresponding geographical coordinates.



(a) (b)

Figure 3: Spatial aggregation based on provincial administrative boundaries diagrammatically represented. (a) Individual data records are depicted geographically with arrows connecting origins to their respective destinations. (b) The result of aggregation is a set of weighted connections between provinces.

Next, the data is aggregated so that province level flow patterns are emphasized. Figure 3 provides a diagrammatic representation of spatial aggregation on the basis of provincial administrative boundaries. To compare the volume of flow between rural provincial areas and urbanized zones, train stations in major cities were aggregated into seperate groups. Simiarly, stations located outside of Belgium were aggregated to distinct group. A diagrammatic breakdown is depicted in Figure 2 where color and shape is used to demarcate each group. The following types of flows are observable from the outcome of aggregation:

- Travel from any provincial station to a major city.
- Travel between any two major cities.
- Travel between any two provincial stations.
- Travel between any international station to a major city.
- Travel between any international station to a provincial station.

4.2 Visualization

The chord diagram is a visualization technique based on the node-link metaphor that arranges nodes along the circumference of a circle. Each node is represented by an arc where its length is proportional to the total volume of incoming and outgoing flows. Chords or curved line segments are drawn to connect nodes. The width at the head or tail of each chord indicates the amount of movement relative frequency of movement from a certain location to another. In our implementation of the chord diagram, nodes are colour coded to indicate individual provinces and the major cities located within their administrative boundaries. Interactive filtering is introduced to simplify the visualization and details are provided in pop-up dialogue when upon mouse-over.

5. RESULTS

In this chapter, we report on the results of studying visualizations with experts of *Trein Tram Bus*[7], a not for profit passenger federation.

5.1 Time Based Analysis

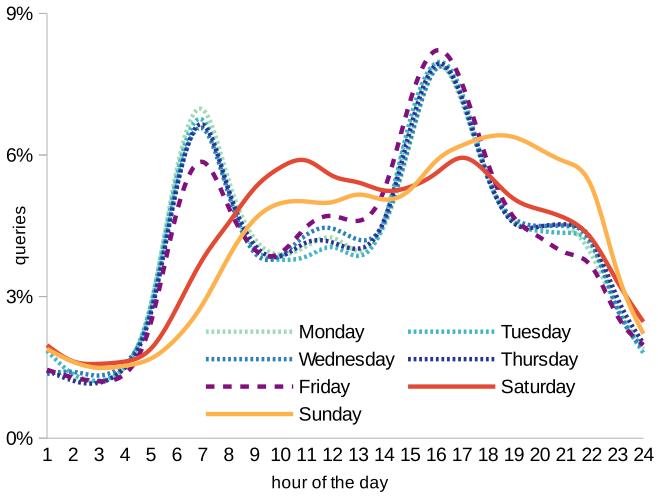


Figure 4: Distribution of iRail route planning API queries on average per day of the week between October 2012 and December 2015

Results from the analysis of this dataset reveal trends in route planning queries that may provide insights into rail travel demand. First, the daily and hourly distribution of queries appear to correspond with known commuting patterns. Figure 4 depicts a break down of the queries that occur, on average, on an hourly basis for each day of the week. Several trends can be observed in this chart. Morning and evening peak periods in particular, are clearly distinguishable. At noon, a small dent is noticable, which may be attributed to part time workers and meetings on location after or before noon. Peak hours on weekdays are also distinctively different from those of weekends. Similarly, the frequency of queries during the evening

[7] *http://treintrambus.be*

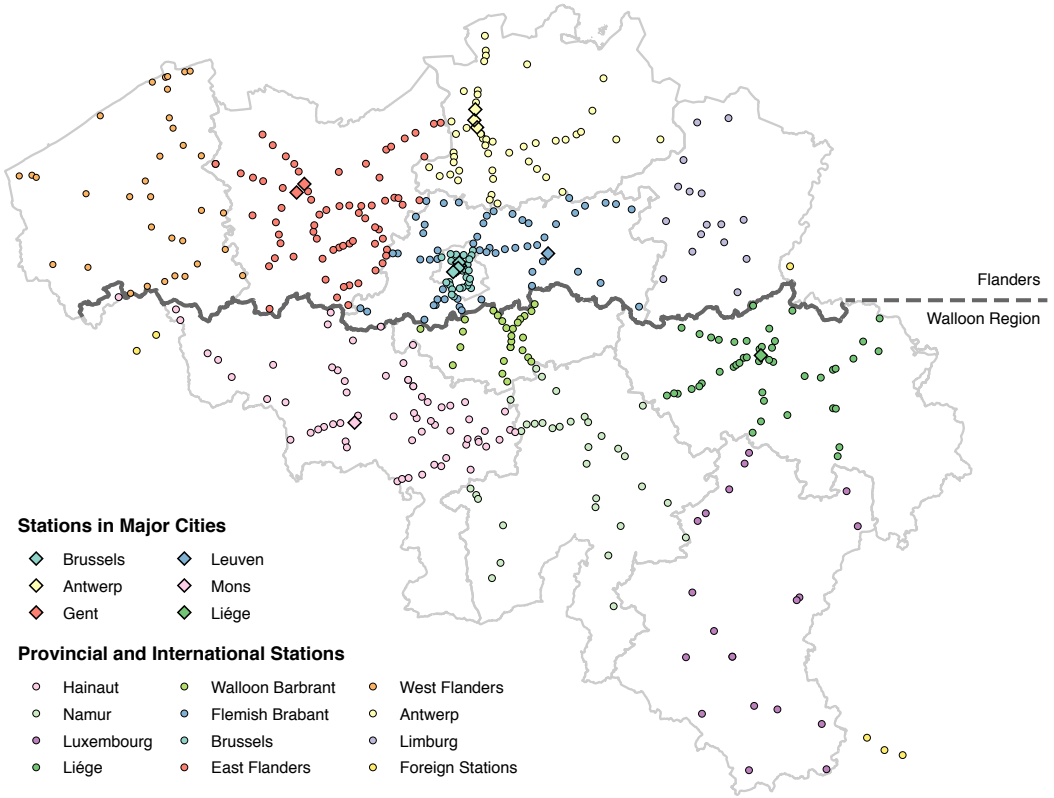


Figure 2: Each station is color coded to indicate the province it will be aggregated to. Stations in major cities as well as stations outside of Belgium are aggregated to seperate groups.

peak period on Fridays appear to be substantially higher than those that occur between Monday and Thursday. The discrepancy maybe explained by students that travel home from their student homes over the weekends. The absence of a clear peak on Saturdays can be attributed to lack of journeys to work queries while the evening peak on Sundays may emerge from students returning to college accommodations.

Moreover, the distribution of queries on public holidays is observed to deviate from that of an average day. Figure 5 provides a breakdown of queries per hour on Christmas eve, December 24th 2015, in comparison to an average Thursday of the same year. As illustrated, the evening peak on Christmas eve occurs earlier than that of an average Thursday.

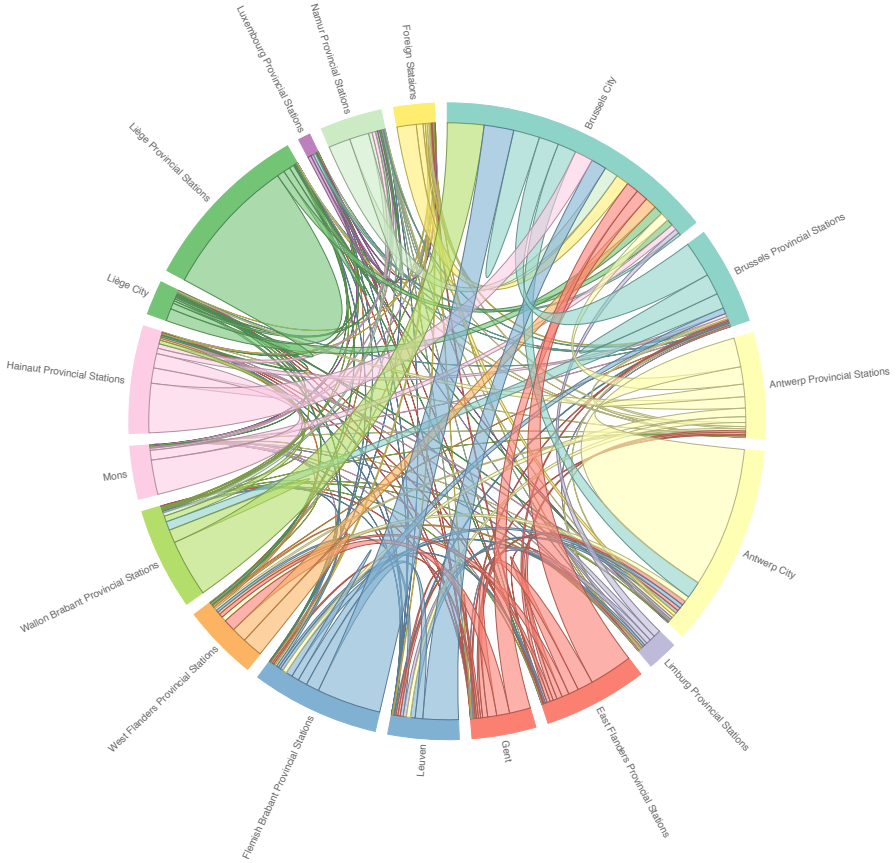
5.2 Structure of Flows in Belgium

The chord diagrams shown in Figure 6 depict the aggregated number of queries made between any two stations on weekdays (See Figure 6a) and weekends (See Figure 6b) respectively. Visualizing our dataset in this manner reveals the complexity of flows on the Belgian rail system and the significance of cities in daily travel. Brussels City stands out as the most distinctive visual element in both figures, indicating its function as the principal centre of rail activity (see Figure 6a). On a regional level, its function as a centre appears to be more distinctive for the Walloon region than Flanders. With exception to Liège city (see Figure 6e), queries from the Walloon region are generally made from provincial stations towards Brussels instead of major cities within their administrative boundaries. Queries from Flanders, on the hand, tend to be distributed among four major cities otherwise known as the Flemish Diamond, a network of metropolitan areas in Belgium comprised of Brussels, Antwerp (see Figure 6b), Gent (see Figure 6c) and Leuven (see Figure 6d). These insights indicate that the structure of flows in Flanders appear to follow a polycentric pattern while that of the Walloon region is relatively monocentric, with Brussels city as a major centre. The difference between both regions appear to correspond with existing measures of population density and degree of urbanisation, providing valuable insight as well as alternative perspectives into the function of cities in rural and urban settings.

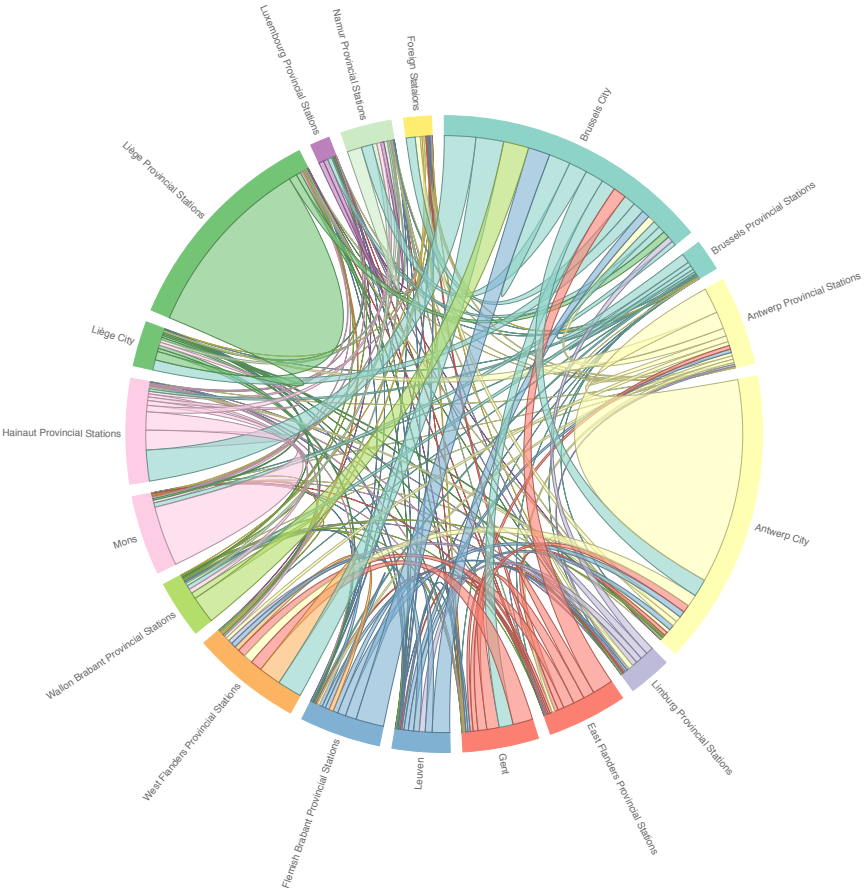
5.3 Counter Intuitive Results

There are several inconsistent patterns in the data that do not correspond with an existing understanding of flows in Belgium. In particular, the large number of trips made between Antwerp city and other stations within the province of Antwerp in the morning, is not observable in other major cities in Flanders. Further investigation led to the identification of an isolated connection originating from Antwerp city towards a suburb, suggesting anomalous usage of the iRail API that were not flagged as bots or a high concentration people who utilise route planning applications on a regular basis. Alternatively the large number of trips may have originated from users who repeatedly request for updates on delayed trains or from a concentration of people who utilise route planning applications on a regular basis. Accordingly, the reverse is identified in Liège where an unusually large connection originating from a suburb towards Liege city. This prompted speculation, as well as discussions, on how such outliers should be treated or what may have led to their existence. As with any form of analysis based on appropriation of data, we acknowledge that more work is required to identify caveats and understand the validity of the findings presented.

6. PUBLISHING TRANSPORT DATA

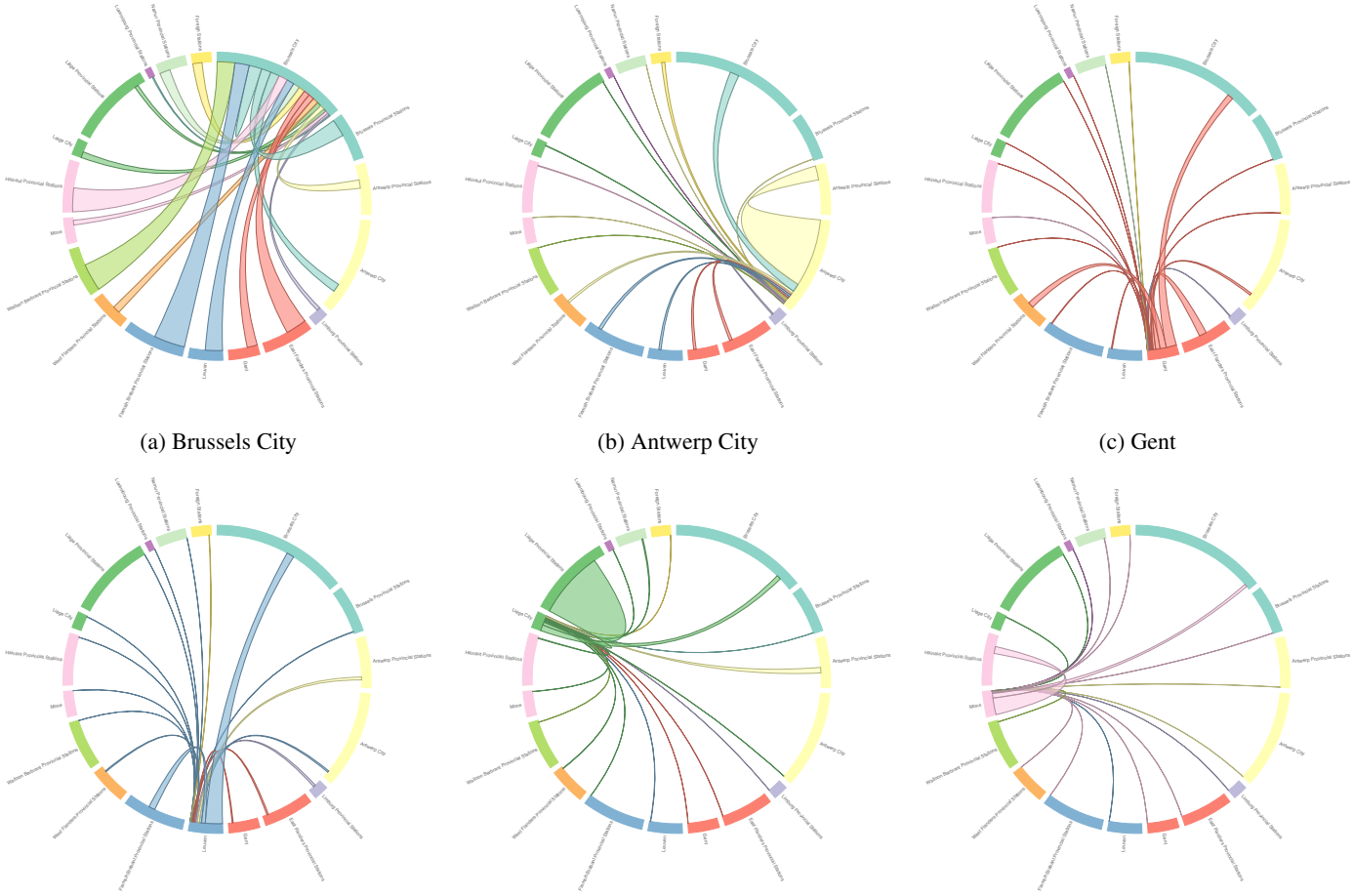


(a) Weekdays



(b) Weekends

Figure 6: Comparison of morning rush hour travel flows on weekdays and weekends.



(a) Brussels City (b) Antwerp City (c) Gent

(d) Leuven (e) Liege City (f) Mons

Figure 7: Comparison of morning rush hour travel flows on weekdays originating from or towards six major cities in Belgium.

The expressiveness of a server affects the way logs can be gathered [14]. We have shown in previous sections that the query logs of a *highly expressive server*, such as the iRail API, are interesting: each request contains an entire query which can be interpreted as a travel intention. The possibility that two requests are exactly the same is low, as there are many URLs which can be requested. Nevertheless, we cannot fully guarantee that each HTTP GET request to the server will trigger a log entry, due to caching mechanisms on the Web [7], which might lead to a false representation.

Hosting a route planning API as the only way to publish transport data comes with three identified limitations, as the server will need to handle the requests from different use cases with different needs:

1. When an application developer would like a *new feature*, such as taking wheelchair accessibility information into account, the feature would have to be implemented on the server of the data publisher.
2. Keeping the server *highly available* is costly, as any question can be asked by anyone for any purpose.
3. Federated querying, which would allow for *intermodal* route planning for route planning APIs, is inexistent up to date.

In order to overcome these limitations, the General Transit Feed Specification (GTFS)[8] can be used. GTFS is a compressed ZIP-file containing a couple of CSV files, describing the rules for when a public transit vehicle will pass by on a certain location. It is supported among all current open-source route planning software systems and it is used in products/apps such as *CityMapper*, *Ally*, *Navitia.io*, *Google Maps* and *Bing Maps*. It succesfully enabled reuse for intermodal travel, engineers can rely on the data dumps even if the servers of the transit agency are offline and there is no limitation to the features that can be implemented. The public transit agency however has no access any longer to an indication of travel demand.

In Figure 8 we illustrate these two options as two extremes, with other options that are yet to be discovered. When a server only allows to set e.g., a departure station and a departure time, then the server cannot log the arrival station, yet the client is still able to plan a route by executing the algorithm on the client-side [5]. We would be able to fully rely on the query logs if the expressivity would be maximal (extreme right) and caching would be turned off. Turning off caching may work in a private setting where you know how much and what kind of queries you can expect, yet on the open Web, caching helps decrease load on servers at peak moments.

7. CONCLUSION AND FUTURE WORK

Firstly, we studied query logs to find travel patterns in Belgium. We conclude that this is up to now, the best representation of travel flows over the Belgian railway network.We studied data which represent a maximum of 1.66% of the passengers. By visualizing the average distribution of requests on workdays and Saturday and Sunday, we were able to recognize the patterns we would expect: a.o.

[8] *https://developers.google.com/transit/gtfs/reference*

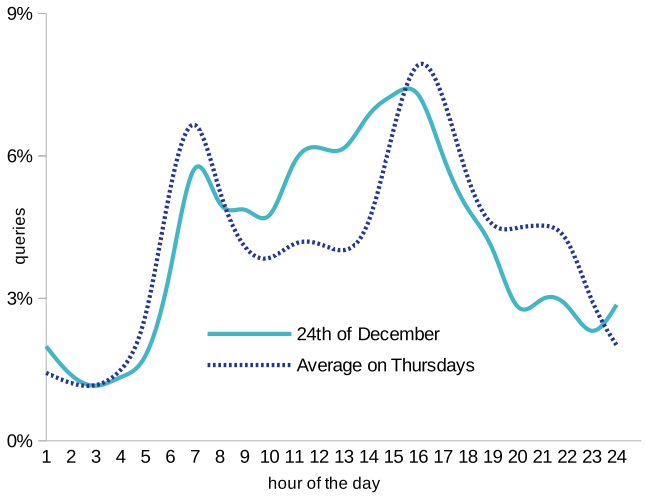


Figure 5: Comparison of route planning queries on December 24th (Christmass Eve) 2015 with the average on a Thursday between October 2012 and December 2015. The illustration shows that people started going home earlier than usual.

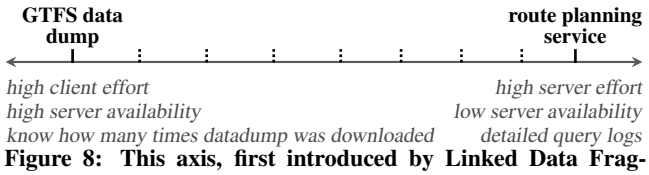


Figure 8: This axis, first introduced by Linked Data Fragments [15], illustrates the possibilities within publishing transport data. Different ticks on the axis illustrate client effort versus server expressivity: on the far left, data dumps offer high server availability, yet the effort needed by data reusers is high, and query logs do not reach the server. Moving to the right, we identify Linked Connections [5] as an in-between solution. On the far right, route planning services require high server effort, leading to detailed query logs.

a morning and evening peak on workdays and a bigger evening peak on Friday and Sunday. More interestingly, we were able to see that the evening peak on Christmass Eve, Thursday the 24th of December 2015, started earlier than average on Thursdays. Visualizing our dataset using the origin and destination reveals the complexity of movement on the Belgian rail system and the significance of cities in weekday travel. Furthermore, we found evidence that Flanders is polycentric, while Walloon traffic is monocentric.

Secondly, we also concluded that there are obvious caveats associated with the use of such data as proxy for actual statistical counts. We identified a couple of gaps:

1. The data only captures an intention of an end-user: it is unsure whether the person actually took the train.
2. Multiple user queries may be needed to cover a travel intention.
3. Caching may mask peak hours, as many of the same requests in one minute is only stored in the logs once.

Nevertheless, we published the query logs as open data at *http://api.irail.be/logs* for other scientists to continue to research possibilities with this dataset. For instance, predicting trip congestion on the basis of this data, looks promising.

Related work [13] suggested to add additional metadata in the HTTP headers in order to track the user's intention. While this is an interesting suggestion, it does not overcome the problem of caching. In order to create a better representation of travel intentions, future work can gather logs by POST requests with the only purpose to gather analytics of travel intention. This is a similar approach than what happens on the Web of Documents today with *Piwik* or *Google Analytics*.

8. REFERENCES

[1] N. Andrienko and G. Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, page 1473871612457601, 2012.

[2] M. Arias, J. D. Fernández, M. A. Martínez-Prieto, and P. de la Fuente. An empirical study of real-world sparql queries. *arXiv preprint arXiv:1103.5043*, 2011.

[3] R. Beecham, J. Wood, and A. Bowerman. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47:5–15, 2014.

[4] W. Christaller. Some considerations of tourism location in europe: the peripheral regions-underdeveloped countries-recreation areas. *Papers in Regional Science*, 12(1):95–105, 1964.

[5] 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*, Oct. 2015.

[6] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2149–2158, 2013.

[7] R. T. Fielding. *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, Irvine, 2000.

[8] M.-J. Kraak. Exploratory visualization. In *Encyclopedia of GIS*, pages 301–307. Springer, 2008.

[9] A. C. Robinson. Collaborative synthesis of visual analytic results. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pages 67–74. IEEE, 2008.

[10] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS one*, 6(1):e15923, 2011.

[11] L. A. Servillo, R. Atkinson, A. P. Russo, L. Sỳkora, C. Demazière, and A. Hamdouch. Town, small and medium sized towns in their functional territorial context, draft final report. 2014.

[12] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1—2):1–174, 2010.

[13] R. Verborgh. The lonesome LOD cloud. In *Proceedings of the 4th USEWOD Workshop on Usage Analysis and the Web of Data*, 2014.

[14] R. Verborgh. DBpedia's Triple Pattern Fragments: Usage patterns and insights. In F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker, and A. Zimmermann, editors, *The Semantic Web: ESWC 2015 Satellite Events*, volume 9341 of *Lecture Notes in Computer Science*, pages 431–442. Springer, June 2015.

[15] R. Verborgh, M. Vander Sande, P. Colpaert, S. Coppens, E. Mannens, and R. Van de Walle. Web-scale querying through Linked Data Fragments. In *Proceedings of the 7th Workshop on Linked Data on the Web*, Apr. 2014.