Propagation Measure on Circulation Graphs for Tourism Behavior Analysis

Hugo Prevoteau, Sonia Djebali, Zhao Laiping and Nicolas Travers

September 30, 2022

Propagation Measure on Circulation Graphs for Tourism Behavior Analysis*

Hugo Prevoteau
College of Intelligence and Computing, Tianjin University, Tianjin, China
Léonard De Vinci Pole Universitaire, Research Center, 92916 Paris La Défense, France
hugo.prevoteau@edu.devinci.fr

Sonia Djebali
Léonard De Vinci Pole Universitaire, Research Center, 92916 Paris La Défense, France
sonia.djebali@devinci.fr

Zhao Laiping
College of Intelligence and Computing, Tianjin University, Tianjin, China
laiping@tju.edu.cn

Nicolas Travers
Léonard De Vinci Pole Universitaire, Research Center, 92916 Paris La Défense, France
nicolas.travers@devinci.fr

ABSTRACT

Social network analysis has widespread in recent years, especially in digital tourism. Indeed the large amount of data that tourists produce during their travels represents an effective source to understand their behavior and is of great importance for tourism stakeholders. This paper studies the propagation effect of tourists on the territory thanks to geotagged circulation graphs.

A new weighted measure is introduced for circulation characterization based on both topologies and distances. This measure helps to determine the behavior of tourists on local and global areas. An optimization strategy based on spanning trees is applied to reduce the computation on the whole graph while keeping a good approximation of the behavior.

CCS CONCEPTS

• **Information systems** → *Geographic information systems*; • **Theory of computation** → **Sparsification and spanners**.

KEYWORDS

Spanning Trees, Graph Data Mining, Digital Tourism

ACM Reference Format:
Hugo Prevoteau, Sonia Djebali, Zhao Laiping, and Nicolas Travers. 2022. Propagation Measure on Circulation Graphs for Tourism Behavior Analysis. In *BDA'2022: October 24–27, 2022, Clermont-Ferrand, France.* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Due to the growth of the tourism industry [1], the spatiotemporal data platforms (*e.g.,* `Instagram`, `Flickr`, `Tripadvisor`, etc.) have emerged to capture tourists' experience.

These platforms constitute an interesting observation field to analyze tourists' behavior. It has become crucial and a real challenge for tourism stakeholders to understand the flow of tourists and how they propagate themselves on the territory.

Data produced with these platforms can be modeled as a network, making the application of graph theory algorithms possible [2, 3]. However, those methods make absolute assumptions about the manner that a graph behave, so applying a measure to a given circulation flow characteristics to another different flow will generate a loss of ability to fully interpret results.

We propose in this work a new propagation measure, the `Remoteness Influence Factor (RIF)` that takes advantage of existing methods on the topology of the graph and enhances them with the geodesic distance to analyze the propagation effect on both time and space. The `RIF` measure captures both the topology of propagation (*i.e.,* star *vs.* deep trips) and geodesic distances (*i.e.,* long trip *vs.* excursionism).

The `RIF` measure evaluates tourists propagation in a network of a given area. However, analyzing the whole graph is highly computational due to the volume of data. To scale-up our approach we reduce the time complexity by extracting sub-graphs. Indeed, subgraphs, like `Spanning Trees`, are ideal solutions to analyze huge graphs considering only a subset of the graphs without losing their essential properties.

2 A TOURISM CIRCULATION GRAPH

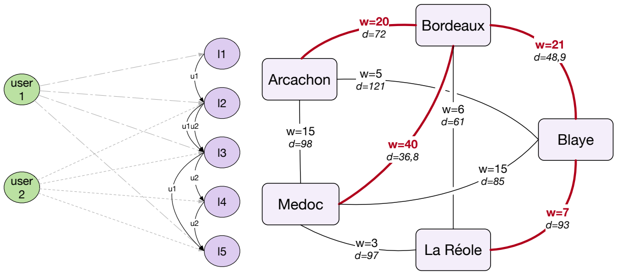
Graph Data Models Our database is composed of geolocalized locations of a type (hotel, restaurant, attraction), localization aligned with administrative areas (GADM)[1] and users with nationnality.

The circulation graph model $C(V, E)$ relies on the fact that tourists can review several locations during their trip. The figure 1a illustrates the transformation of bi-partie graph by connecting directly locations on users' trip (plain edges).

Aggregated Graphs To produce the multi-weighted aggregated graph, our case study relies on various aggregation levels on nodes. Nodes correspond to locations grouped together while sharing area properties (e.g., cities, district). Edges from the circulation graph are merged together on review properties (e.g., nationality). Two properties are associated with edges: 1) the cardinally of circulation edges is represented as the weight w and 2) the geodesic distance d between two nodes. We use the notation $AC(\mathcal{V}, \mathcal{E}(w, d))$ to denote the multi-weighted aggregated graph AC for a given area,

[*] The original version of this article has been published at ACM SIGAPP@SAC'22 [4]

[1] https://gadm.org/index.html - 386,735 administrative areas.



(a) Circulation Graph

(b) Aggregated Graph

Figure 1: Multi-Weighted Aggregated Graph Construction

where $\mathcal{V}$ is a set of vertices aggregated on a given *area* level, and $\mathcal{E}(w, d)$ a set of multi-weighted edges by two properties w and d, where $w, d \in \mathbb{R} | w > 0 \wedge d > 0$, as illustrated in figure 1b.

3 THE REMOTENESS INFLUENCE FACTOR

The RIF measures the correlation between weights w and distances d in the graph by combining Weighted Betweenness Centrality (WBC) (nodes influence *wrt.* the behavior) and Shortest Paths (covered distance). WBC is defined as the number of paths between node pairs that run through a specific node, divided by the total number of paths from any node to all other nodes in AC.

Definition 3.1 (Remoteness Influence Factor). Consider the multi-weighted aggregated graph AC$(\mathcal{V}, \mathcal{E}(w, d))$ on areas and s a source node of the graph. $RIF(AC, s)$ measures the remoteness of vertices combined with their influences in the graph. For each node $n \in \mathcal{V}$, it computes its normalized distance from the source node s combined with the inverse of its WBC of n, defined as:

$$RIF(AC, s) = \sum_{n \in \mathcal{V}} \left(\frac{\log_{dist_{max}} \left(\sum_{e \in path(s,n)} dist(e) \right)}{|\mathcal{V}| - 1} \times \frac{\frac{1}{1+WBC(n)}}{|\mathcal{V}|} \right)$$

where s is the source of the graph AC, being the node with the highest WBC based on weights w (*i.e.,* the main starting node of trips), $dist_{max}$ is the maximum path length d and $WBC(n)$ is the WBC of node n based on weights w.

The RIF highlights a node that is topologically badly connected and far away in terms of distance from the source of the graph. While the RIF helps to define a propagation score for a multi-weighted graph, its computation with shortest paths and betweenness centralities is polynomial. Our approach requires to optimize the computation. To achieve this, we propose to apply the RIF on a Spanning Tree extracted from the AC graph.

Spanning Trees provides an interesting way to enhance propagation analysis by extracting the strongest connections between nodes. It is considered to be the best representation of the tendencies in a network [5]. Since our approach aims at analyzing the propagation of tourists, we need to have Spanning Trees from which the strongest connections represent the highest tourists' flow. Hence we use a Maximum Spanning Tree (MaxST) that maximizes the sum of edge weights. MaxST represents the most important routes for information flow in the graph. We must notice that with our circulation graph, the construction of the MST by taking a highest betweenness centrality for the root (often a capital city) or the highest weighted edge lead to the same total weight.

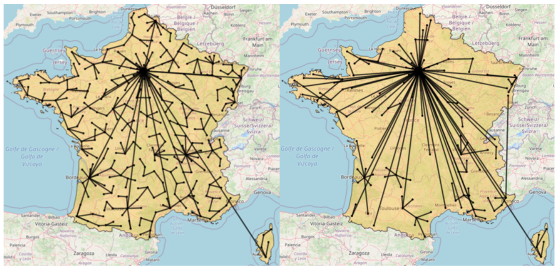


Figure 2: French and American Maximum Spanning Trees in 2018 over *France* (district-scale)

Thanks to this, the RIF computation requires only to focus on remaining edges which reduce considerably the complexity both for paths and centralities.

The analysis of the maps in figure 2 made it possible to identify very different behaviors: exploratory for the French (homogeneity and depth), day-trippers for the Americans (star propagation over long distances).

4 CONCLUSION

This paper formalizes a methodology to extract and study the propagation effect on multi-weighted graphs, through the topology and edge weights on circulation graphs.

We hence proposed the RIF measure for propagation which relies on both topology and geodesic data. We also proposed an optimization strategy of this computation by relying on a MST which keeps a good approximation of this measure.

Our case studies have shown that computing the RIF at different scales in the same area helps to capture local and global behaviors on the propagation effect.

For future work, we would like to study an unconnected trees, also called forests to observe how different tourist regions behave and how they interact within a territory.

Acknowledgments. This work has been funded by the *Métropole Européenne de Lille.*

REFERENCES

[1] Chris Cooper and C Michael Hall. 2007. Contemporary tourism. Routledge, London, UK.

[2] Ronald L Graham and Pavol Hell. 1985. On the history of the minimum spanning tree problem. Annals of the History of Computing 7, 1 (1985), 43–57.

[3] Fei Hu, Zhenlong Li, Chaowei Yang, and Yongyao Jiang. 2019. A graph-based approach to detecting tourist movement patterns using social media data. Cartography and Geographic Information Science 46, 4 (2019), 368–382.

[4] Hugo Prevoteau, Sonia Djebali, Laiping Zhao, and Nicolas Travers. 2022. Propagation measure on circulation graphs for tourism behavior analysis. In SAC '22: The 37th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, April 25 - 29, 2022. ACM, 556–563. https://doi.org/10.1145/3477314.3507070

[5] CJ Stam, P Tewarie, E van Dellen, ECW Van Straaten, A Hillebrand, and P van Mieghem. 2014. Trees and the forest: characterization of complex brain networks with minimum spanning trees. Int. J. of Psychophysiology 92, 3 (2014), 129–138.