




Indicators for Measuring Tourist Mobility

Sonia Djebali^(✉), Nicolas Loas^(✉), and Nicolas Travers^(✉) 

Leonard de Vinci, Research Center, 92 916 Paris La Défense, France
{sonia.djebali,nicolas.travers}@devinci.fr, nicolas.loas@edu.devinci.fr

Abstract. Digital traces left by active users on social networks have become a popular means of analyzing tourist behavior. The large amount of data generated by tourists provides a key indicator for understanding their behavior according to various criteria. Analyses of tourists' movement have a crucial role in tourism marketing to build decision-making tools for tourist offices. Those actors are faced with the need to discern tourists' circulation both quantitatively and qualitatively. In this paper, we propose a measure to capture tourist mobility on various areas which relies on a flow network of data from TripAdvisor into a Neo4j graph database. Thanks to this representation, we produce aggregated graphs at various scales and apply deep tourists' analysis. One centrality aspect of graphs is used to propose a key indicator of tourists mobility.

Keywords: Location and trajectory analytics · Mobile data analytics · Mobile location-based social networks · Graph databases · Neo4j

1 Introduction

In present days, tourism is considered as one of the widest and fastest growing industries [5]. Tourism is a displacement phenomenon that fully participates in the global traffic of people's experiences, norms and indeed, tourism represent a major vector of globalization, mobility and traffic. The tourist's mobility brings into play relations that need to be analyzed and understood.

Studying tourism through the "circulation flow" object means taking into account the diversity of contemporary mobility with Web technologies. In fact, e-Tourism becomes a means to identify tourism circulation flow, via digital traces. Tourism takes advantage of the social network like TripAdvisor, Booking, Facebook, Instagram, Flickr, etc. It needs to take into account both space and time. With millions of comments and photos on locations, it becomes a real challenge for tourism actors to analyze enormous volumes of data to understand how tourist circulation evolves [15]. Analyzing tourist travel behavior and knowledge of travel motivation plays a key part in tourism marketing to create a broader vision and assist tourists in decision-making [17].

By modeling the tourist flow as a graph of areas' interconnections, it becomes possible to analyse and measure the quality or capacity of the network by applying graph theory concepts like centrality, modularity, ranking, etc. However, those methods make absolute assumptions about the manner that a graph

behaves and not on a precise flow in the network like counting shortest paths [8], or multiple paths like information of infections [2].

By applying a measure to a given set of circulation flow characteristics to another different flow will consequently generate a loss of ability to fully interpret results and get poor and inappropriate answers. In this context, it becomes a real challenge to identify a correct key indicator that turns out to be appropriate to a given graph. As a matter of fact, it becomes important to produce measures based on the network structure while it witnesses a continuous evolution.

This paper proposes an approach to extract and interpret tourist mobility on geographic areas. Based on a graph-oriented database, we model tourists' reviews from social networks as a circulation graph which can be scaled at various levels of granularity over a geographic area. We propose the *Circulation Factor* which captures locally and globally how populations behave over a given area. Our contributions can be resumed as the following:

- A circulation graph data model which can be aggregated on time and space,
- The **Circulation Factor** to capture tourists flow on the circulation graph,
- The implementation and the analysis of tourists mobility at various scales.

This paper is organized as follows. We first detail in Sect. 2 the related work on flow modeling with graphs. Then, we formalize our graph data model and explain how to aggregate it in Sect. 3. Section 4 presents the *Circulation Factor* to highlight mobility in the graph. To finish with, Sect. 5 details its integration in Neo4j and analyze the factor on the TripAdvisor dataset.

2 State of the Art

Many graph theory algorithms and concepts are used in network analysis to measure the importance of nodes, to understand interactions in the network, to show information circulation or to deduce communities of nodes that share some characteristics. Each measure follows a specific definition and rules to target important nodes in order to have a network understanding.

In literature, most used concepts for social network analysis which lead to tourists' indicators are based on the identification of nodes importance, clustering nodes into communities or extracting patterns as trajectories in the graph. Identifying nodes importance in graphs is highly dependent on its definition. In our context, nodes importance can be used in order to comprehend how tourists circulate all over a territory represented as a graph.

A first family focuses on clustering algorithms to identify collections of nodes which share some characteristics and produce communities. HCS [11] focuses on the maximization of connectivity within clusters HCS and is used to detect communities of individuals. The *Louvain* algorithm [1] is a hierarchical algorithm that maximizes cluster modularity by merging nodes into high level in a hierarchical tree. *Label Propagation* [18] binds a unique label to each node which tries to spread their own label to neighbor nodes. *Chameleon* [14] overcomes the limitations of existing agglomerative hierarchical clustering algorithms by

adapting to the cluster characteristics. Even if clustering methods could identify groups of behaviors, it does not target the circulation issue or groups too many nodes between each other which do not help to identify nodes independently.

A second family of algorithms focuses on spanning trees [10]. They are used defining the cheapest subset of edges that keeps the graph in one connected component or finding frequent patterns in a graph like with Mining Maximal Frequent sub-graphs from Graph Databases [12]. However, in our context spanning trees will only give the main path and not a global sight on a territory.

Last but not least, centrality algorithms aim at providing relevant analytical information about nodes in the graph and then nodes importance into a graph. *Closeness centrality* [6] scores each node based on their closeness to all other nodes in the network. *Betweenness centrality* [6] measures how much a node lies on the shortest path between other nodes in a network which helps to find the node that influences the data flow. *Degree centrality* [6] assigns an importance score based simply on the number of links held by each node, which shows very connected, popular or informational nodes. *EigenCentrality* [6] measures nodes influence based on the number of links but consider nodes connectivity as well as the neighborhood. However, those measures focus on single nodes to identify most representative ones but do not integrate this notion of circulation. One variant of the EigenCentrality is the well-known *PageRank* [16] mostly used by the Google search engine to rank web pages to have more accurate search results. Even if the PageRank score is a good indicator of circulation, it can only be compared with other nodes of a given graph and not between different graphs. In fact, PageRank scores are highly dependent on the composition of the graph and two different graphs even with the same set of nodes can produce very different scores and are hardly comparable which is the main objective of our circulation measure. However, PageRank will be integrated further in our measure since it enhances the circulation.

All methods presented above can be applied to any graph. However, those methods make absolute hypotheses about the manner of the graph behave and not on specific flows in the network. Applying a measure to a given set of circulation flow characteristics to another different flow will consequently generate a loss of ability to fully interpret results and get poor and inadequate answers.

On a broader analysis aspect of tourists flows extraction, some researches have been proposed like [4] for flow visualization, pattern mining [21], extraction of Point-of-Interests [19], or Kernel density estimations [20]. However, those solutions are either static focusing on hot spots hardly comparable with other points, or hardly flexible to compare various densities of paths.

To offset this problem, the main object is the formalism of graphs in a graph-oriented database model that takes into account tourists' circulation flow. Smart circulation graphs that will be able to scale out various levels of granularity can be produced. Then, we propose a measure that helps to identify the mobility of each node for a given population in order to develop a valuable indicator in terms of mobility and centrality.

3 A Tourism Circulation Graph

Modeling tourism data requires to take into account locations information, users' properties and their interactions. We propose the circulation graph data model in order to deal with interactions on locations. Graphs rely on links between users and locations through their reviews. A circulation graph is thus modeled with all the properties associated to the users.

3.1 Graph Data Models

Data Types. Our database is composed of geolocalized locations, reviews and users. A location is composed of a type (hotel, restaurant, attraction), localization (lat, long) and a rating ($r \in \mathbb{R} \wedge r \in [1.0, 5.0]$). To characterize localization, each location has been aligned with administrative areas (GADM)¹.

Each location l is linked to an area a if its geo-localization is contained into the area's shape (*SpatialPolygon* function SP), such that the $SP(l.lat, l.long) = a$. This administrative area is composed of a country, a region, a department, a district and a city. Thus, each location l is identified by: $l \in \mathcal{L}(type, r, a)$.

A review represents a note ($n \in \mathbb{N} \wedge n \in [1, 5]$) given by a user u on a location l at time t (t is in the discrete time domain \mathcal{T}). Each review is then defined by an event r_t such that: $r_t = (l, u, n)$.

Graph Data Model. To understand tourists' behavior and mobility in the study of a given destination (*e.g.*, department, region, country), we need to target tourists. For this, we focus only on users u who visit a destination at least once. Then we get all the reviews r_t they made, even elsewhere, in order to gather their circulation all over the world.

The initial graph data model T is a natural bipartite graph which links users to locations, as illustrated in Fig. 1 with dotted edges.

Even if this huge graph contains relevant information in order to produce analyzes, it will not ease the way to manipulate it or extract circulation of tourists. Therefore, we need to provide a new graph data model based on T that will allow the analyzes.

Circulation Graph Data Model. Since our analyzes require several levels of studies (*i.e.*, international to local), we need providing a generic graph data model. To achieve this, based on graph T a new graph data model is built by focusing on circulation between locations.

The circulation graph model $C(V', E')$ relies on the fact that tourists can review several locations during their trip. Consequently, the sequence of reviews from user u can generate new edges between locations. However, we consider that a trip is composed of reviews r_{t_1} and r_{t_2} written at most at 7 days apart [9]

¹ GADM: <https://gadm.org/index.html> - 386,735 administrative areas (country, region, department, district, city and town).

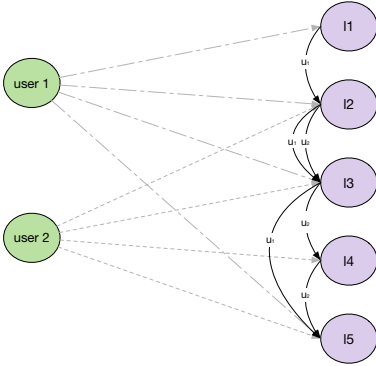


Fig. 1. Circulation graph

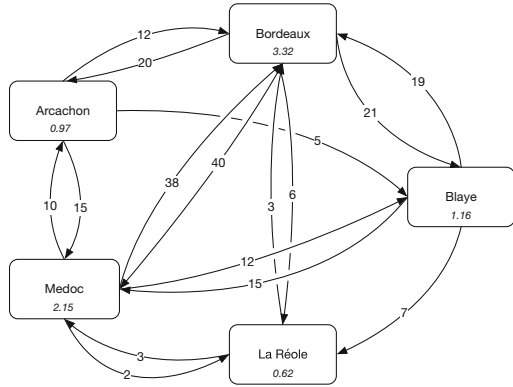


Fig. 2. Filtered circulation graph

$(t_2 - t_1 < 7d)$. Two consecutive events from user u that occur within 7 days generate an edge e between nodes l_1 and l_2 . Figure 1 illustrates the transformation of graph T by connecting directly locations on users' trip (plain edges).

3.2 Aggregated Graphs

Tourism actors need to focus their studies in a geodesic point of view, on both time and space. According to that, we need to provide fine grain studies on the circulation graph C by aggregating vertices and edges, while filtering on properties (*i.e.*, users, locations, time). For this, we produce new graphs where nodes are aggregated according to a property P on areas (*e.g.*, district, city) and produced edges give the number of edges in E' between aggregated nodes [7].

We can notice that time is discretized on both years and months. It will enable the focus on long time periods for studies, at a minimum at month scale and show the evolution over time of tourists behavior.

Moreover, we can aggregate nodes and edges from the circulation graph C to obtain an aggregated graph on other shared properties (between V' nodes). This aggregated circulation graph will be denoted by AC . Then, the study will focus on circulation between groups of locations (*e.g.*, districts, cities) on a given zone. Figure 2 illustrates this aggregation between cities of the Gironde's department.

In the following, the AC aggregate circulation graph is denoted as $ac_{n,e}^{p,f}$ where indices give nodes and edges aggregating properties and exponent as the filtering predicate on edges. Here an example of graph AC on aggregated on city nodes, years and nationality edges focused on Americans in 2018:

$$ac_{cities,year\&nat}^{USA,2018} = AC(cities, year\&nat, \sigma_{nat="USA" \wedge year=2018})$$

4 The Circulation Factor

The goal of our study is to provide a novel way to characterize the flow of touristic circulation with a valuable indicator. To achieve this, we propose the Circulation Factor CF which relies on both the circulation graph AC and the combination of PageRank computations [22].

As we saw in Sect. 2, the PageRank score of a node represents the current best solution in our context to represent the fact that tourists tend to go through an area during their journey. However, even if this score is a good indicator of circulation, it can only be compared with other nodes of a given graph and not between different graphs. In fact, PageRank scores are highly dependent on the composition of the graph and two different graphs even with the same set of nodes can produce very different scores and are hardly comparable.

To give an example, we wish to study the score's evolution of the city of Bordeaux for American tourists over years. Thus, we can compute PageRank scores on the extracted graph from $ac_{cities,year}^{2018,USA}$. Thus, "Bordeaux" PageRank score $PR_{Bordeaux}(ac_{cities,year}^{USA,2018})$ has a meaning according to other nodes (cities) in the graph like $PR_{Blaye}(ac_{cities,year}^{USA,2018})$. But the comparison is useless with $PR_{Bordeaux}(ac_{cities,year}^{USA,2017})$. We propose in this article a new measure which helps to compare various flows of circulation in AC .

4.1 The Transient Circulation Factor

To cope with this issue, we propose the **Transient Circulation Factor** which gives for a node, a value that represents how much a population circulates in an area compared to the whole population.

Definition 1. *The Transient Circulation Factor $TCF_{n,e}^{p,f}$ is a factor applied on an aggregate graph $AC = ac_{n,e}^{-,f}$. The factor $TCF_{n,e}^{p,f}(AC, \nu)$ is the comparison between the PageRank PR of a node $\nu \in AC$ where edges are only filtered out by the context f , with its PageRank in AC where edges are filtered out by both the population p and the context f :*

$$TCF_{n,e}^{p,f}(AC, \nu) = \frac{PR_{\nu}(ac_{n,e}^{p,f})}{PR_{\nu}(ac_{n,e}^{-,f})}$$

The **Transient Circulation Factor** is an impact factor of a population circulation flow over a directed graph. It represents how much a population circulates compared to other populations. The following example illustrates the *Transient Circulation Factor* of the city of "Bordeaux" for American in 2018.

$$TCF_{cities,year}^{USA,2018}(AC, Bordeaux) = \frac{PR_{Bordeaux}(ac_{cities,year}^{USA,2018})}{PR_{Bordeaux}(ac_{cities,year}^{-,2018})}$$

Remind that the Weighted PageRank [22] is based on the computation of navigation probabilities with weighted for both in and out links.

In our approach, *TCF* compares the same set of nodes and edges but updates weights according to a given population. Therefore, the comparison can be summarized to the variation of ratios of weights from a given population with all populations.

Thus, the computation of focused in/out weights on a given population can be higher or lower than 1. The *TCF* of the whole population is equal to 1. Consequently, a *TCF* value over 1 says that a given population tends to circulate through this node more than others. At the opposite a value below 1 says that the node is less central in the circulation of this population.

Thanks to this **Transient Circulation Factor**, we can compare populations on a given area, but also for all the areas on the whole graph. Thus, we can give the circulation profile of a population to say if they are more mobile than others or remain central in a narrow area.

Moreover, this factor can also be used to study the evolution of a population over years. In fact, the evolution of proportions from incoming and outgoing arcs' weight computed by PageRanks in the *TCF* can be compared between two years for a given population. The following statement means that Americans focus more on Bordeaux in 2018 than 2017.

$$TCF_{cities,year}^{USA,2017}(AC, Bordeaux) < TCF_{cities,year}^{USA,2018}(AC, Bordeaux)$$

We can also compare two populations in *AC* to identify if the first one tends to circulate more through a given node than the second population.

$$TCF_{cities,year}^{French,2018}(AC, Bordeaux) < TCF_{cities,year}^{USA,2018}(AC, Bordeaux)$$

4.2 The Global Circulation Factor

Since the *TCF* captures the intrinsic flow value of tourists circulation on a given area, we need to produce an indicator of a global sight on the behavior of a given population. The **Global Circulation Factor** computes all *TCFs* of a given population on the whole graph to show how much this population circulates on the territory compared to other people.

Definition 2. *The Global Circulation Factor $GCF_{n,e}^{p,f}$ is a factor which computes the mean value of *TCF* values for all nodes $\nu \in AC$ with aggregated nodes on property n and edges e , and filtered out by the context f and the population p :*

$$GCF_{n,e}^{p,f}(AC) = \frac{\sum_{\nu \in AC} TCF_{n,e}^{p,f}(AC, \nu)}{|AC|}$$

The mean of all *TCF* values integrates all local behaviors to provide a broad sight of all populations p in a given zone (*i.e.*, graph *AC*). The capability to manipulate the level of aggregation on nodes n and edges e helps to capture different kinds of behavior: local circulation (city) to global (district or department), evolution (years) to seasons (months), etc. To finish with, filter f on edges allows us to focus on a specific aggregated contexts (*e.g.*, years, months).

Section 5 will validate our approach with various settings. It will focus especially on the capabilities to enlighten behaviors with *TCF* and *GCF* at various scales and aggregations. Even if the number of possibilities is really wide, we have produced most significant observations which enhance our contributions.

5 Experiments

5.1 The Neo4Tourism Implementation

Neo4Tourism [3] is a framework which helps to manipulate graphs by aggregating and filtering them by taking into account geographic data. Graphs are stored in a *Neo4j* server dedicated to the tourist circulation characterization. It transforms bipartite graphs in circulation graphs *C* and its aggregations *AC*. Aggregated graphs are also materialized as new graphs for optimization purposes. Since data are stored incrementally (time dependency), materialized graphs do not have to be updated, only new edges are added to graphs.

In our circulation characterization, we will focus especially on aggregated graphs *AC*. The *Cypher* query language used in *Neo4j* helps to manipulate graphs and produces *TCFs* and *GCFs* values.

Aggregated Graphs Materialization. *TCF* and *GCF* require several granularities of studies (*e.g.*, regions, districts, cities, etc.), it is then necessary to compute queries on various aggregate graphs. The first *AC* graph focuses on aggregated nodes with the smallest area: towns. Every location belonging to a town is merged in a single node. All edges which share the same properties (*i.e.*, nodes, year, month, nationality, age) produce an edge with a new property *NB* that represents the number of merged edges. This first graph at town scale is built by a Java program and stored in *Neo4j*. It reads all series of reviews from each user to generate the circulation between locations and consequently towns.

Other aggregated graphs are built from the first one, they merge nodes that share a same property (department, district, city), so do the edges. To achieve this in *Cypher*, the *MERGE* clause is used to produce derived graphs:

```
MATCH (t1:Town) -[t:trip]-> (t2:Town)
MERGE (c1:City{name:t1.city}) MERGE (c2:City{name:t2.city})
MERGE (c1)-[ct:trip{year:t.year,month:t.month,nat:t.nat,age:t.age}]->(c2)
ON CREATE SET ct.NB=t.NB ON MATCH SET ct.NB=ct.NB+t.NB
```

This query relies on the first graph composed of *Town* nodes where edges are typed as *trip*. Since each node is labeled with GADM administrative zones, we can merge them according to various areas, here city names. Then, edges are merged when they share same properties and *NB* values are summed.

All graphs are generated at all scales: city, district, departments, regions and countries. Each time, nodes from a given scale contains areas information to be filtered out and to focus on a specific zone. Thus, we can extract subgraphs on a given zone like a department or a region.

Table 1. Tripadvisor dataset

Table	# instances
Locations	4.8×10^4
Users	1.31×10^6
Reviews	3.58×10^6

Table 2. Different AC graphs characteristics

AC Graphs	Aggregation	# nodes	# edges
Nouvelle-Aquitaine	Cities	482	382,266
	Districts	41	170,403
Hauts-de-France	Cities	297	153,345
	Districts	27	75,429
Gironde	Cities	55	94,032
	Districts	6	26,704

Advanced Manipulations. Now we have circulation graphs at all scales; we can compute *PageRanks* on them by applying prepared statement queries. The following query integrates a callable function from the *graph data science* package² to produce PageRank scores for each district within the “Gironde” department.

```
CALL gds.graph.create.cypher("CypherProjection",
  "MATCH (c:City{department:'Gironde'}) RETURN id(c) as id",
  "MATCH (c1:City)-[t:trip{year:2018,nationality:'USA'}]->(c2:City)
  RETURN id(c1) as from, id(c2) as to, sum(toFloat(t.NB)) as weight")
CALL gds.pageRank.stream("CypherProjection",
  {dampingFactor:0.85,iterations:50,weightProperty:true}) YIELD node, score
RETURN node.city, sum(score) as score;
```

We can see that the *AC* graph uses a Cypher projection to get *Gironde*’s graph (Bordeaux’s department in France) where only edges in 2018 done by Americans are kept and merged (sums of NBs). Then, the PageRank is computed on this sub-graph “CypherProjection” with PageRank scores for each node/city of the sub-graph. Figure 2 gives an example of this result where PageRank scores are above each node.

We must notice that we keep edges that link a node to itself. In fact, this edge represents the reality that tourists circulate within an area.

To provide $TCF_{n,e}^{p,f}(AC, \nu)$, it requires to compute two PageRanks. The first one is given by the query above, and the second one just removes the filter on nationality. TCF values are then associated to each node in *AC*. Then, GCF values are computed with a simple mean query on considered nodes.

Graphs Characteristics. To support our approach, we need constituting a dataset which represents the best notion of circulation over a territory by visiting various locations. Several e-tourism websites were considered. *Booking* focuses only on accommodation and cannot be used at small scales (between cities). *Flickr* is really interesting for precise locations. However, the public dataset is

² Neo4j 4, GDS: <https://neo4j.com/docs/graph-data-science/1.2/>.

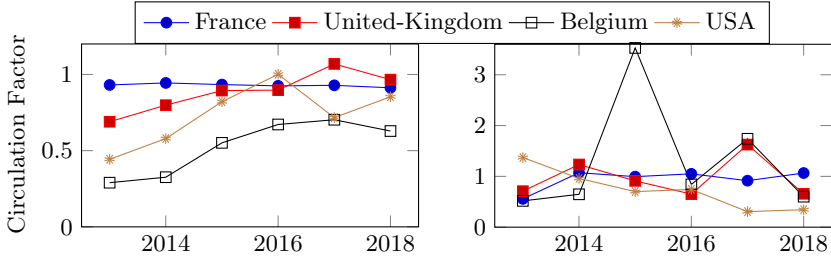


Fig. 3. TCFs of 1) Bordeaux and 2) Blaye (city-scale)

too small to be representative for diverse populations. Finally, we chose *TripAdvisor* which gives precise information on locations, populations and constitutes a sufficient amount of data to begin to be representative.

Table 1 gives the initial dataset gathered from *TripAdvisor* focused on two French regions: *Nouvelle-Aquitaine* and *Hauts-de-France*. This dataset contains 3.58×10^6 reviews on 4.8×10^4 locations.

The setting of our experiments tries to enhance both graph data manipulations at various scales and the capability of TCF and GCF to witness the circulation of tourists. To achieve this we have extracted six *AC* graphs (Table 2) with three zones aggregated on both cities and districts: *Nouvelle-Aquitaine* (region), *Hauts-de-France* (region) and *Gironde* (department). This will help to understand local and global behaviors. Notice that the number of edges here is the sum of edges' weight within the graph. This loss between the number of reviews and the number of edges' weight corresponds to the fact that we focused only tourists circulation (seven-day trip as mentioned previously).

5.2 Transient Circulation Factor's Evaluation

TCF at City-Scale. Figure 3 shows the TCF evolution of *Bordeaux* and *Blaye* cities for different nationalities. We can see that the ratio of PageRanks for *Bordeaux* is almost equal to one for French while witnessing a small decrease of interest. At the same time, British and American populations grow significantly to reach 1 in 2016. This effect is correlated to the opening of the new high-speed train line between Paris and Bordeaux making this city more central in tourist trips. Belgians have a lower TCF but tend to grow in the past years.

The city of *Blaye* known for its castle and wines witnesses an interesting aspect of the TCF: event detection. The French population is on average more represented in this area which confirms the fact that they do prefer the countryside. But interestingly Belgian people in 2015 has a factor reaching almost 4. This anomaly is explainable by an event organized by a tour operator that occurred in 2015 with a consequent group of belgians. Consequently, this score grows up significantly at once. This fact is also observed locally for various populations.

Maps that are shown in Fig. 4 correspond to TCF values of *Gironde* cities in 2014 for respectively French and British. The colors scale helps to highlight

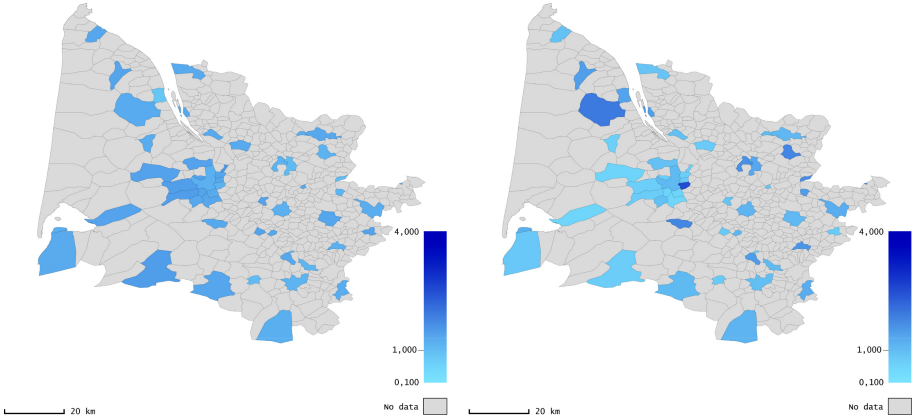


Fig. 4. 1) French and 2) British TCFs for Gironde’s cities in 2014

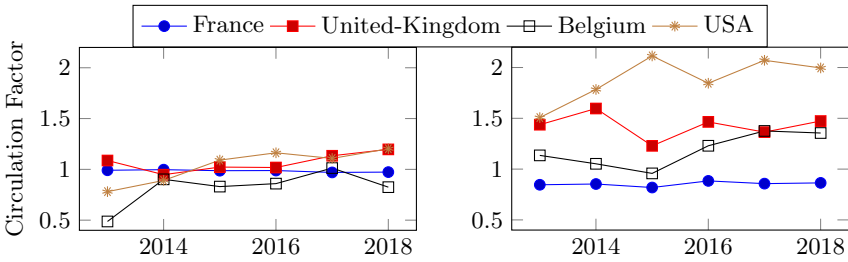


Fig. 5. TCFs of 1) Bordeaux and 2) Libourne (district-scale)

areas where those populations are more mobile. Of course we identify touristic zones easily like *Bordeaux* (hypercenter), *Arcachon* (South-west), *Lespar-Médoc* (North), *Blaye* (South), *Langon* (South-East) or *Libourne* (North-East).

For French, we can see that most cities have a score around 1 which means that they homogeneously circulate all over cities. At the opposite, British are less uniform while they are focusing mainly on the area of *Lespar-Médoc* and *Libourne*. *Arcachon* and *Bordeaux*’s suburbs are less central in their journey.

It is interesting to bring out mobility differences of local behaviors on the territory. The map identifies clearly where each population concentrates their journey. More importantly those distributions are comparable to each other.

TCF at District-Scale. We now aggregate the graph at district scale with six significant areas in Gironde. It produces a smaller graph in which mobility is also concentrated between those nodes.

As we can see, scores are more homogeneous in Fig. 5 for the Bordeaux district. It is due to the fact that nodes and edges are aggregated which leads to less variability on mobility. We can confirm the fact that French people are less rep-

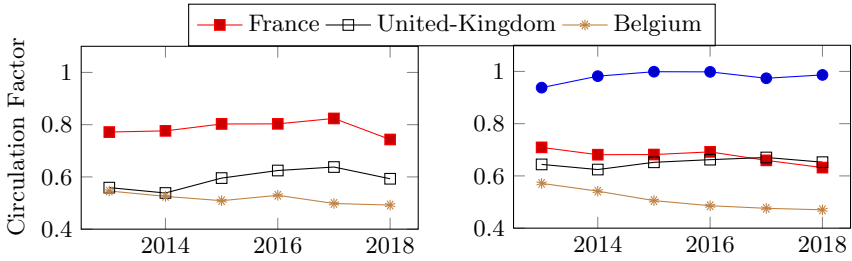


Fig. 6. GCFs of 1) *Nouvelle-Aquitaine* and 2) *Hauts-de-France* (city-scale)

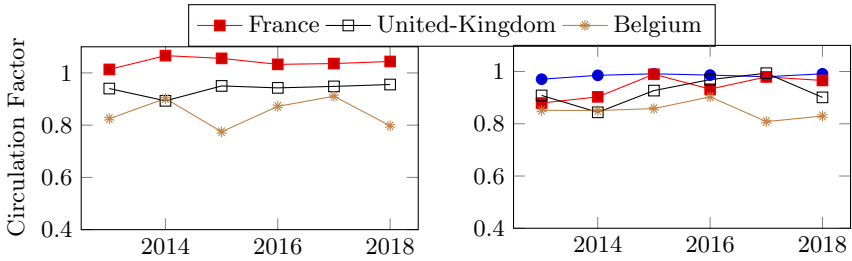


Fig. 7. GCFs of 1) *Nouvelle-Aquitaine* and 2) *Hauts-de-France* (district-scale)

representative of the circulation around *Bordeaux*. Belgians witness a better score than at city scale, this means that the mobility is higher with more exchanges between districts (back and forth to *Bordeaux*'s area).

However, the attractiveness growth of the *Libourne* district is really significant with higher scores for all populations except for the French one. The whole area concentrates many castles, wine tasting and tour operator activities and thus attractive for foreign tourism.

It is interesting to see that the TCF brings different conclusions at each scale. The city-scale helps to extract events for a given population and more local mobility. At the opposite, the district-scale gives tendencies for populations but also the typology of places that are considered. We observed similar results on other departments and the graph on “Hauts-de-France”.

5.3 Global Circulation Factor’s Evaluation

While TCF shows hotspots within a circulation graph, GCF focuses on the whole graph in order to characterize a population all over the territory.

To enhance this circulation indicator, we now apply the computation on wider graphs and try to compare both scaling of areas and two different zones. To achieve this we apply the GCF on region graphs on *Hauts-de-France* and *Nouvelle-Aquitaine*. Having two different geographic zones will be useful to study differences and common points between them. And for each graph we have two scales: city scale and district scale.

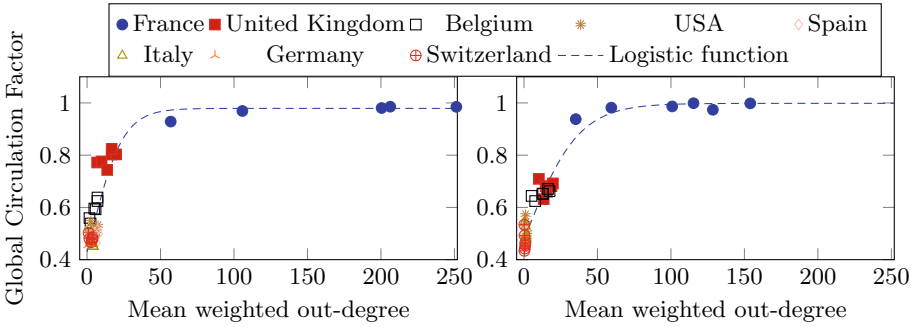


Fig. 8. Logistic function of 1) *Nouvelle-Aquitaine* & 2) *Hauts-de-France* (city-scale)

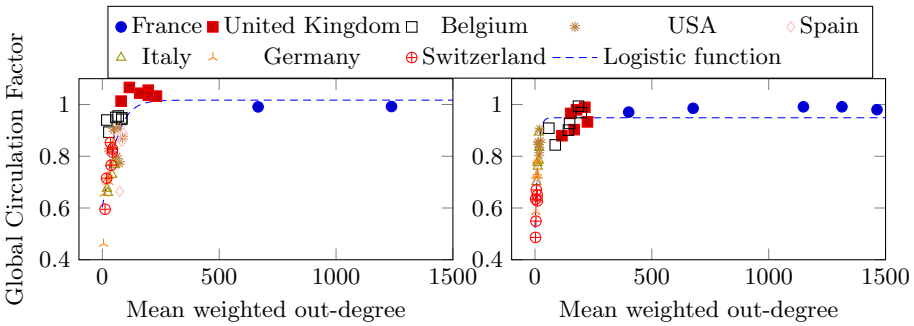


Fig. 9. Logistic function of 1) *Nouvelle-Aquitaine* & 2) *Hauts-de-France* (district-scale)

Figure 6 gives the GCF evolution on regions at city scale. This scale is interesting since tourists usually do not circulate in such a wide zone (region) with small stops (cities). Consequently, this evolution shows the global interest of a population to visit a region. From *Nouvelle-Aquitaine* (left) and *Hauts-de-France* (right) we can especially see that British tourists are more interested in the first one (old British country). At the opposite Belgians tend to visit more places in *Hauts-de-France* which is next to their own country, even though the British circulate more than the latter in this region.

At **district scale** in Fig. 7, since cities are aggregated into bigger areas, the circulation effect is higher and less fluctuating. Differences between British and Belgians are less visible. However, Americans witness a significant growth from the previous analysis. It is due to the fact that Americans target specific zones within each district, especially on Second World War memorials (*i.e.*, memory tourism [13]).

5.4 The GCF Property

As we saw, GCF allows highlighting populations behavior at various scales. One can argue that some other centrality measures can bring out similar results. Remind that the most similar measure is based on the PageRank. However

Table 3. Logistic parameters and errors

Parameters	NA Cities	HF Cities	NA District	HF District
L	9.791×10^{-1}	9.984×10^{-1}	1.017×10^0	9.488×10^{-1}
k	9.215×10^{-2}	5.611×10^{-2}	2.250×10^{-2}	1.062×10^{-1}
x_0	2.667×10^0	1.598×10^0	-1.653×10^1	-1.937×10^0
MSE	1.620×10^{-1}	7.229×10^{-2}	2.871×10^{-1}	1.231×10^{-1}
MAE	4.540×10^{-2}	3.165×10^{-2}	5.660×10^{-2}	4.366×10^{-2}
MAPE	8.204 %	5.688 %	7.136 %	5.753 %

as stated previously, it cannot compare various scales, zones, populations or evolution.

Another similar solution is the degree centrality [6] by computing the average out-degree centrality of all nodes for different populations as well as its evolution.

This centrality gives an approximate solution (for space reason we only present this one). We tried to find a correlation between the weighted out-degree centrality and the GCF. Figures 8 (city scale) and 9 (district scale) show a logistic regression of mean weighted out-degrees x : $f(x) = \frac{L}{1+e^{-k(x-x_0)}}$ where L denotes the upper bound, k the growth rate and x_0 its midpoint.

Since the PageRank is a logarithmic measure, it is natural to follow this law. But this scaling effect is all the more important since it gives an exploitable indicator of circulation. In fact, we can see that French out-degrees can be very high, fluctuating (axes are cut for *Nouvelle-Aquitaine*) and hardly comparable, likewise low out-degrees make all small populations packed all together. The GCF helps to differentiate them in a logarithmic scale.

Table 3 gives parameters which belong to each distribution (Nouvelle-Aquitaine/Hauts-de-France, cities-districts). L values are bounded to 1.02 which corresponds to British and French circulations and is as expected higher for district scale. k is the tendency of the curve which is the opposite of the growth, then values tend to grow faster at city scale. x_0 gives the midpoint, the lower the value is, the lower is the minimum GCF. To validate our correlation between the weighted out-degree centrality and the GCF, we used three forecasting accuracy techniques MAE, MSE and MAPE³. MAPE is the most precise measure to compare the accuracy between different items since it measures the relative performance. In our case MAPE values are lower than 8.2% which is an excellent accurate.

6 Conclusion

We have formalized in this article a methodology to produce and manipulate circulation graphs from a digital trace of users on Social Networks. This approach

³ MAE: Mean Absolute Error, MSE: Mean Squared Error, MAPE: Mean Absolute Percentage Error.

helps to produce various aggregated graphs by zooming on geographic areas and filtering on population characteristics. We also proposed the *Circulation Factor* which enables the mobility comparison from a population to another either on space and time. Our approach has been integrated in the *Neo4j* database which easily produces various aggregated graphs and applies graph theory algorithms on them. Our experiments showed that the TCF can highlight events and local mobility while the GCF enhances global tendencies and population behavior.

For future works, we wish to propose a prediction model that takes into account for each population its tendency and predicts the next circulation factor on a zone. On the other hand, it should be interesting to focus on detecting global propagation patterns for a given population like spanning trees by taking into account coverage. To finish with, we wish to focus on community extraction on the graph to compare how much linked cities of a cluster can be correlated to an administrative zone and thus representative of its impact in the area.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**(10), P10008 (2008)
2. Borgatti, S.P.: Centrality and network flow. *Soc. Netw.* **27**(1), 55–71 (2005)
3. Chareyron, G., Quelhas, U., Travers, N.: Tourism analysis on graphs with Neo4jTourism. In: U, L.H., Yang, J., Cai, Y., Karlapalem, K., Liu, A., Huang, X. (eds.) WISE 2020. CCIS, vol. 1155, pp. 37–44. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-3281-8_4
4. Chua, A., Servillo, L., Marcheggiani, E., Moore, A.V.: Mapping Cilento: using geotagged social media data to characterize tourist flows in southern Italy. *Tour. Manag.* **57**, 295–310 (2016)
5. Cooper, C., Hall, C.M.: Contemporary Tourism. Routledge, Abingdon (2007)
6. Das, K., Samanta, S., Pal, M.: Study on centrality measures in social networks: a survey. *Soc. Netw. Anal. Min.* **8**(1), 13 (2018). <https://doi.org/10.1007/s13278-018-0493-2>
7. Endriss, U., Grandi, U.: Graph aggregation. *Artif. Intell.* **245**, 86–114 (2017)
8. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: a measure of betweenness based on network flow. *Soc. Netw.* **13**(2), 141–154 (1991)
9. Gössling, S., Scott, D., Hall, C.M.: Global trends in length of stay: implications for destination management and climate change. *J. Sustain. Tour.* **26**(12), 2087–2101 (2018)
10. Graham, R.L., Hell, P.: On the history of the minimum spanning tree problem. *Ann. Hist. Comput.* **7**(1), 43–57 (1985)
11. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. *Inf. Process. Lett.* **76**(4–6), 175–181 (2000)
12. Huan, J., Wang, W., Prins, J., Yang, J.: Spin: mining maximal frequent subgraphs from graph databases. In: ACM SIGKDD 2004, pp. 581–586 (2004)
13. Jacquot, S., Chareyron, G., Cousin, S.: Le tourisme de mémoire au prisme du “Big Data”. Cartographier les circulations touristiques pour observer les pratiques mémorielles. *Mondes du Tourisme*, June 2018

14. Karypis, G., Han, E.H., Kumar, V.: Chameleon: hierarchical clustering using dynamic modeling. *Computer* **32**(8), 68–75 (1999)
15. Keng, S.S., Su, C.H., Yu, G.L., Fang, F.C.: AK tourism: a property graph ontology-based tourism recommender system. In: KMICe, pp. 83–88. UUM (2018)
16. Langville, A.N., Meyer, C.D.: Deeper inside pagerank. *Internet Math.* **1**(3), 335–380 (2004)
17. March, R., Woodside, A.G.: *Tourism Behaviour: Travellers' Decisions and Actions*. Cabi, Wallingford (2005)
18. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **76**(3), 1–11 (2007)
19. Spyrou, E., Korakakis, M., Charalampidis, V., Psallas, A., Mylonas, P.: A geo-clustering approach for the detection of areas-of-interest and their underlying semantics. *Algorithms* **10**(1), 35 (2017)
20. Sun, Y., Fan, H., Helbich, M., Zipf, A.: Analyzing human activities through volunteered geographic information: using Flickr to analyze spatial and temporal pattern of tourist accommodation. In: Krisp, J. (ed.) *Progress in Location-Based Services*. LNGC, pp. 57–69. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-34203-5_4
21. Vu, H.Q., Li, G., Law, R., Ye, B.H.: Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tour. Manag.* **46**, 222–232 (2015)
22. Xing, W., Ghorbani, A.: Weighted pagerank algorithm. In: CNSR 2004, pp. 305–314, May 2004