informatics

MDPI

Article

From Data to Rhizomes: Applying a Geographical Concept to Understand the Mobility of Tourists from Geo-Located Tweets

Federica Burini [1], Nicola Cortesi [2,3] and Giuseppe Psaila [3,*]

[1] Department of Foreign Languages, Literatures and Cultures, University of Bergamo, Via Salvecchio 19, 24129 Bergamo, Italy; federica.burini@unibg.it
[2] Consortium for Technology Transfer C2T, Corso di Porta Vittoria, 28, 20122 Milano, Italy; nicola.cortesi@consorzioc2t.it
[3] Department of Management, Information and Production Engineering, University of Bergamo, Viale Marconi 5, 24044 Dalmine, Italy
* Correspondence: giuseppe.psaila@unibg.it; Tel.: +39-035-205-2355

Abstract: In geography, the concept of "rhizome" provides a theoretical tool to conceive the way people move in space in terms of "mobility networks": the space lived by people is delimited and characterized on the basis of both the places they visited and the sequences of their transfers from place to place. Researchers are now wondering whether in the new era of data-driven geography it is possible to give a concrete shape to the concept of rhizome, by analyzing big data describing movement of people traced through social media. This paper is a first attempt to give a concrete shape to the concept of rhizome, by interpreting it as a problem of "itemset mining", which is a well-known data mining technique. This technique was originally developed for market-basket analysis. We studied how the application of this technique, if supported by adequate visualization strategies, can provide geographers with a concrete shape for rhizomes, suitable for further studies. To validate the ideas, we chose the case study of tourists visiting a city: the rhizome can be conceived as the set of places visited by many tourists, and the common transfers made by tourists in the area of the city. Itemsets extracted from a real-life data set were used to study the effectiveness of both a topographic representation and a topological representation to visualize rhizomes. In this paper, we study how three different interpretations are actually able to give a concrete and visual shape to the concept of rhizome. The results that we present and discuss in this paper open further investigations on the problem.

Keywords: mobility networks of people; geo-located tweets; itemset mining; concrete shapes for rhizomes

Citation: Burini, F.; Cortesi, N.; Psaila, G. From Data to Rhizomes: Applying a Geographical Concept to Understand the Mobility of Tourists from Geo-Located Tweets. *Informatics* **2021**, *8*, 1. https://dx.doi.org/10.3390/informatics8010001

Received: 23 October 2020
Accepted: 17 December 2020
Published: 24 December 2020

1. Introduction

The massive production of personal data, thanks to the diffusion of the Internet, social media and location-based services, could be exploited in order to gather information about mobility of citizens. This is made possible by mobile apps for social media, such as Twitter, that allow users to post geo-located messages. Consequently, a completely new possibility is offered to researchers, if compared with the past: it is possible to collect data sets about traveling (or simply moving) people, obtaining possibly large data sets that could become precious for analysts. Similarly, a plethora of data sets concerned with people habits could be extracted from social media, which could be very useful for any kind of human science; in practice, Big Data [1] are bridging all sciences to adopt the same data-driven approach that has characterized studies in physics for a long time (observations provide data, which inspire theories that are verified by means of data collected by making new observations). Geographical studies are touched by this paradigmatic change as well [2]: in fact, they have started recently the transition from a data-scarce context to a data-rich context. As an example, if geographers and spatial analysts were provided with traces of tourists that

visited a given city or territory, derived from geo-located messages that tourists posted on social media, they could try to understand how tourists experienced the territory, so as to reveal important drivers that are able to attract tourists. This kind of analysis is quite important for small and tourist-oriented cities, whose administrators would like to increase the number of tourists.

Geographers developed theoretical frameworks to perform such an analysis. In our opinion, currently the most interesting concept for understanding mobility of people is the concept of *rhizome* [3]: it characterizes the space lived by people, following a networked approach. The concept of rhizome is a theoretical framework, thought to provide a key to read a modern society based on mobility; to the best of our knowledge, no one has tried to use it in a data-driven analysis of geo-located messages (posted on Twitter) concerning mobility of people, based on the places people visited, and on transfers they made from one place to another; in other words, mobility networks can be derived and further analyzed, allowing important connections between places to emerge. On this basis, considering again the case of tourist-oriented cities, geographers could wish to "discover mobility networks of tourists, to let (possibly virtual) connections concerning touristic destinations emerge". This is clearly a multi-disciplinary work, where geographers and computer scientists have to jointly contribute to address the problem.

Can informatics provide the right tools to validate the concept of rhizome in the data-rich context, in order to give a concrete shape to the concept of rhizomes? Is it possible to move from the theoretical intuition to a visualization of it? The attempt to address this research problem is the main goal of this paper.

The research problem reminded us of the well-known technique for data mining called *itemset mining* [4], that is the basis for mining association rules: the goal of this technique is to extract frequent associations of items, i.e., groups of items that frequently appear together in transactions (the technique originated as a tool for market-basket analysis in the retail market). If we think of transactions as trips and we think of items either as visited places or as transfers from one visited place to another, it appears that this technique could be applied to give a practical interpretation to the concept of rhizome. Nevertheless, visualization techniques are necessary to make rhizomes "visible": in fact, the concept of rhizome (in botany, the rhizome is the way roots of plants tangle) is a "visual" concept. We argued that, in this context, visualization is essential to give results obtained by itemset mining the shape of rhizome.

The research problem has emerged during the multi-disciplinary research project named *Urban Nexus* [5], which has investigated how to define "a data-driven methodology to study mobility of city users based on data coming from Social Media sources, to let (possibly virtual) connections among places emerge". In fact, in this project we argued that geo-located messages posted by city users could provide an unexpected source of information and an unexpected perspective about mobility, in a way that traditional statistical methodologies are not able to do.

In the project, we identified the case study of "studying mobility networks of tourists in a city and its surrounding area", and we built the *case-study data set*.

We defined the following methodological approach. First of all, we defined three practical interpretations of the concept of rhizome by means of the technique for itemset mining, coupled with both topographic and topological representations. Second, we applied them to the case-study data set. Third, we analyzed what it is possible to obtain and whether these representations are actually able to give a concrete and visual shape to the concept of rhizome.

We discovered that three different interpretations of the concept of rhizome (based on itemset mining) highlight different aspects that characterize mobility. We also observed that the topographic and topological representations of itemsets jointly contribute to give a concrete and visual shape to rhizomes: the former one is able to put their physical boundaries in evidence; the latter is able to show the strength of connections and the poly-centric nature of rhizomes.

It is very important to clarify the exact contributions of the paper. First of all, the aim of this paper is not to provide a definite solution to the problem of studying mobility networks by means of social-media data, as well as its goal is not to provide a comprehensive pool of techniques to pre-process and geo-code data from social media. The main contribution is the following: provided that analysts have a data set containing traces of moving people, which they consider of satisfactory quality, it provides a a practical application of the concept of rhizome based on itemset mining and visualization techniques; this is a (little but possibly crucial) first step towards the application of this concept to data-driven analysis of mobility networks, which could inspire further discussions and investigations.

Nevertheless, we do not forget that the transition to a data-driven approach is enabled by the availability of data that can be obtained from social media. In fact, these considerations inspired us to associate the concepts of rhizome and itemset mining, while working on the problem of studying mobility networks of tourists detected through Twitter. Consequently, we can consider a kind of "secondary contribution" of the paper: showing how the application of the concept of rhizome to traces of people gathered through Twitter can provide geographers with novel sources and tools. In fact, this way we show how the joint contributions of computer science and geography can foster the exploitation of these novel data sources.

The paper is organized as follows. Section 2 introduces the background of our research; in particular, in Section 2.1 we introduce the concept of rhizome interpreted as mobility networks of people; in Section 2.2, we shortly present the technique for itemset mining; in Section 2.3, we present the related literature as far as analysis of Twitter messages for studying mobility is concerned. In Section 3, we investigate how to apply rhizomes and itemset mining to study how tourists experienced a given city and its surrounding regional area. Specifically, Section 3.1 briefly presents the research project and details the case study that we considered for this research, along with the data set we built to validate the approach; Section 3.2 explains the methodological approach; Sections 3.3–3.5 provide three different specific interpretations of the concept of rhizome based on itemset mining and investigate how discovered itemsets should be visualized, both by adopting a topographic representation and a topological one, to let mobility networks visually emerge. Section 4 presents the application of two of the three discussed interpretations to a large data set of traces, in order to show the potential application of the presented approach to study traces of people gathered through social media (in particular, through Twitter). Finally , Section 5 draws the conclusions, by discussing learned lessons and open issues.

2. Background

In this section, we describe the multidisciplinary background of our work. In fact, the research activity is across geography and computer science. For this reason, we need to present both the basic geographic concept we are investigating (Section 2.1) and the notion of itemset mining (Section 2.2).

2.1. From the Actor-Network Theory (ANT) to the Concept of Rhizome

The *Actor-Network Theory (ANT)*, elaborated by Bruno Latour [6,7], can be considered the first approach adopted in a multidisciplinary way by sociology, geography and other disciplines, making networks the focus of attention for the study of spatialities. Geographers have adopted ANT because it goes much further than traditional absolute analysis of space; through ANT, geographers promote a relational approach, based on networks and relations among humans and non-humans. Bruno Latour claimed that this division needs to be overcome; the perspective is the world as it presents itself to us in the form of networks, relations and hybrids that cross the artificial boundaries drawn between culture and nature, and between the worlds of people and of things (for a clear analysis of the Actor-Network Theory, see [8–12]).

ANT is an ideal framework to deal with spatiality, because it clearly directs our attention to the effects produced by the fluidity of spatial configurations of a variety

of actors. However, from our perspective, ANT is also important for understanding the effects of their connections in movement. Thus, ANT is an excellent framework to describe the complex and mutable composition of networks of heterogeneous actors, especially in movement. In this respect, Sheppard argued [9] that networks are non-hierarchical spaces, that is, spaces that are not important for quantitative reasons, such as the area, the population or the metric distance (the largest regions, the most inhabited places, the nearest places to cities and so on) that can be analyzed in a vertical way. In contrast, networks are spaces that are important for their connectivity and networking in a horizontal way. This means that all the involved human and non-human actors have the same importance, producing a lack of attention to their internal differentiation. On the contrary, relations in a network are not all the same, and their differences produce different spatialities [10].

On the basis of the network perspective introduced by ANT, Jacques Lévyi and Michel Lussault in [12] have proposed a more specific approach that focuses on networked spaces produced by humans and things in their movement. The two French geographers have started an important phase of social sciences of space, based on the importance of movement in the globalized world. In particular, they claimed that "contemporary urban" assumes a poly-centric and reticular configuration: it is no longer divided into center and periphery; rather, it is viewed as an "osmotic-centered system" of mobility; in fact, contemporary urban is inserted into a globalized network, where local scale and global scale interact by reconfiguring centrality and axes, and internal and external connections of the city [13]. The creation of networks among the multiple places of contemporary urban is one of the processes that characterize the mobility of inhabitants, and more in general, of any kinds of city users: a new reticular dimension emerges, based on connections activated among places, exploited by individuals in their life experience; connections could be either real (transportation infrastructures) or virtual (information published either on the web or on social media about places, possibly produced by citizens). Such networking, produced by experience of individuals in urban space, is termed *rhizomatic*, by resuming a concept born in the field of botany and then re-elaborated in the philosophical field: "Compared to centric (even poly-centric) systems, hierarchical communication and predetermined connections, a rhizome is an a-centric, non-hierarchical and not meaningful system" ([3], page 33). The concept of rhizome was refined in spatial terms by Jacques Lévy: "A rhizome is the space of individual action in mobility, but also in the multiform relationship with other individuals" ([14], page 19). A further definition could be the following: "A rhizome is a family of networks, characterized by the absence of identifiable boundaries and a meeting between topological metric inside and topographic metrics outside" ([14], pages 18–19). In other words, a rhizome belongs to the topology metric that is to a discontinuous space, based on nodes and connections that produce a network without beginning, without end and without well-defined boundaries, because it is the result of the experience in space of individuals.

In order to explain our view of the concept of rhizome, consider three cities denoted as A, B and C. First of all, notice that the concept of distance is not only a matter of metric distance, but it is also a matter of accessibility. Suppose that the distance between the city centers of A and B is 30 km, but there is a train connection with fast and frequent trains that allow people to move from A to B and vice versa in 20 min every 15 min. Suppose now that the distance between the centers of A and C is only 10 km, but only a mountain road connects them, without public transportation: people without cars cannot reach A from C. Thus, the availability of an efficient transportation infrastructure makes cities A and B "closer" than cities A and C. Consequently, transfers from A to B and vice versa are more frequent than transfers from A to C and vice versa; thus we expect that, by analyzing data of moving people, cities A and B should be more frequently associated together than cities A and C. Thus, a metric distance is not always a valid parameter because accessibility is more important.

Figure 1 depicts the situation: on the left-hand side, the topographic representation

shows that cities A and B are farer than cities A and C (based on their metric distance). On the right-hand side, the topological representation shows that cities A and B are much more connected than cities A and C: the points representing cities A and B are closer than the points representing cities A and C; furthermore, the line connecting cities A and B is thicker than the line connecting cities A and C. Thus, the reader can have a concrete shape to what was argued by Levy ([14], pages 18–19): the rhizome is topographic outside-, i.e., places lived by people are physically located in the space; nevertheless, the rhizome is topological inside, because the strength of connections does not depend only on the physical distance between the connected places. If we consider connections between many places, both the topographic and the topological representations assume a reticular shape, where some place can be more attractive than the other ones, but no center clearly emerges; the reader can find examples in Section 3.

Figure 1. A topographic representation of connections among cities A, B and C (left side) and a topological representation of connections among the same three cities based on strength of connections (right side).

Another aspect of interest is related to the fact that connections are not only related to physical accessibility but also to virtual connectivity. One example is related to promotional activities: suppose that in city A there is a restaurant whose fame is increasing, because some people living in city B are promoting it through social networks; friends of promoting people (living in city B too), could decide to go to that restaurant as well. This consideration has an impact on the visualization strategy to adopt for visually analyzing connections. Again, it appears that a topographic representation is necessary but it is not enough to enhance the strength of connections; a topological representation in which two very connected cities are depicted closer than two loosely connected cities (as in the right-hand side of Figure 1) could provide a very useful visualization perspective, able to help analysts understand networks of connections.

By moving from the above considerations, it is possible to guess that "a rhizome is the space with a set of places frequently lived by a single person and by many people, on the basis of material and virtual connections among them".

2.2. Mining Itemsets and Association Rules

Since the 1990s, a large number of data mining techniques have been developed to address a large variety of problems. One of the most famous data mining techniques is called *mining of association rules* [4]. Born for market-basket analysis, its original goal was to find frequent associations of (sold) items, i.e., items that frequently appear together in (commercial) transactions. An association rule has the form $\{A, B, C\} \rightarrow D$, meaning

that when a customer buys products A, B and C, he/she also buys product D. Each rule has two numerical weights, called *support* and *confidence*. The *support* is the percentage of transactions that contain the rule. The *confidence* is the conditional probability that the whole rule is found in a transaction having found the body (in the sample rule, the body is $\{A, B, C\}$, while D is the head). Since the number of rules could be extremely high, it is necessary to prune the search space, in order to get only rules that could be really meaningful; pruning is done by setting a minimum threshold for support.

$\{A, B, C\}$, $\{A, B, C, D\}$ and so on are called *itemsets*; in particular, $\{A, B, C\}$ is called 3-itemset because it contains three items, while $\{A, B, C, D\}$ is called a 4-itemset because it contains four items. Notice that the rule $\{A, B, C\} \rightarrow D$ is obtained from itemset $\{A, B, C, D\}$; this means that both the itemset $\{A, B, C, D\}$ and the rule $\{A, B, C\} \rightarrow D$ have the same support. Given a minimum threshold for support named $minsupp$, itemsets having support greater than or equal to $minsupp$ are called *large itemsets*.

The basic step to compute association rules is called *itemset mining*: an algorithm extracts those sets of items that frequently appear together in transactions (large itemsets). This problem is not so easy to solve, particularly when the data set is very large and the number of items per transaction is large too. The reader can refer to [15,16] for some well-known algorithms developed to efficiently mine large itemsets. In this work, we used the implementation developed within the *Hints from the Crowd* project [17], which is a main memory algorithm able to deal with generic items. A reader willing to experiment the approach can exploit any algorithm available on the Internet for mining itemsets or association rules; a few lines of code written in some procedural programming language will usually be enough to pre-process the data set, so as to transform it into a format that is suitable for the specific implementation of the algorithm.

In order to detach from the context of commercial transactions, we can give the following generic formulation of the problem of itemset mining, originated from the semantics of the `MINE RULE` operator introduced for relational databases [18] and for XML databases [19].

Definition 1 (Data set and items)**.** *Consider a data set $\mathcal{D}$. This is a set of groups; i.e., $\mathcal{D} = \{g_1, g_2, \ldots, g_n\}$. Each group g_i is, in turn, a finite set of items, i.e., $g_i = \{a_{(i,1)}, a_{(i,2)}, \ldots\}$. Groups are not disjoint; i.e., they can share items.*

As an example, in the case of market-basket analysis, items are products, while groups are single commercial transactions.

Definition 2 (Large itemset)**.** *An itemset $h = \{a_1, a_2, \ldots\}$ is a finite set of items that appear together in some group in $\mathcal{D}$. The support of h is the number of groups in $\mathcal{D}$ that contain the whole itemset, divided by the total number of groups in $\mathcal{D}$. Formally,*

$$supp(h) = |\{g \in \mathcal{D} \mid g \cap h = h \wedge h \neq \emptyset\}| / |\mathcal{D}|.$$

Given a minimum threshold $minsupp$ such that $0 \leq minsupp \leq 1$, the itemset h is said to be large if its support is greater than or equal to the minimum threshold, i.e., if $supp(h) \geq minsupp$.

After these premises, we can say that the problem of large itemset mining is to compute all large itemsets from within a data set $\mathcal{D}$, given a minimum threshold for support denoted as $minsupp$.

As an example, consider the data set reported in Table 1. If we set $minsupp = 3/5 = 0.6$, we obtain the large itemsets reported in Table 2. Notice itemset h_8, with support $3/5 = 0.6$, that appears in three groups, i.e., groups g_2, g_4 and g_5. This is the itemset with the largest cardinality (number of items) that we can extract from the data set: itemsets having a larger number of items have no sufficient support.

Table 1. Sample data set D, organized in groups and items.

Group ID	Items
g_1	C, D, E, G, H
g_2	A, B, C, E, F, G
g_3	E, G, H
g_4	B, C, F, G
g_5	B, C, G

Table 2. Large itemsets, con minsupp $= 3/5 = 0.6$.

Itemset ID	Items	Support
h_1	B	3/5
h_2	C	5/5
h_3	E	3/5
h_4	G	5/5
h_5	B, C	3/5
h_6	B, G	3/5
h_7	C, G	4/5
h_8	B, C, G	3/5

2.3. *Related Work on Analysis of Twitter Messages for Studying Mobility*

To complete the background of the paper, we consider the secondary contribution of the paper, i.e., how social media can push data-driven approaches to study mobility of people. Notice that we focus on works made on data gathered from Twitter, which are numerous: this is due to the fact that Twitter API does not pose any obstacle to gather data, while the other social media usually do.

Many studies have been published concerning analysis of tweets, particularly for studying mobility. In fact, researchers of many human sciences consider now micro-blogs (such as Twitter) a precious source of information for their studies. In fact, as stated in [20], people blog to provide a record of their life to share with followers. Obviously, it is necessary to be aware of doubts concerning the representativeness of data obtained by analyzing traces of Twitter users [21], because they represent only Twitter users, which are a subset of all moving people, and they are a subset of social media users.

Anyway, traces of Twitter users can complement other sources of information, in order to let a possibly unexpected perspective about studied phenomena emerge. Furthermore, often traces of Twitter users are the only source of information that describes paths of moving people, as in the context of tourism [22,23]. Notice that many researchers are interested in exploiting social-media sources to study how people move. For example, in [24] the authors studied how cities influence mobility, by defining statistical metrics of centrality that they applied to data produced by Twitter users. In [25], the authors analyzed digital footprints, such as data produced by phone networks, by cross analyzing them with geo-referenced photos posted on micro-blogs, to study the movement of tourists during their visit to Rome (Italy).

On a world-wide scale, geo-located posts by Twitter users can be also used to study global mobility patterns. In [26], statistical approaches were proposed to study country-to-country patterns, by considering several aspects, such as country-to-country networks, temporal patterns of mobility and so on. Furthermore, mobility patterns could be very useful to analyze migrations as well. In [27], the authors addressed the problem of understanding which countries migrants to the EU actually come from. They adopted a clustering algorithm with the goal of discovering the provenance of migrants, in spite of the (possibly false) countries they declared they came from.

The evolution of a city to become a *smart city* can be significantly fostered by analyzing Big Data (in general) and traces of Twitter users (more specifically). In [28], the authors tried to relate the choices made by tourists in the pre-trip phase with the experiences they

shared on social media (post-trip). In [29], the authors tried to create value by analyzing social media: by estimating kernel density and latent Dirichlet allocation, they showed that it is possible to investigate how social media can provide a platform to develop smart services for urban tourism. Always through social media analysis, in [30] the authors explained that Big-Data analysis can help improve decision-making processes, and create marketing strategies with more personalized offers. Additionally, in [31], the authors explored how the use of intelligent tourism technologies such as travel-related websites, social media and smartphones in travel planning, can improve traveler satisfaction. Micro-blogs can be effective also to influence tourists when they form their perceptions of the chosen destination; in [32], the authors tried to demonstrate, through the use of *Sina Weibo*, a Chinese micro-blogging site, how the choice of a touristic destination is influenced by information published on the social network.

Traces generated by Twitter users were used to analyze flows of people. For example, in [33] the authors exploited traces of Twitter users to study flows within a city. They relied on a topographic representation of flows, and applied clustering techniques to identify places that were more active, i.e., where Twitter users mostly posted tweets. However, they did not exploit a topological representation and did not use itemset mining. At a regional scale, traces of Twitter users were used to analyze traffic [34], in order to study critical areas and routes on a spatio-temporal basis, i.e., correlating traffic jams, routes and time. In this work, the authors adopted clustering techniques and topographic representations.

An interesting paper that proposed an approach significantly related to our approach is [35]. The idea of the authors was to study mobility of individuals by clustering them instead of locations, in order to discover moving patterns. They adopted a method that clusters individuals having similar visitation rates for each location. Our approach is similar, since we try to associate places and transfers that frequently appear together in traces of Twitter users; however, the adoption of itemset mining provides a different perspective.

A paper that addresses the study of urban characterization is [36]. The authors partitioned the space by adopting a clustering method that exploits density of tweets posted in the different areas. Then, numerical features based on the number of tweets in the different areas and the number of moving users were computed and temporally evaluated. With respect to our approach, it can be considered complementary, although focused on similar themes.

We also want to highlight that many approaches can be adopted for studying traces of Twitter users. One choice could be to develop automatic clustering techniques. However, specific techniques must be developed in this respect, because traces are sequences of visited places. In [37–39], a clustering technique for traces of Twitter users (or trips) is proposed: the idea is to evaluate different similarity metrics between trips that were previously geo-partitioned on the basis of categorical coding systems (such as ZIP code) of the area containing coordinates denoting geo-coding of messages. Then, a multi-level fuzzy-clustering algorithm is applied to discover clusters of most popular trips. In those works, the reticular view of lived space is not considered, whereas that is the goal of this paper.

The technique of itemset mining was used in other works to analyze micro-blogs. For example, in [40], the authors used this technique to extract patterns from messages, in order to use frequent itemsets for query expansion, when users formulate queries to find posts of interest. Similarly, in [41] the authors adopted itemset mining for opinion mining and sentiment analysis of micro-blogs. Specifically, they proposed an opinion-descriptive model, that is the basis for an opinion mining method. This way, posts in the micro-blogs are classified on the basis of their sentiment.

To the best of our knowledge, itemset mining has been rarely used to study how people live space. We found only [42], where the authors adopted itemset mining to address the problem of *geo-social co-location*. Suppose that people that are found in the same place in a given time slice are described by features such as the university they are studying in, the course and so on; the itemset-mining approach is used to find out the most

frequent associations of personal features that characterize people that frequently are in the same places. For example, if data describe students of universities, features are the names of the universities; thus, the results are the set of university names whose students are often found in the same places. In our work, we adopt the opposite approach: we want to obtain places that are often visited in the same trip by many tourists.

The work in [43] addressed the problem of analyzing micro-blogs for business applications, namely, context-aware service profiling. To do that, they introduced the notion of *strong generalized flipping itemset* that is able to highlight the existence of outliers in terms of the polarity of a relationship between concepts extracted from messages. In fact, the idea is that, given a generalized itemset (obtained, given a taxonomy, at a level higher than the leaf level) and its polarity (positive, negative or neutral), if one or more of its descendant itemsets (i.e., extracted at a lower level) shows a different polarity, this means that an anomaly is occurring. Again, they did not work with geo-location of messages to study mobility of users.

Another study that applied itemset mining to micro-blogs was [44]. Specifically, the authors addressed the problem named *WTF* (*who to follow*): the idea is to recommend to users, other users to follow, on the basis of the topical users (popular users such as singers or politicians) and the semantic categories topical users belong to. The authors exploited itemset mining to profile users on the basis of semantic categories associated with topical users they follow. The work is quite interesting, but they considered neither posted messages nor geo-location.

3. Rhizomes as a Data-Driven Approach to Discover Mobility Networks

In this section, we present the main contribution of the paper. In Section 3.1, we preliminarily present the research project in which we had the ideas discussed in this paper, and the case study we considered to experiment the approach. In Section 3.2, we introduce the methodological approach we identified. Then, in Sections 3.3–3.5 we separately discuss three different interpretations of the problem.

3.1. Research Project and Case Study

The research work presented in this paper is part of a research project led by University of Bergamo (Italy) called *Urban Nexus* [5]. The goal of the project is to develop a methodology and tools for studying mobility of city users. The novelty of the approach is the exploitation of geo-tagged messages posted on Twitter, to discover movements of city users.

Since the project involved universities and city councils of three tourist-oriented cities, i.e., Cambridge (UK), Lausanne (Switzerland) and Bergamo (Italy), we decided to define a case study related to tourism. In fact, city councils of tourist-oriented cities lack tools to study mobility of tourists, in order to comprehend and create territorial assets capable of attracting tourists.

Geo-tagged tweets possibly posted by tourists were gathered through the *FollowMe* suite [45,46]. It is a software tool that we developed to gather data of moving people from Twitter. By exploiting Twitter API, it detected possibly traveling people by looking for messages posted in an airport area; then, such Twitter users were tracked for the next 8 days, by gathering their geo-located posts by directly accessing their timelines, i.e., the history of their posted messages.

We started gathering traces on May 1st, 2015, to the end of studying people visiting EXPO 2015 held in Milan (Italy). Gathering was stopped at the end of 2018; we collected around 3,000,000 messages, describing people traveling all over the world.

We collected geo-located posts only from Twitter, because it is the only social network that does not provide a privacy mechanism: in fact, through Twitter API, any application can collect messages posted by any user. Other social media pose obstacles to do that: some require applications to be explicitly approved; others ask for an explicit consensus by users to share their messages.

During the project, we developed many tools for analyzing the gathered data. In particular, we built a tool for visual analysis of traces and messages, called *Treets* (see [5]), a fuzzy clustering algorithm to aggregate traces based on ZIP codes [37–39], and a framework for manipulating collections of geo-tagged JSON data sets (named *J-CO*, see [47,48]).

In order to have a concrete case study, we considered the city of Bergamo (Italy). Bergamo is a city with 100,000 inhabitants, located 40 km north-east from Milan (Italy). We chose Bergamo because we know the city (we work and live in Bergamo), so it is the perfect choice to validate results. Thus, the case study can be formulated precisely as "studying mobility networks of tourists that experienced the area of Bergamo and the surrounding regional area".

By means of the above-mentioned tools, we built the case-study data set. We decided to keep its size compact, so that it is easy to understand what the itemset-mining technique produces. In fact, the relatively small size allowed us to easily inspect the results and the data set jointly, to validate results. Starting from the whole set of 3,000,000 posts that describe 400,000 world-wide trips, we selected trips that involved Bergamo. Among all these trips, we selected those trips that contain significant text in messages, in order to validate them: in fact, it was important for our research project to be able to characterize messages with respect to the location they were posted from (for example, whether the message tells something about a place close to the geo-location of the message, whether the message provides meaningful information or details and so on and so forth). This way, we were able to extract traces of real tourists.

After this filtering activity, we obtained the case-study data set, which contains 38 trips; the total number of places (locations) in trips is 711; the average number of places per trip is 19. Places are not only in the Bergamo area, but also in Milan, and some easily reachable cities, such as Verona. We consider this a good size for the case-study's data set. Table 3 reports the most popular places among the 38 trips, i.e., places that occurred in at least 4 trips. We can see that, in spite of the relatively small size of the data set, we can get a significant number of itemsets to validate our practical interpretations of the concept of rhizome.

Table 3. Places the appear in the case-study data set, with the number of trips in which they occur.

Place	Number of Trips
Bergamo Railway Station	30
Bergamo Cathedral	24
Orio al Serio Airport	23
Milan	21
Città Alta	16
Milan Malpensa airport	16
San Vigilio Hill	15
Bergamo	11
Piazza Vecchia	10
Basilica di Santa Maria Maggiore	10
Funicolar Città Bassa	8
Colleoni Chapel	7
Dome Square	6
Berlin	6
Pontida Square	6
Campanone	6
Fornaio	5
Funicolar San Vigilio	5
Vineria Cozzi	4
London	4
Porta San Giacomo	4
Meina - Lago Maggiore	4
Lorenzo Mascheroni Square	4
Birreria di Città Alta	4
Il Maialino di Giò	4

The size of the case-study data set is good for proving the concept, which is the goal of the paper. Obviously, in order to provide geographers with significant information to study how tourists live the city, the size of the data set to study should be larger; a larger data set will be considered in Section 4.

Before continuing, we want to highlight some critical issues concerning the data.

The first issue concerns the capability of messages posted by traveling people to describe all places they visited. This is due to the fact that they do not post geo-located messages in every single visited place; in contrast, we can expect that they post messages from places that they find particularly interesting to their eyes. Thus, it is clear that collected traces are incomplete by nature; anyway, we can figure out that this is not a problem but a positive aspect, for analysts that are interested in discovering virtual connections among places that were judged as interesting by tourists.

Another issue to consider is the size of the case-study data set, in relation to itemset mining. Itemset mining was born to extract frequent association rules from large data sets, so it can appear strange to adopt it on a data set that contains only 38 trips. Nevertheless, we do not think that the outcomes from this data set could be considered relevant for analysts or for decision makers. The data set was built to prove our ideas about the concept of rhizome interpreted by means of itemsets. Much larger data sets would not allow for deeply understanding the effectiveness of the three interpretations, because there would be too many itemsets to validate on thousands of places. Remember that the focus of the paper is not to give decision makers indications to make decisions; the goal of the paper is to validate a concept, to further apply it on larger and/or different data sets.

3.2. Methodology

In the Urban Nexus project, we built the case-study data set as previously described, because we had to focus on movement of tourists. Specifically, we defined a methodology to discover how tourists experienced the territory, in order to discover (possibly virtual) connections between places both inside and outside the observed territory.

We understood that the network-based approach introduced by the Actor-Network Theory (ANT) and declined as rhizome was promising. However, we realized that a practical interpretation of the concept of rhizome was missing: although the concept is intuitively a visual concept, a concrete shape for it has not been provided yet. Consequently, we decided to investigate how to give a concrete shape to the concept of rhizome, by applying it to study mobility of tourists, which is the contribution of the paper.

In particular, one specific aspect had to be addressed: how to evaluate the strength of connections between places. To solve this problem, we had the intuition of adopting itemset mining, which provides associations of items often present together in groups (see Section 2.2), by applying it to trips of tourists. The intuition is that the reticular view behind the concept of rhizome should let associations of places that are common among tourists emerge, and common associations of transfers made by tourists from one place to another. However, this is the same idea behind itemset mining applied to commercial transactions; by using the generic formulation of the problem of itemset mining reported in Section 2.2, if groups are viewed as trips and items are viewed either as places or as transfers, we can discover the most frequent associations of them; in fact, if one single tourist visits three places during the trip, this can be an outlier, originated by his/her particular interests; in contrast, if many tourists visited the same three places, this is not a case and can reveal a virtual connection among places. Consequently, the greater the number of itemsets a pair of places appears in, the stronger their connection.

Thus, the research question we addressed can be summarized as follows: "can itemset mining be effectively used, together with proper visualization techniques (topographic and topological representations) to give a concrete visual shape to rhizomes, applied to give a methodology for discovering the mobility networks of tourists by analyzing geo-located tweets posted during their trips?"

The methodological approach we followed to address the research question is the following one:

1. We identified three different interpretations for the concept of rhizome on the basis of itemset mining;
2. For each interpretation, we applied the algorithm for itemset mining and visualized the results both by means of a topographic representation and a topological representation;
3. We studied the effectiveness of each interpretation, based on both the topographic representation and the topological representation for letting the rhizome of tourists emerge.

Each interpretation is tied to a specific perspective we can adopt to inspect the data set: by associating visited places (Interpretation 1, presented in Section 3.3); by associating transfers from one place to another (Interpretation 2, presented in Section 3.4); by associating transitive transfers from one place to another (Interpretation 3, presented in Section 3.5).

To let virtual connections emerge (recall Section 2.1), locations outside the surrounding regional area of the city of interest are collapsed either to the center of the city whose area contains the location, or to a specific point of interest near the location (such as an airport). This mixed focus of the analysis has the following rationale: we want to simulate a typical interest of local administrators, i.e., understanding connections of specific places in the administrated city with respect to the surrounding territory. Obviously, the mixed focus stresses the capability of the approach.

Specifically, we performed geo-coding of tweets as reported hereafter.

- *Geo-coding with specific places.* Depending on the specific points where a post is sent from, coordinates of points sent from the same place are different; furthermore, errors performed by GPS antennas could also affect precision. Consequently, it is not the case to refer to pure coordinates: it is necessary to geo-code tweets, by associating them to places they were posted from.
  Not only, not necessarily tweets concerning a given place are posted from the place they refer to: they can be posted form a street or square close by the place, for example, a picture of a building is taken and posted from the square in front of it. Thus, the place the post talks about is the building, not the square.
- *Airports and stations.* These are places with large areas; it is not at all relevant in what part of an airport or station a person posted a tweet. What really matters, as far as the analysis is concerned, is the fact that the user was in the airport/station.
- *Default coordinates.* If the geo-location service is disabled on the smartphone, the Twitter app associates default coordinates to geo-located tweets, i.e., coordinates of the city center. In this case, it happens that many posts are conventionally located in the city centers; if their texts do not mention any specific place, it is not possible to associate them to a specific point of interest.
- *Geo-coding with city center (and city name).* In our experiments, tweets posted from cities outside the area of interest were geo-coded with the name of the city, without distinguishing with respect to the places they actually referred to. Of course, we are aware that this choice affects the results of the analysis, but it is coherent with the methodological goal of our work, i.e., studying connections of places in the area of interest with cities in the surrounding areas. In case an analyst were interested in studying connections with specific places, it would be enough to change the label assigned to tweets actually posted from or talking about this place and repeat the analysis.

Notice that the above-mentioned activities could be performed in an automatic way, in case of very large data sets containing, e.g., millions of geo-located messages, by exploiting thesauri about public places. Of course, in this case we can expect some case of wrong geo-coding, that should be compensated by the size of the data set and by suitable

minimum thresholds for extracting large itemsets. The contribution of the paper is not concerned with gathering, pre-processing, completeness and quality of data, but with proving the concept of rhizome.

Nevertheless, it is worth noticing that every decision about the way messages are geo-coded affects the outcomes of the analysis. However, it is not possible to give a unique and definite guideline, because this depends on specific needs. As far as the specific decision we made to prepare the case-study data set, they were motivated by the need to understand the potentiality of our ideas, i.e., the adoption of itemset mining to give a concrete shape to the concept of rhizome. In other words, given geo-coded traces of moving people, whose completeness and quality are accepted by analysts, the approach can be applied to study their common movements.

3.3. Associating Visited Places (Interpretation 1)

A trip is a finite sequence of places $t_i = \langle p_{(i,1)}, p_{(i,2)}, \dots \rangle$ (where $p_{(i,j)}$ denotes a place). We look at it as a finite set of visited places, irrespective of the temporal factor. By applying itemset mining, it is possible to formulate the problem as follows: *we want to extract all significant associations of places, such that associated places are visited by the same tourists*.

We can formalize the itemset mining problem as follows:

- The data set $\mathcal{D}_1$ is a set of groups, i.e., $\mathcal{D}_1 = \{g_1, g_2, \dots\}$.
- A group $g_i = \{a_{(i,1)}, a_{(i,2)}, \dots\}$ corresponds to a trip $t_i = \langle p_{(i,1)}, p_{(i,2)}, \dots \rangle$; an item $a_{(i,j)}$ in g_i is a place visited by trip t_i, i.e., $a_{(i,j)} = p_{(i,k)}$, with $p_{(i,k)} \in t_i$.
- Set a minimum threshold for support that provides a representative set of itemsets.

This way, an itemset describes places visited together by a few tourists. The greater the cardinality of an itemset and its support, the higher the relevance of the itemset. The pool of extracted itemsets is an interpretation of the concept of mobility network.

Expectations. By means of this interpretation, mobility networks of tourists are seen as the frequent associations of places visited by tourists; this means that the association is relevant for tourists and characterizes the way they visit a city.

From the visual analysis of extracted itemsets, we expect to obtain the following outcomes:

- We expect the topographic representation to provide a clear delimitation of the areas covered by mobility networks;
- We expect the topological representation to provide the strength of the interconnections between single places.

In Section 3.3.1, we discuss whether the expectations were confirmed in the case study.

3.3.1. Case Study: Topographic and Topological Representations

We ran the algorithm for itemset mining on the case-study data set with the minimum threshold for support set as: $minsupp = 2/|\mathcal{D}_1|$; we obtained 1284 itemsets, distributed as reported in Table 4. This choice, which allows for extracting itemsets shared by at least two tourists, is determined by the limited number of trips in the case-study data set. If the analyzed data set had thousands of trips, the choice could be at least 100 tourists (i.e., $minsupp = 100/|\mathcal{D}_1|$). The right threshold can be obtained only by starting with a very high threshold value and by progressively reducing it, stopping when a reasonable number of itemsets (not too few and not too many) is obtained. What do we mean with "a reasonable number of itemsets"? It is difficult to provide an absolute number. Certainly, too many itemsets do not provide information, because many itemsets would have low support and low significance; too few itemsets would let only very strong associations emerge. The investigation necessarily requires one to perform several attempts, so as to tune the minimum threshold for support, not only based on the number of extracted itemsets but also based on the informativeness of representations.

Table 4. Distribution of itemsets for interpretation 1.

# of Items in the Itemset	Support (# of Tourists)	# of Itemsets
2	2	184
2	3	48
2	4	19
2	5	3
2	6	2
2	7	1
2	8	2
2	10	1
3	2	342
3	3	26
3	4	2
4	2	303
4	3	2
5	2	199
6	2	102
7	2	38
8	2	9
9	2	1

Figures 2 and 3 adopt the topographic representation to provide a network view of associations: each place (point) in an itemset is connected with all other places in the same itemset; each itemset is depicted with a different color. The thickness of the itemset, i.e., the support in terms of number of tourists that share the same association of places, corresponds to different thickness: the greater the number of tourists sharing the same places, the greater the thickness of the lines depicting the same itemset. From Figure 2, we focus at the level of regional territory surrounding Bergamo area. Recall that all places in Milan and, in general, other cities outside Bergamo area, were geo-coded with the same place, i.e., either the center of cities or the center of airports, in order to constitute a single item. In Figure 3 we focus on the city of Bergamo.

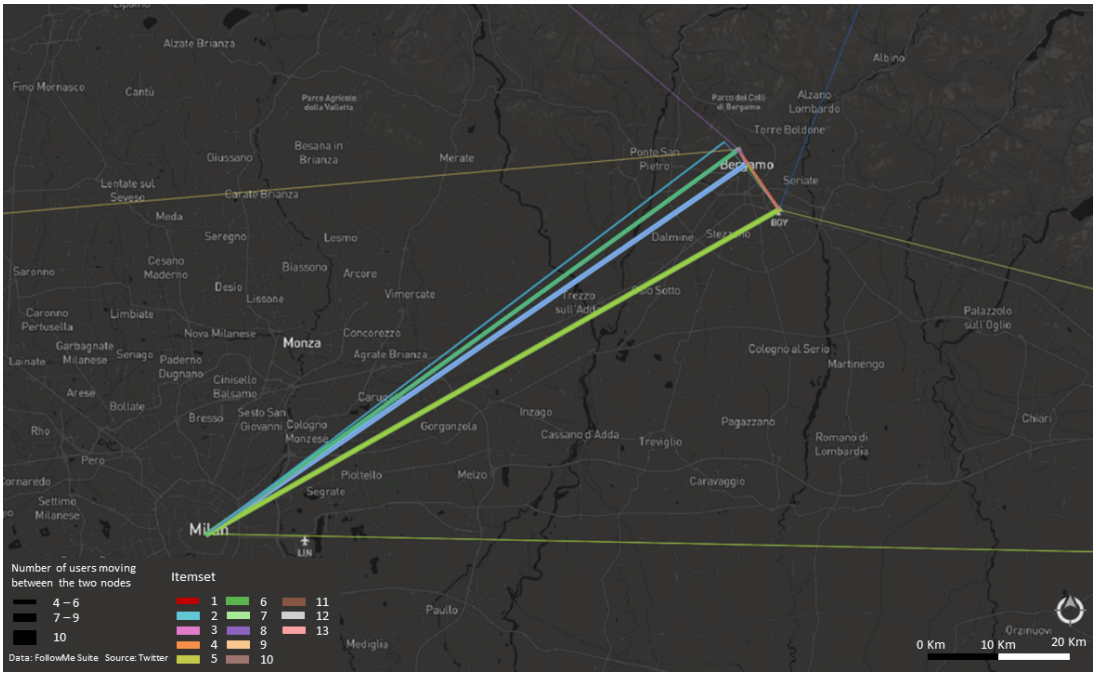

Figure 2. Topographic representation for discovered itemsets associating places (Interpretation 1), focused at the level of surrounding regional territory of Bergamo.

Figure 3. Topographic representation for discovered itemsets associating places (Interpretation 1), focused at the level of Bergamo area.

The topological representation depicting itemsets is shown in Figure 4. The goal is to let the strength of associations between places emerge clearly. In this representation, the sizes of points and lines are proportional to their relevance. In particular, the size of a point, which depicts a place, is proportional to the number of tourists that visited that place. The strength of a line is proportional to the support of the itemset; the support is the number of tourists that share the association of places. Positions of places are not casual: we adopted a tool called `D3.js` (URL: https://d3js.org/), which is an off-the-shelf generic JavaScript library, which determines the position of two points on the basis of the strength of their connection: the stronger the connection, the closer they are. In practice, two points attract themselves if they are strongly connected. This way, the analyst can get a visualization of connections that clearly put strongly connected places together.

Outcomes from the case-study data set. Consider the topographic representation reported in Figure 2: The map covers Bergamo area (on the right-hand side) and Milan area (on the left-hand side), but a few trips involve also the city of Verona (east side), which is not reported in the map. It is possible to see how many tourists visited several places in Bergamo and visited Milan too: this is an example of virtual connection emerged from the data.

From Figure 3 (obtained by zooming in Figure 2 on the Bergamo area), we focus on the local area of Bergamo. It is possible to see that not all tourists in Bergamo posted messages concerning the same places (notice the different lines having different colors). We can argue that tourists are interested in different touristic attractions. Notice that the alignment of places of interest in Bergamo (from south-east to north-west) does not help in visualizing associations on the map.

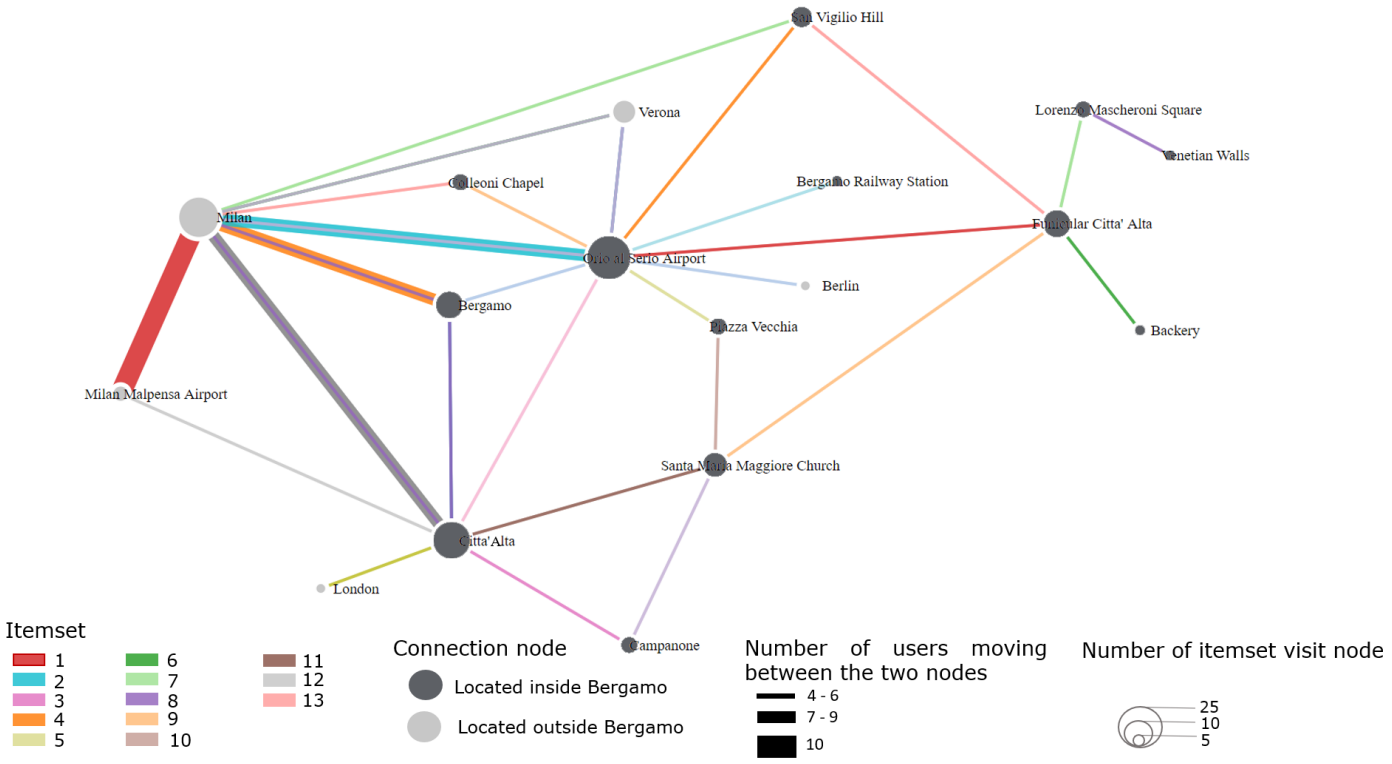

Figure 4. Topological representation depicting the itemsets obtained by running the itemset-mining algorithm on the case-study data set with Interpretation 1.

Consider now the topological representation reported in Figure 4. Notice that the point corresponding to Milan (on the left-hand side) is one of the three biggest points, because we collapsed all locations in Milan to one single point (see Table 3). Similarly, two airports clearly emerge, i.e., Milan Malpensa Airport (on the left-hand side) and Bergamo Orio al Serio Airport (in the center of the figure): in fact, many tourists posted messages while they were at the airport. Notice two places labeled as "Bergamo" and "Città Alta": the first one is associated with any message that does not characterize any specific place in Bergamo, but was posted in Bergamo; the second one characterizes all messages posted in the ancient area of Bergamo (called "Città Alta", i.e., "Upper Town", because it is on the top of a hill) that were clearly posted in that area but do not refer to any specific place.

Notice the very thick line between Milan and Milan Malpensa Airport, which we expected because Milan Malpensa Airport is the most important airport for Milan; another important connection is between Bergamo Orio al Serio Airport and Milan, meaning that many visitors of Milan landed at Bergamo Airport to reach Milan (Bergamo Airport is served by Ryanair). Another important connection is between Milan and the generic place associated with Bergamo, meaning that many tourists visited both Milan and Bergamo. The reader can notice the thin connection with London (on the bottom and left-hand side corner), meaning that a few tourists came from the area of London (UK).

Furthermore, the fact that lines with different colors are overlapped (see, in particular, the two lines connecting Milan and Bergamo) allows analyst to get information about the variety of different itemsets that involve the two places.

3.3.2. Summary of Interpretation 1

We can now summarize what we obtained by experimenting Interpretation 1.

- As far as visualization is concerned, topographic representations are good at getting the spatial focus of connections, because they give evidence of the relative positions of connected places. For example, adding information layers describing roads and railways could help the analysts highlight infrastructures that motivate the strength

of connections. In contrast, the topological representation actually reveals virtual connections, their strength and the mobility networks.
- Interpretation 1 is good at revealing places that are central in many mobility networks of tourists.
- Connections between places emerge (in the topological representation), but neither order nor time are considered.
- Imprecision of geo-coding in tweets affects the meaning of results (note the generic point labeled as "Bergamo", in which we collapsed tweets which are not specifically related to any place in the city).
- The network of connections appears to be quite intricate, especially in the topological representation. This appearance strongly recall the botanic rhizome. Thus, we can say that a mobility network is actually "rhizomatic", so the geographic concept of rhizome clearly assumes a concrete and visual shape.

3.4. Associating Direct Transfers (Interpretation 2)

With Interpretation 2, we look at a trip as a finite set of transfers from one visited place to the next one. By applying itemset mining, it is possible to formulate the problem as follows: *we want to extract all significant associations of direct transfers from one place to another place, such that associated direct transfers are performed by the same tourists*.

We can formalize the itemset mining problem as follows:

- The data set $\mathcal{D}_2$ is a set of groups, i.e., $\mathcal{D}_2 = \{g_1, g_2, \dots\}$.
- A group $g_i = \{a_{(i,1)}, a_{(i,2)}, \dots\}$ corresponds to a trip $t_i = \langle p_{(i,1)}, p_{(i,2)}, \dots \rangle$; an item $a_{(i,j)}$ in g_i is a transfer $a_{(i,j)} = (from_{(i,j)}, to_{(i,j)})$ from place $from_{(i,j)}$ to place $to_{(i,j)}$ such that $from_{(i,j)}$ is the k-th place $p_{(i,k)} \in t_i$ visited in trip t_i and $to_{(i,j)}$ is the $(k+1)$-th place $p_{(i,k+1)} \in t_i$. If a trip t_i contains multiple occurrences of the same transfer, this appears only once in g_i.

Expectations. This way, an itemset describes common transfers performed by a few tourists. The greater the cardinality of an itemset and its support, the higher the relevance of the itemset. The idea is the following: with Interpretation 2, the network is the characterization of how people move through the territory. In case of tourists, we can say that networks of tourists are the frequent direct transfers from one place to another place, which should reveal how tourists move through the territory.

As far as the graphical representations of the itemsets are concerned, we expect that:

- The topographic representation should clearly show frequent direct transfers from one place to another;
- The topological representation should clearly show direct virtual interconnections between places.

3.4.1. Case Study: Topographic and Topological Representations

As for Interpretation 1, we set the minimum threshold for support as $minsupp = 2/|\mathcal{D}_2|$ to run the algorithm for itemset mining on the case-study data set. We obtained the itemset distribution reported in Table 5. Figures 5 and 6 adopt a topographic representation to depict the obtained itemsets. This time, items are arrows, so that arrows having the same color are associated together in the same itemset. The size of points is proportional to the number of itemsets that share the point, and the thickness of arrows is proportional to the number of tourists that share the same set of transfers. Figure 7 presents the topological representation of the itemsets discovered on the basis of Interpretation 2. With respect to Figure 4, lines become arrows, because we now represent direct transfers.

Table 5. Distribution of itemsets for Interpretation 2.

# of Items in the Itemset	Support (# of Tourist)	# of Itemsets
2	2	12
2	3	4
2	4	1
3	2	4

Figure 5. Topographic representation for discovered itemsets associating direct transfers (Interpretation 2).

Outcomes from the case-study data set. Let us analyze what it is possible to obtain by adopting Interpretation 2. First of all, the reader certainly noticed that Interpretation 2 gives rise to a smaller number of itemsets with smaller cardinality, if compared with itemsets obtained for Interpretation 1 (associations of places); in fact, the maximum number of items in an itemset is three, while for Interpretation 1 it was nine. When depicting these itemsets on the map, the effects become more evident. In fact, from Figure 5 (surrounding territory of Bergamo area), the reader can see that now Milan emerges to be strongly connected with Milan Malpensa Airport (green line from the top and left-hand side of the figure), meaning that many tourists that visited Milan arrived at Milan Malpensa Airport. Bergamo (right-hand side of the map) is also well connected with Milan, meaning that a significant number of tourists transferred from Milan to Bergamo, and once in Bergamo, moved within the city.

Focusing on Bergamo area (Figure 6), it is possible to see how tourists moved in the city. Again, recall that we considered only direct transfers, so the number of tourists that made the same direct moves is low.

Considering the topological representation (Figure 7), notice that now the graph is fragmented, places are connected with fewer other places. Again, the connection between Milan and Milan Malpensa Airport emerges, along with the connection between Milan and Bergamo Railway Station: the consequence is that the train connection between the two cities is effective. In contrast, Bergamo Orio al Serio Airport (on the top and left-hand side of Figure 7) emerges to be exploited by tourists who landed in Bergamo Airport and visited Bergamo and its ancient town (Città Alta).

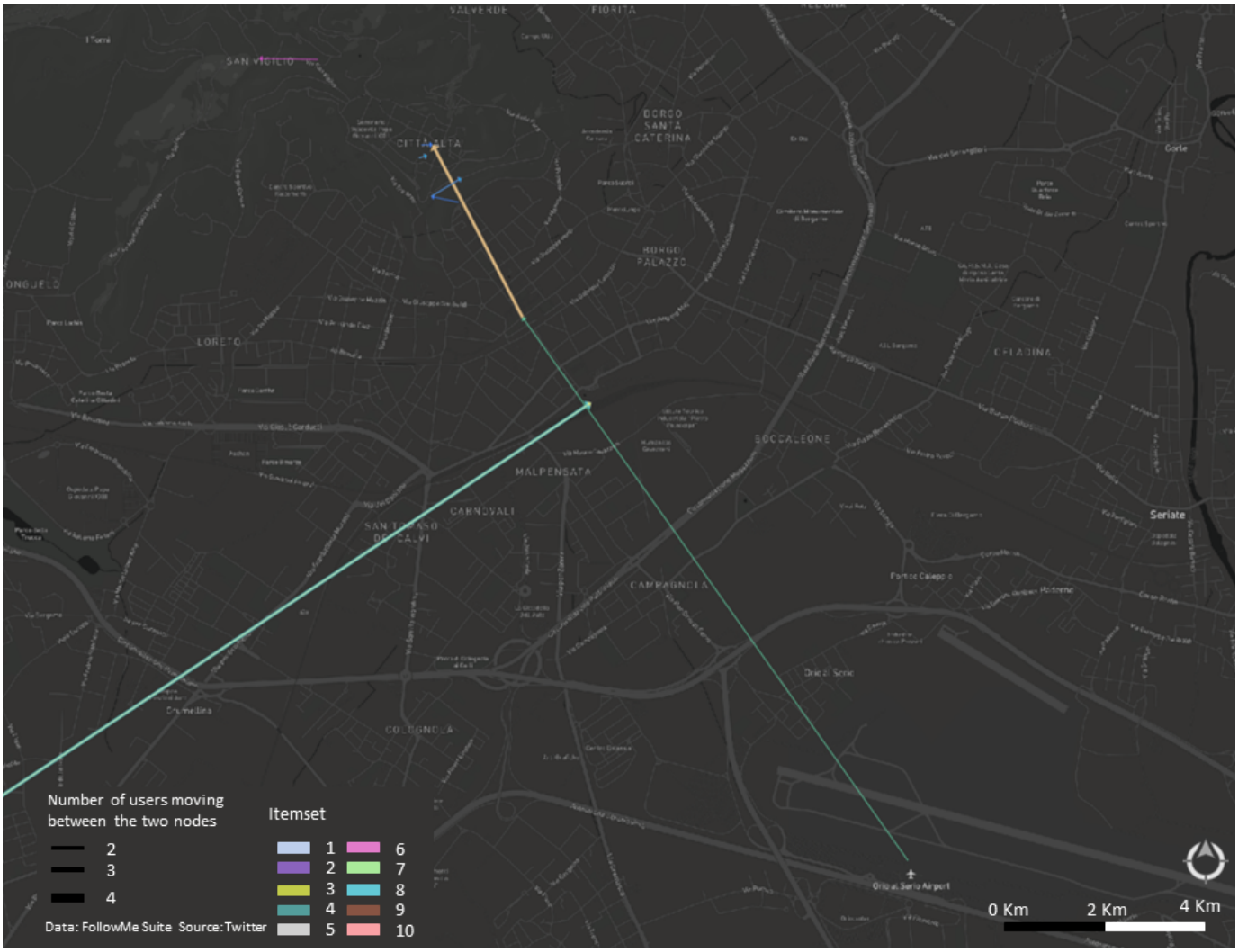

Figure 6. Topographic representation for discovered itemsets associating direct transfers, focused on Bergamo area.

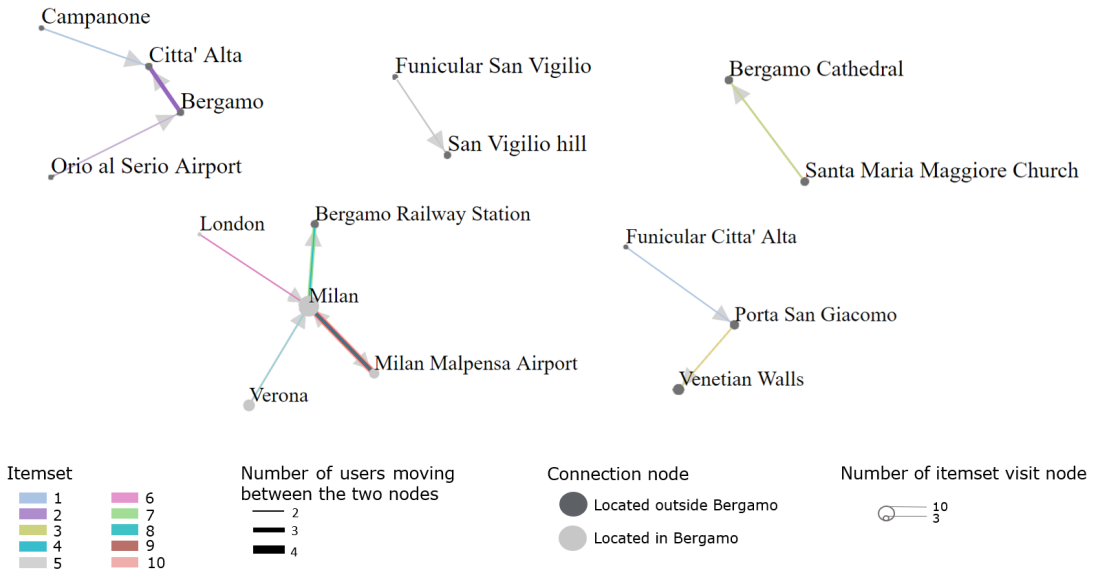

Figure 7. Topological representation for discovered itemsets associating direct transfers (Interpretation 2).

3.4.2. Summary of Interpretation 2

We can now summarize what we obtained by experimenting Interpretation 2.

- With respect to Interpretation 1, now the time dimension of trips is somehow considered, because transfers are ordered on the basis of posting time.
- In the visualization, lines have become arrows, because now itemsets represent a very different concept, i.e., not just associations of places but common transfers from one place to another.
- The number of specific direct transfers commonly performed by single tourists is not high, at least if we maintain the focus at the local area.
- Consequently, Interpretation 2 promises to be effective to reveal mobility networks where nodes are cities, not specific places in cities, i.e., all locations are geo-coded with the center of the city whose area contains the location.

The above-mentioned considerations suggested us to modify the interpretation, by considering not only direct transfers, but also indirect (transitive) transfers.

3.5. Associating Transitive Transfers (Interpretation 3)

Interpretation 2 (presented in Section 3.4) considers only direct transfers, i.e., a transfer from a place $p_{(i,k)}$ to place $p_{(i,k+1)}$ in a trip t_i. However, it is possible to guess that such specific transfers could be not particularly frequent in trips. For this reason, in Interpretation 3 we consider all transitive transfers (obtained by composing direct transfers) in a trip. The itemset-mining problem can be formulated as follows: *we want to extract all significant associations of (either direct or indirect) transfers from one place to another place, such that associated transfers are performed by the same tourists*. This way, we expect common visiting paths to emerge from trips.

We formalize the itemset-mining problem as follows:

- The data set $\mathcal{D}_3$ is a set of groups, i.e., $\mathcal{D}_3 = \{g_1, g_2, \dots\}$.
- A group $g_i = \{a_{(i,1)}, a_{(i,2)}, \dots\}$ corresponds to a trip $t_i = \langle p_{(i,1)}, p_{(i,2)}, \dots \rangle$; an item $a_{(i,j)} \in g_i$ is a transfer $a_{(i,j)} = (from_{(i,j)}, to_{(i,j)})$ from place $from_{(i,j)}$ to place $to_{(i,j)}$ such that $from_{(i,j)} = p_{(i,k)}$ is the k-th place $p_{(i,k)} \in t_i$ visited in trip t_i and $to_{(i,j)} = p_{(i,k+h)}$ is the $(k+h)$-th place $p_{(i,k+h)} \in t_i$ (with $h \geq 1$) in the same trip, such that there are h direct transfers. If the same transitive transfer $A \rightarrow D$ appears more than once in the trip, it appears only once in the group.

Expectations. An itemset describes common transfers performed by a few tourists. The greater the cardinality of an itemset and its support, the higher the relevance of the itemset. The idea is the same as for Interpretation 2, but since now we consider indirect transfers too, hidden connections between places should be revealed in a better way.

By depicting the itemsets on the map, we expect to obtain the following outcomes:

- From the topographic representation, we expect to get a clear reticular view of the space visited by tourists, in such a way each network clearly delimits the lived space;
- From the topological representation, we expect to obtain strong evidence of virtual interconnections among places.

3.5.1. Case Study: Topographic and Topological Representations

We ran the itemset-mining algorithm on the case-study data set with Interpretation 3 by setting the minimum threshold for support as $minsupp = 3/|\mathcal{D}_3|$, to contrast the explosion of large itemsets. We obtained a larger number of itemsets with more items (Table 6 reports the distribution of itemsets) than in the case of Interpretation 2 (see Table 5).

Table 6. Distribution of itemsets for Interpretation 3.

# of Items in the Itemset	Support (# of Tourist)	# of Itemsets
2	3	59
2	4	9
2	5	1
2	6	2
2	7	1
2	8	1
3	3	30
3	4	5
4	3	6
4	4	1
5	3	1

Figures 8 and 9 adopt the topographic representation to depict the discovered itemsets: as for Interpretation 2, items are arrows from one place to another. Figure 10 reports the topological representation for the discovered itemsets: items are still arrows from one place to another.

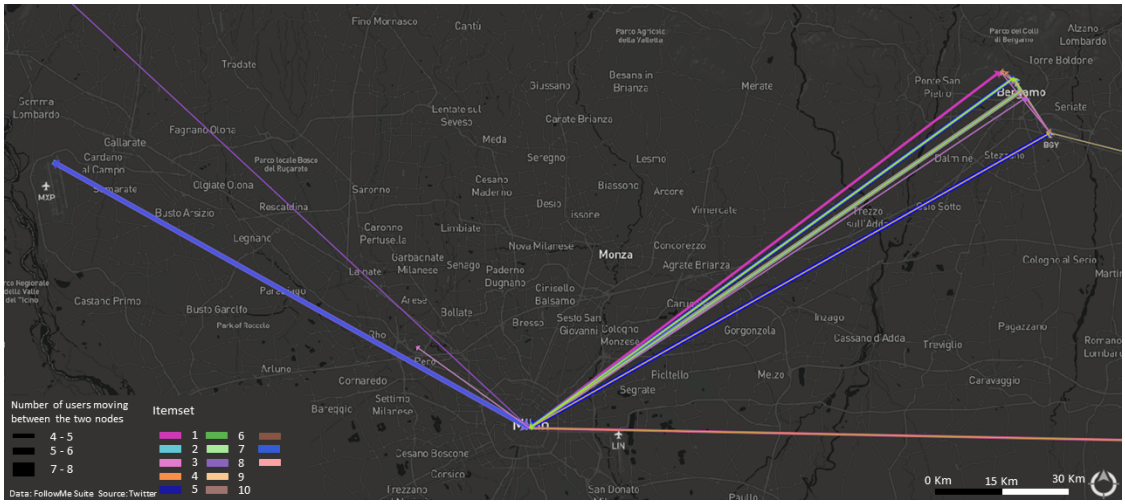

Figure 8. Topographic representation for discovered itemsets associating indirect transfers (Interpretation 3), focused at the level of surrounding regional territory of Bergamo.

Outcomes from the case-study data set. The larger number of itemsets obtained by Interpretation 3 (with respect to Interpretation 2) was due to the fact that groups contained a much larger number of items than in the case of Interpretation 2, i.e., all indirect transfers obtained by composing direct transfers.

Figures 8 and 9 report the topographic representations of the itemsets obtained for Interpretation 3. In particular, Figure 8 is focused at the level of surrounding territory of Bergamo area, while Figure 9 is focused at the level of Bergamo area. Now, the analyst is provided with much richer information, because many tourists made common indirect transfers. In fact, this way we are able to let the following situation be revealed: consider two tourists; the first one made the direct transfers (A, B) and (B, D), while the second tourist made the direct transfers (A, C) and (C, D); with Interpretation 3, the fact that both tourists moved from place A to place D emerges. This is why, in Figure 8, the number of depicted itemsets is much larger than those depicted in Figure 5.

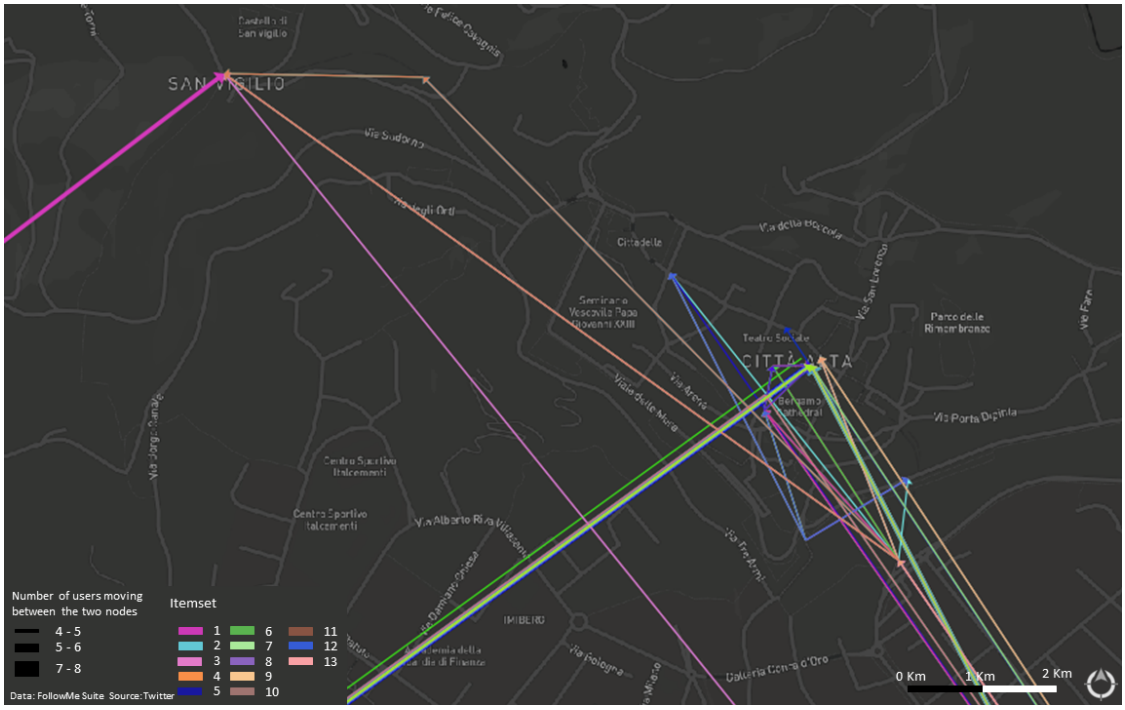

Figure 9. Topographic representation for discovered itemsets associating indirect transfers (Interpretation 3), focused at the level of the Bergamo area.

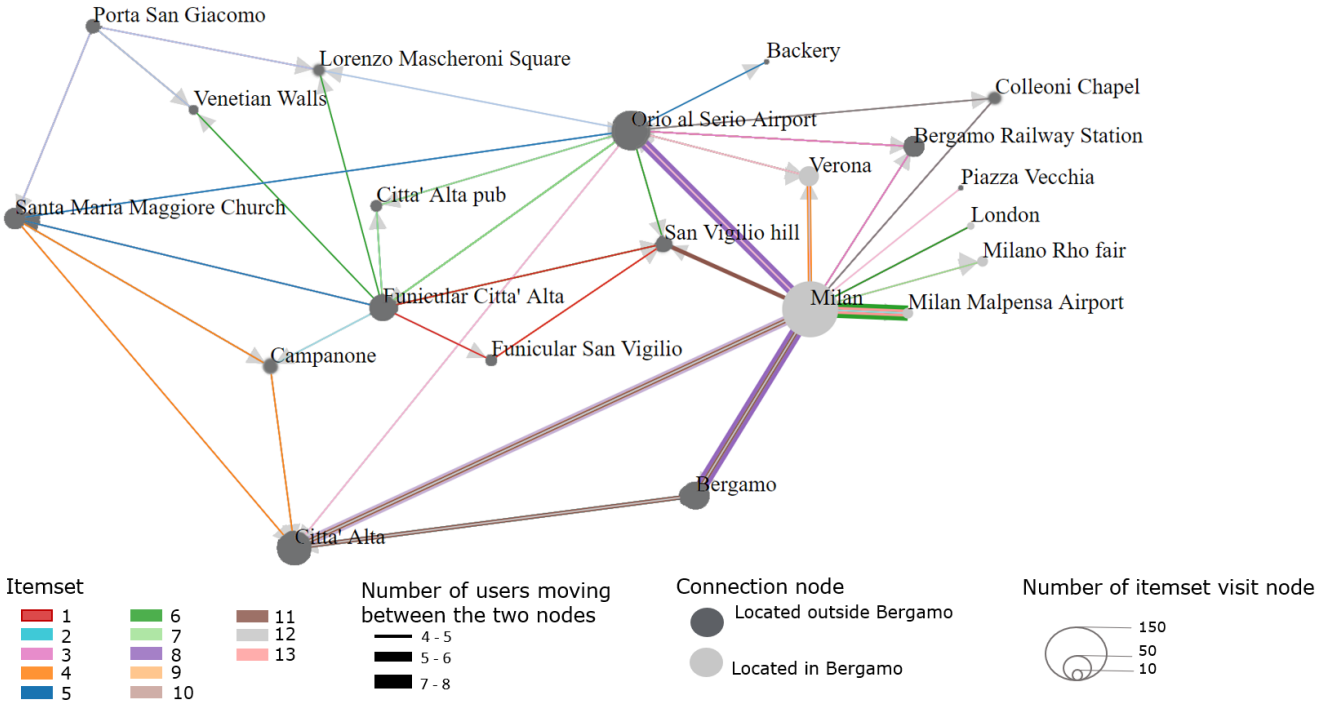

Figure 10. Topological representation for discovered itemsets associating indirect transfers.

While Figure 8 confirms what emerged with Interpretation 2, Figure 9 (focused at the level of local area) is able to show how tourists visited the different places in the city. We do not want to go into a deep analysis of the map, since it is related to the specificities of the territory and is outside the scope of the paper. Nevertheless, it is clear that a network view, based on movements of tourists, emerges and could be the starting point for further

analysis, based either on automatic (such as the *Node Rank* algorithm presented in [49]) or manual and visual analysis tools to develop.

Figure 10 shows the topological representation of the itemsets obtained for Interpretation 3. Notice that now the network structure of interconnections among places is clear. Now it is possible to study (since indirect transfers are represented as arrows) common moving paths.

Furthermore, strongly connected places attract themselves, so that they are shown close each other. Specifically, the reader can see that while the connection between Milan and Bergamo is confirmed, and the connection between Milan and Milan Malpensa Airport (top and left-hand side of the figure), the connection between Milan and Bergamo Orio al Serio Airport clearly emerges, meaning that many people reached Milan from Bergamo (Orio al Serio) Airport, passing through intermediate destinations possibly located in Bergamo.

Finally, notice that many apparently-satellite places emerge: now, the directed arrows clearly show the paths they belong to, i.e., the paths followed by tourists to visit the local area. This kind of information could be very important for further analyses.

3.5.2. Summary of Interpretation 3

We can now summarize what we obtained by experimenting Interpretation 3.

- Deriving all transitive transfers from direct transfers overcomes the limitation of Interpretation 2, giving now a means to study common paths of tourists, in particular with the focus at the level of local area.
- The large number of items that now are in groups has a negative effect; i.e., a very large number of itemsets could be extracted.
- Since itemsets to extract may be too many, making the algorithm unable to compute them, it is necessary to determine a higher and suitable value for minimum support that strongly depends on the data set.
- Interpretation 3 strongly enhances the mobility network of people, as is shown by the topological representation.
- Interpretation 3 is able to deal with the mixed focus during the analysis, i.e., maintaining detailed places in the local area and collapsing places outside: this way, it is possible to reveal connections between single places in the local area and connections between single places in the local area and cities in the surrounding regional area.

4. Sample Application to a Large Data Set

In Section 3, we presented the methodological approach and investigated the potentiality of the concept of rhizome to study mobility networks. In this section, we want to show its application to the large data set from which the case-study data set was extracted.

This data set was built by monitoring users who posted at least one tweet from a pool of 30 airports, chosen because they were directly connected to Bergamo International Airport. Among all traces, we selected those that posted at least one tweet from Bergamo area. These traces are 1129 and the total number of tweets is 47,290 associated with 4765 different locations; the average size of each trace is 42. Geo-coding was performed by means of "Nominatim" (URL: https://wiki.openstreetmap.org/wiki/Nominatim), which relies on *OpenStreetMap*. The geo-coding service provided address, city and country of each position.

After geo-coding, we observed that most of trips had tweets posted from abroad Italy, so we decided to apply Interpretation 2 in order to let international interconnections with Bergamo and surrounding area emerge. For such a kind of analysis, addresses provided by geo-coding tools are not relevant: consequently, we used city names as place labels.

We ran the algorithm for itemset mining with a minimum support threshold $minsupp = 0.005$: this way, we extracted common transfers that appear in at least 5 trips. Table 7 reports the number of itemsets having cardinality greater than 1; the total number of extracted itemsets is 1446, which we thought to be the right size for obtaining significant visualizations.

Table 7. Distribution of itemsets for Interpretation 2 applied to the large data set.

# of Items in the Itemset	# of Itemsets
2	552
3	500
4	270
5	102
6	21
7	1

Figure 11 shows the topological representation we obtained. Notice that it is possible to discover the European cities that are mostly connected with Orio al Serio (the municipality whose territory hosts Bergamo International Airport). In contrast, with the bare geo-coding provided by external tools, without additional knowledge, it is not possible to identify Milan Malpensa Airport, because its area is distributed on the territory of four different municipalities. Clearly, we do not want to discuss about specific connections depicted in the figure; in contrast, we want to point out that we obtained a Europe-wide rhizome, based on traces of Twitter users: even though they represent only partially moving people, they actually are a mine of new data for geographers and spatial analysts, previously not available, that strongly contributes to foster a data-driven approach to geography and spatial analysis.

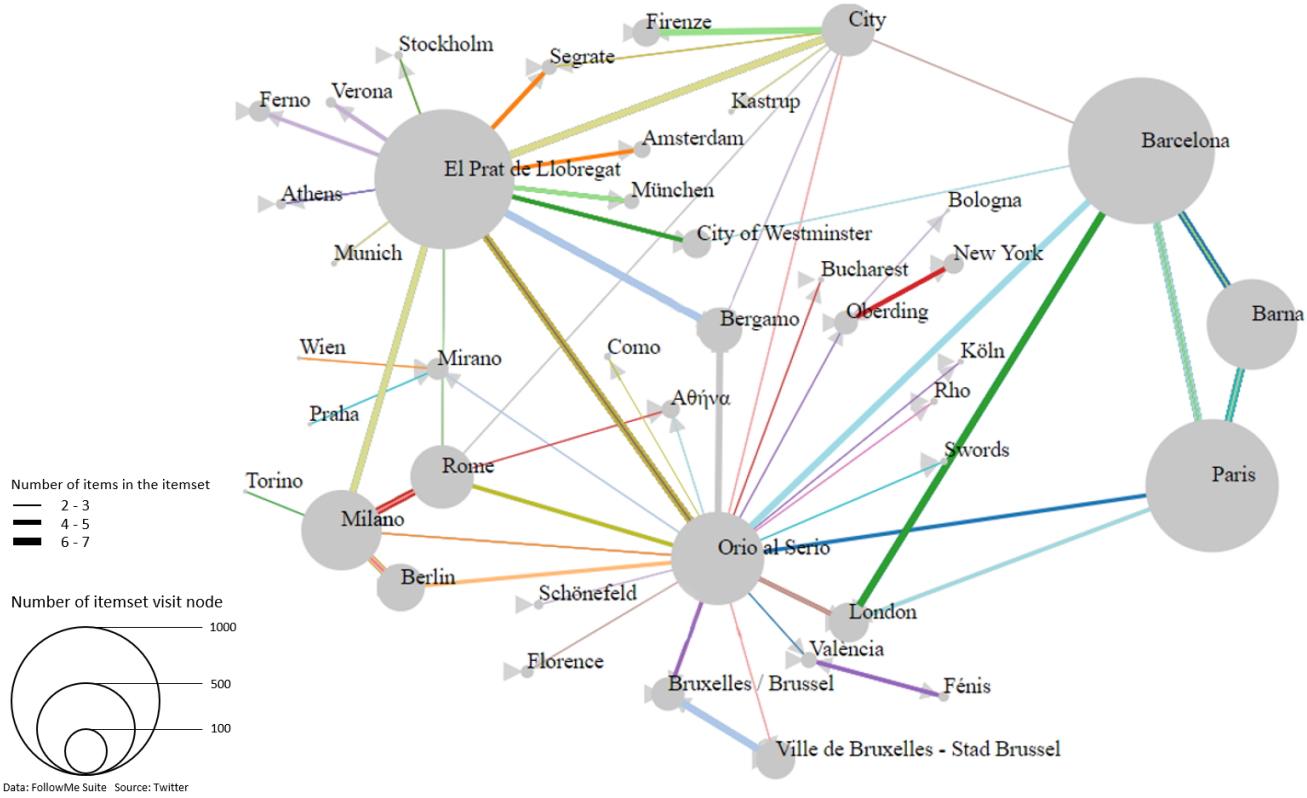

Figure 11. Topological representation for discovered itemsets from the large data set by applying Interpretation 2.

Figure 12 reports the topographic representation of discovered associations of direct transfers, on a zoom level that partly covers the east coast of the USA (on the west side notice "Newark", that hosts a very large international airport), centered on Bergamo area: this way, it is possible to have a clear evidence of provenances/destinations of tourists that visited Bergamo area. However, if we zoom in, we can discover some interesting details: Figure 13 is centered on the area of Amsterdam (The Netherlands) and highlights how people moved from the city of Amsterdam to its airport and vice-versa. This is an

interesting example of what it is possible to obtain by analyzing social media data by adopting the rhizomatic approach.

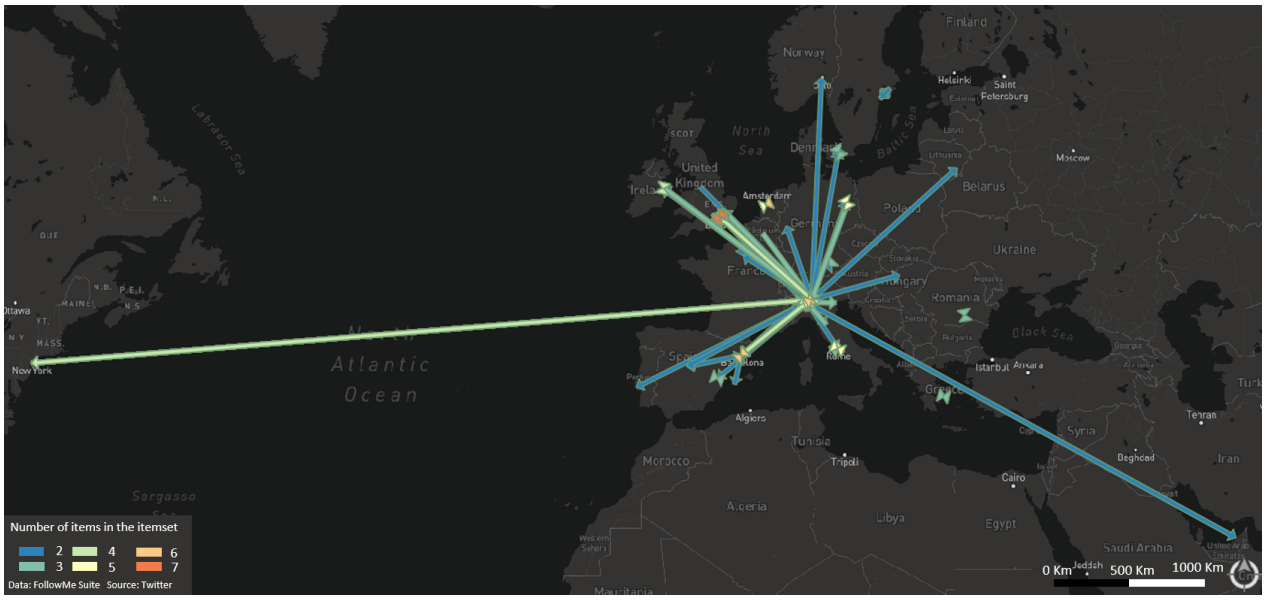

Figure 12. Topographic representation of discovered itemsets from the large data set by adopting Interpretation 2.

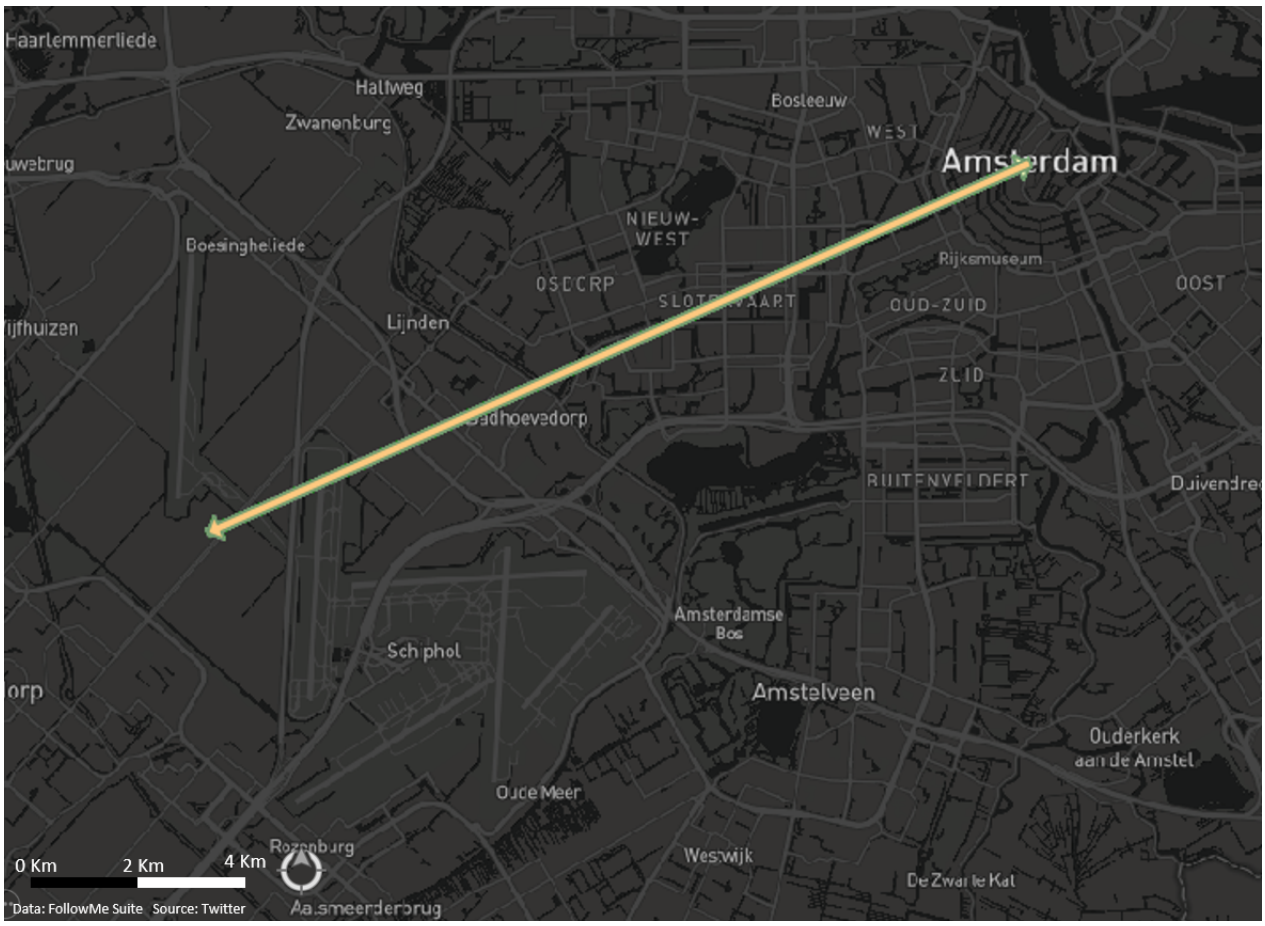

Figure 13. Topographic representation for discovered itemsets from the large data set by applying Interpretation 2, focused on Amsterdam area.

What happens if we apply Interpretation 1 by keeping the same minimum threshold for support, i.e., $minsupp = 0.005$? Table 8 reports the distribution of itemsets on the basis of their cardinality. Notice that they increased significantly, but the largest cardinality is the same, i.e., seven items. The total number of itemsets is now 4240, that is about four times the itemsets obtained for Interpretation 2.

Remember that the representation strategy we adopted for Interpretation 1 depicts a line for each pair of items in an itemset (see Section 3.3.1). However, this strategy, combined with the large number of itemsets we obtained (more than 4000), would produce a useless topological representation, in which edges would form a substantially uniform mixture of colors. For this reason, we decided to consider only itemsets having seven items in Figure 14: notice that the most relevant aspect is the size of a point, which corresponds with the degree of interconnection with other places.

Table 8. Distribution of itemsets for Interpretation 1 applied to the large data set.

# of Items in the Itemset	# of Itemsets
2	1000
3	1494
4	1151
5	492
6	97
7	6

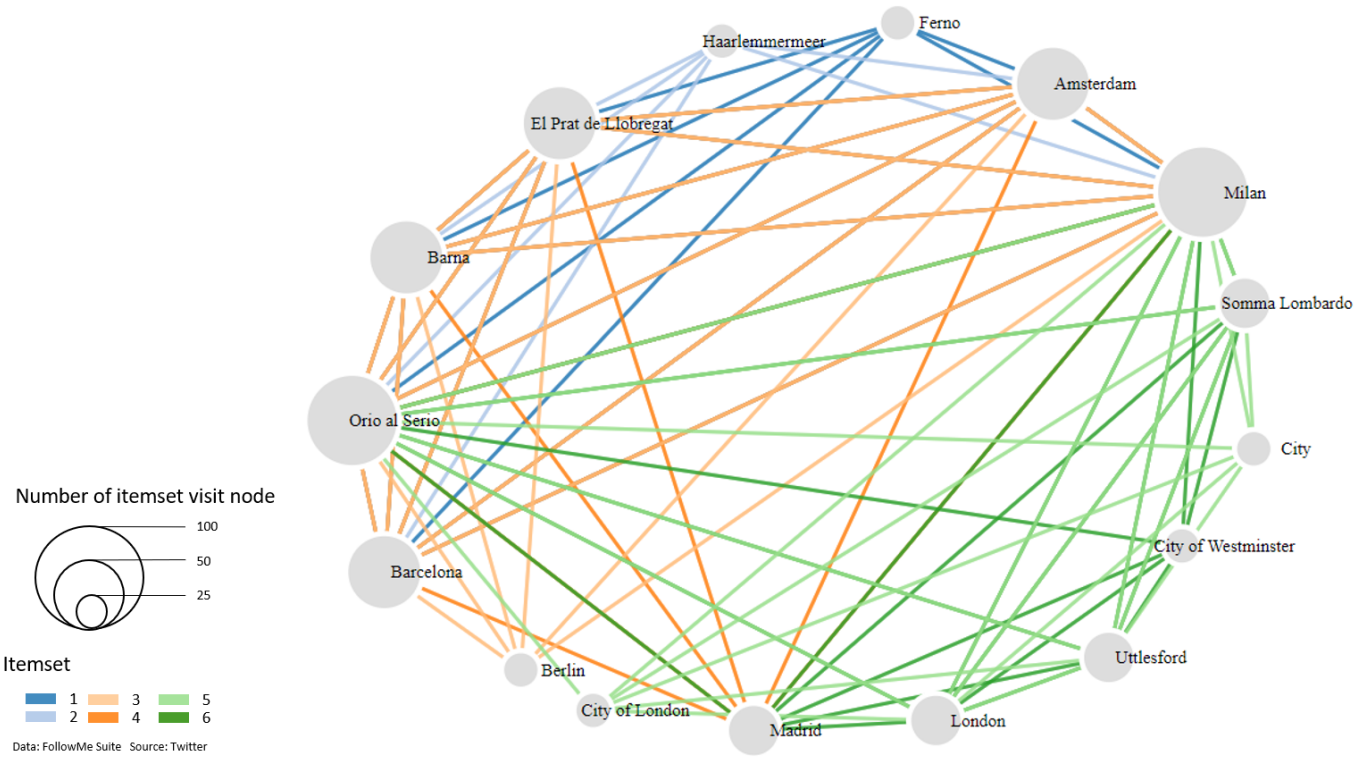

Figure 14. Topological representation for discovered itemsets from the large data set by applying Interpretation 1.

In Figure 15, we depict the topographic representation of the overall set of more than 4000 extracted itemsets, zoomed at a world-wide scale. Notice how the topographic representation helps us understand how areas are virtually connected; in particular, some more connections that were not discovered by Interpretation 2 emerge (notice, in particular, the west coast of the American continent).

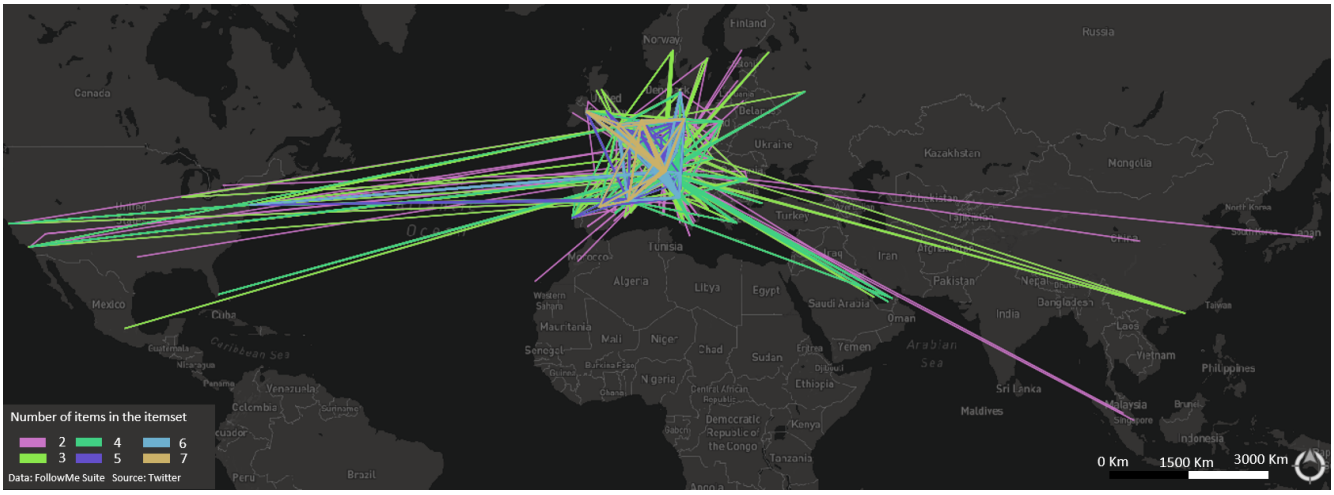

Figure 15. Topographic representation for discovered itemsets from the large data set by applying Interpretation 1.

However, by zooming in on specific places, even this representation can give interesting highlights. Figure 16 is focused on the London (UK) area: the reader can notice the poly-centric structure of connections that concern that area. This is due to the fact that geo-coding tools provide several names as city names and not only "London" (probably, this is due to the administrative structure of municipalities). The poly-centric structure that emerges from the topographic representation clearly reveals this situation, and can provide significant insights about movements of people.

This section has shown the secondary contribution of the paper: rhizomes and itemset mining can be very powerful to study mobility by analyzing traces of people gathered through social media and Twitter in particular. However, several issues concerned with visualization of possibly numerous itemsets must be further investigated, so as to provide effective tools.

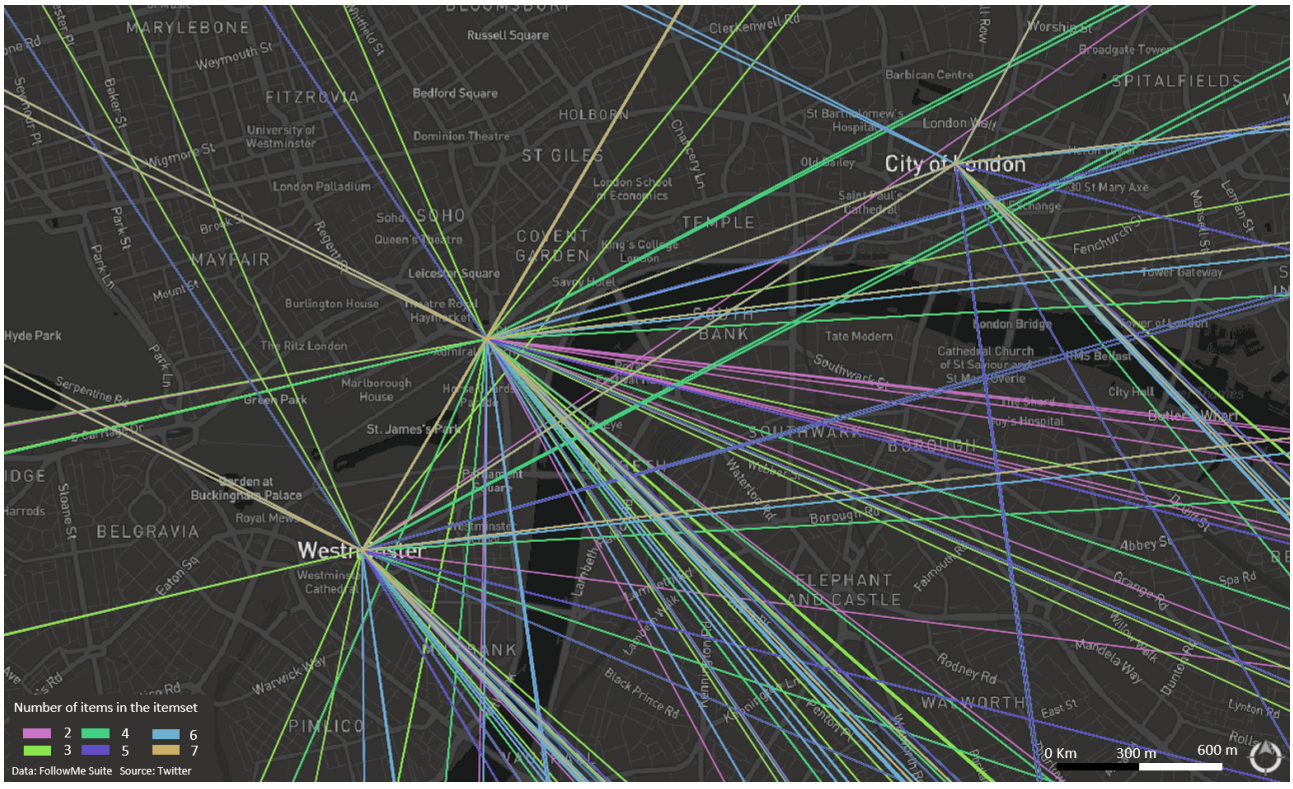

Figure 16. Topographic representation for discovered itemsets from the large data set by applying Interpretation 1, focused on London (UK) area.

5. Conclusions

In this paper, we showed how a classical data mining technique, namely, itemset mining, can be applied to give a data-driven interpretation to the concept of rhizome, defined by geographers.

We built a case-study data set in order to test the approach: traces of geo-located tweets posted by tourists that visited the area of Bergamo (city in the north of Italy) were gathered, filtered and analyzed. The reduced size of the case-study data set, united with our knowledge of the territory, helped us understand the outcomes of the mining technique, so as to prove the concept.

We considered three different interpretations of the problem, corresponding to three different formulations of the mining task. Since the bare application of itemset mining is not sufficient to view the mobility networks, it is essential to visualize itemsets on the map; specifically, two different representations (topographic and topological) are necessary, because they highlight different aspects. Thus, the analysis can be performed only by jointly adopting the proper interpretation and the proper visualization. Nevertheless, they give a concrete and data-driven shape of the concept of rhizome.

Furthermore, in order to show the potential application of rhizomes and itemset mining to study large data sets containing numerous traces of moving people, we presented a sample application to the overall set of traces of potential tourists that posted at least one tweet in Bergamo area, from which we extracted the case-study data set. This sample application highlights the potential usefulness of traces coming from social media (an Twitter in particular) for geographers, enabling novel data-driven investigations.

Hereafter, we summarize the lessons that we have learned.

- The application of itemset mining without a proper support for visualizing discovered itemsets is substantially useless; i.e., only the joint contribution of mining and visualization is effective.
- As argued in the definition of the concept of rhizome (Section 2.1), the topographic representation and the topological representation jointly characterize mobility networks. In particular, while the topographic representation shows the mobility networks as they could be seen and imagined from outside the networks, the topological representation is actually able to show insights of the networks, in some sense giving a look at the intimate structure of the networks.
- Depending on the interpretation we give to the problem, the proposed approach is able to reveal different aspects of mobility networks, suitable for different focuses of the analysis. Interpretation 1 is good for getting preliminary insights about the mobility networks, without considering the time dimension. Interpretation 2 is good for studying mobility networks at the level of interconnected cities. Interpretation 3 is good for studying mobility networks at different focuses at the same time, i.e., local area and surrounding regional area.
- Finally, both the topographic and the topological representations show a "rhizomatic shape". Consequently, we can claim that we have found concrete interpretations for the concept of rhizome that has assumed practical and concrete shapes.

Several open issues to be investigated in the future can be identified (we list the main ones).

- In this work, we focused on mobility networks of groups of people. However, mobile applications provide owners of mobile devices with the possibility to continuously trace their movements (such as Google Timeline). Could the presented approach be used to study personal mobility networks, on the basis of traces voluntarily provided by people?
- Some readers could have concerns about the ethics of using tweets for studying the mobility of users. In this respect, our opinion is that Twitter users should be aware of the fact that Twitter has no privacy mechanism; thus, everything they post can be seen

by everybody. However, often, this is exactly what Twitter users want. We were able to gather traces simply because Twitter APIs allowed us to do that. We did not collect data from other social media, such as Facebook, because their privacy models ask for a direct action by users, who have to explicitly agree to share their data.

- An issue that emerges from Interpretation 1 is the fact that n-itemsets may include k-itemsets, with $2 \leq k < n$, which certainly have support greater than or equal to the n-itemset. Their visualization grows up common edges. Thus, it is legitimate to wonder whether this effect makes connections artificially stronger than they are.
  A possible solution could be to explore the adoption of *closed itemsets* [50]: an itemset y is closed if it is not contained in another itemset z, such that y and z have the same support. Thus, if we extract only closed itemsets, the generated itemsets are significantly less than those obtained by the classical technique. What happens if we visualize only closed itemsets instead of all itemsets? We plan to investigate this issue as future work.
- An important limitation of the visualization techniques we experimented is the fact that they render pair-wise connections. In fact, it is true that different itemsets are depicted with different colors; however, the overlapping effect hides associations of more than two places. This is an important issue to solve, by exploring novel approaches for visualizing itemsets.
  Along the same lines, we can consider the issue emerged by applying Interpretation 1 to the large data set (Section 4): in order to obtain a meaningful topological representation, we had to depict only itemsets containing seven items. As a consequence, we guess that a practical visualization tool should allow analysts to select the cardinality of itemsets to visualize.
- The number of places visited during a trip that we can discover strongly depends on the attitude of the user regarding posting messages in a frequent way. Clearly, it appears from our data set that Twitter users are not so active in posting geo-located messages. What would happen if they were more active?
  It is hard to say. Certainly, their traces would be richer and their movements could be traced more accurately. Clearly, the extraction of itemsets would be strongly affected, resulting in an exponential increase of itemsets, in particular in the case of Interpretation 3, because the number of transitive transfers would dramatically increase. Anyway, the number of extracted itemsets can be taken under control by increasing the minimum threshold for support, and by adopting the technique to extract only closed itemsets.
  However,this is outside the scope of the paper. In fact, its main contribution is to give, for the first time, a data-driven interpretation of the concept of rhizome, provided that a pool pf traces (possibly gathered thorough social media) is available.

We want to remark that the contribution of the paper is neither how to collect data sets with traces of moving people from social media nor how to pre-possess these traces in a possibly automatic way. The main contribution is a first step towards the definition of a practical interpretation of the concept of rhizome (which, until now, was only theoretical) based on itemset mining and visualization techniques, provided that a data set with traces of moving people is available and whose quality is accepted by analysts and geographers. In fact, the outcomes from the case-study data set are not meant to be the contribution of the paper (to be helpful, for instance, for decision makers); they are only a means to understand the potentiality of the approach. Nonetheless, the secondary contribution of the paper is to show how social media (and Twitter in particular) could become precious sources of information for studying mobility of people by applying rhizomes and itemset mining, provided that flexible tools for dealing with the analysis are provided.

Consequently, the future work can continue in two different directions: the first one concerns the tuning of the application of the itemset-mining technique to larger data sets, in particular as far as tuning of the minimum threshold for support is concerned. The second direction concerns the development of new tools to better inspect discovered itemsets

(for example, by providing filtering panels and inspection functionalities), and the development of further algorithms to perform a second-stage analysis on discovered itemsets.

Author Contributions: Conceptualization and methodology, F.B. and G.P.; software, N.C.; writing—original draft preparation, F.B. and G.P.; writing—review and editing, F.B., N.C. and G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mazzocchi, F. Could Big Data be the end of theory in science? *EMBO Rep.* **2015**, *16*, 1250–1255. [CrossRef] [PubMed]
2. Miller, H.J.; Goodchild, M.F. Data-driven geography. *GeoJournal* **2015**, *80*, 449–461. [CrossRef]
3. Deleuze, G.; Guattari, F.; Pérez, J.V.; Larraceleta, U. *Rizoma: (Introducción)*; Pre-Textos: Valencia, Spain, 2003.
4. Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD Rec.* **1993**, *22*, 207–216. [CrossRef]
5. Burini, F.; Cortesi, N.; Gotti, K.; Psaila, G. The Urban Nexus Approach for Analyzing Mobility in the Smart City: Towards the Identification of City Users Networking. *Mob. Inf. Syst.* **2018**, *2018*, 17. [CrossRef]
6. Latour, B. On Actor-Network Theory: A few clarifications. *Soz. Welt* **1996**, *47*, 369–381.
7. Latour, B. On recalling ANT. *Sociol. Rev.* **1999**, *47*, 15–25. [CrossRef]
8. Bosco, F.J. Actor-Network Theory, networks, and relational approaches in human geography. In *Approaches to Human Geography*; SAGE: London, UK, 2006; pp. 136–146.
9. Sheppard, E. The spaces and times of globalization: Place, scale, networks, and positionality. *Econ. Geogr.* **2002**, *78*, 307–330. [CrossRef]
10. Hinchliffe, S. Specifying powers and their spatialities. *Entanglements Power Geogr. Domin.* **2000**, *5*, 219.
11. Paddison, R.; Philo, C.; Routledge, P.; Sharp, J. *Entanglements of Power: Geographies of Domination/Resistance*; Routledge: Abingdon-on-Thames, UK, 2002.
12. Lussault, M.; Lévy, J. *Dictionnaire de la géographie et de l'espace des sociétés*; Éditions Belin: Paris, France, 2000.
13. Lévy, J. *L'invention du Monde*; Presses de Sciences Po: Paris, France, 2008.
14. Lévy, J.; Romany, T.P.L.; Maitre, O.P. Rebattre les cartes. Topographie et topologie dans la cartographie contemporaine. *Réseaux* **2016**, *34*, 17–52. [CrossRef]
15. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference very Large Data Bases, VLDB, Santiago de Chile, Chile, 12–15 September 1994; Volume 1215, pp. 487–499.
16. Wang, K.; Tang, L.; Han, J.; Liu, J. Top down fp-growth for association rule mining. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taipei, Taiwan, 6–8 May 2002; pp. 334–340.
17. Fosci, P.; Psaila, G.; Di Stefano, M. The hints from the crowd project. In Proceedings of the International Conference on Database and Expert Systems Applications, Prague, Czech Republic, 26–29 August 2013; pp. 443–453.
18. Meo, R.; Psaila, G.; Ceri, S. A new SQL-like operator for mining association rules. In Proceedings of the VLDB, Mumbai, India, 3–6 September 1996; Volume 96, pp. 122–133.
19. Meo, R.; Psaila, G. An XML-based database for knowledge discovery. In Proceedings of the International Conference on Extending Database Technology, Munich, Germany, 26–31 March 2006; pp. 814–828.
20. Nardi, B.A.; Schiano, D.J.; Gumbrecht, M. Blogging as social activity, or, would you let 900 million people read your diary? In Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, Chicago, IL, USA, 6–10 November 2004; pp. 222–231.
21. Steiger, E.; Westerholt, R.; Resch, B.; Zipf, A. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Comput. Environ. Urban Syst.* **2015**, *54*, 255–265. [CrossRef]
22. Paraskevopoulos, P.; Palpanas, T. What do Geotagged Tweets Reveal About Mobility Behavior? In Proceedings of the International Workshop on Mobility Analytics for Spatio-Temporal and Social Data, Munich, Germany, 1 September 2017; pp. 36–53.
23. Abbasi, A.; Rashidi, T.H.; Maghrebi, M.; Waller, S.T. Utilising location based social media in travel survey methods: bringing twitter data into the play. In Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Bellevue, WA, USA, 1 November 2015; p. 1.
24. Lenormand, M.; Gonçalves, B.; Tugores, A.; Ramasco, J.J. Human diffusion and city influence. *J. R. Soc. Interface* **2015**, *12*, 20150473. [CrossRef] [PubMed]
25. Girardin, F.; Calabrese, F.; Dal Fiore, F.; Ratti, C.; Blat, J. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Comput.* **2008**, *7*, 36–43. [CrossRef]

26. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [CrossRef] [PubMed]
27. Hübl, F.; Cvetojevic, S.; Hochmair, H.; Paulus, G. Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 302. [CrossRef]
28. Bae, S.J.; Lee, H.; Suh, E.K.; Suh, K.S. Shared experience in pretrip and experience sharing in posttrip: A survey of Airbnb users. *Inf. Manag.* **2017**, *54*, 714–727. [CrossRef]
29. Brandt, T.; Bendler, J.; Neumann, D. Social media analytics and value creation in urban smart tourism ecosystems. *Inf. Manag.* **2017**, *54*, 703–713. [CrossRef]
30. Del Vecchio, P.; Mele, G.; Ndou, V.; Secundo, G. Creating value from social big data: Implications for smart tourism destinations. *Inf. Process. Manag.* **2018**, *54*, 847–860. [CrossRef]
31. Huang, C.D.; Goo, J.; Nam, K.; Yoo, C.W. Smart tourism technologies in travel planning: The role of exploration and exploitation. *Inf. Manag.* **2017**, *54*, 757–770. [CrossRef]
32. Kim, S.E.; Lee, K.Y.; Shin, S.I.; Yang, S.B. Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo. *Inf. Manag.* **2017**, *54*, 687–702. [CrossRef]
33. Azmandian, M.; Singh, K.; Gelsey, B.; Chang, Y.H.; Maheswaran, R. Following human mobility using tweets. In Proceedings of the International Workshop on Agents and Data Mining Interaction, Valencia, Spain, 4–8 June 2012; pp. 139–149.
34. Steiger, E.; Resch, B.; de Albuquerque, J.P.; Zipf, A. Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps. *Transp. Res. Part C Emerg. Technol.* **2016**, *73*, 91–104. [CrossRef]
35. Valle, D.; Cvetojevic, S.; Robertson, E.P.; Reichert, B.E.; Hochmair, H.H.; Fletcher, R.J. Individual movement strategies revealed through novel clustering of emergent movement patterns. *Sci. Rep.* **2017**, *7*, 44052. [CrossRef] [PubMed]
36. Wakamiya, S.; Lee, R.; Sumiya, K. Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from twitter. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, IL, USA, 1–4 November 2011; pp. 77–84.
37. Bordogna, G.; Frigerio, L.; Cuzzocrea, A.; Psaila, G. An effective and efficient similarity-matrix-based algorithm for clustering big mobile social data. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 514–521.
38. Bordogna, G.; Frigerio, L.; Cuzzocrea, A.; Psaila, G. Clustering geo-tagged tweets for advanced big data analytics. In Proceedings of the 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 27 June–2 July 2016; pp. 42–51.
39. Bordogna, G.; Cuzzocrea, A.; Frigerio, L.; Psaila, G.; Toccu, M. An interoperable open data framework for discovering popular tours based on geo-tagged tweets. *Int. J. Intell. Inf. Database Syst.* **2017**, *10*, 246–268. [CrossRef]
40. Aboulnaga, Y.; Clarke, C.L. *Frequent Itemset Mining for Query Expansion in Microblog Ad-Hoc Search*; Technical Report; Waterloo University: Waterloo, ON, Canada, 2012.
41. Lin, L.; Li, J.; Zhang, R.; Yu, W.; Sun, C. Opinion mining and sentiment analysis in social networks: A retweeting structure-aware approach. In Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, UK, 8–14 December 2014; pp. 890–895.
42. Weiler, M.; Schmid, K.A.; Mamoulis, N.; Renz, M. Geo-social co-location mining. In Proceedings of the Second International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data, Melbourne, VIC, Australia, 31 May 2015; pp. 19–24.
43. Cagliero, L.; Cerquitelli, T.; Garza, P.; Grimaudo, L. Twitter data analysis by means of strong flipping generalized itemsets. *J. Syst. Softw.* **2014**, *94*, 16–29. [CrossRef]
44. Faralli, S.; Di Tommaso, G.; Velardi, P. Semantic enabled recommender system for micro-blog users. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 992–998.
45. Cuzzocrea, A.; Psaila, G.; Toccu, M. Knowledge Discovery from Geo-Located Tweets for Supporting Advanced Big Data Analytics: A Real-Life Experience. In Proceedings of the 5th International Conference on Model and Data Engineering, Rhodes, Greece, 26–28 September 2015; Volume 9344, pp. 285–294.
46. Cuzzocrea, A.; Psaila, G.; Toccu, M. An innovative framework for effectively and efficiently supporting big data analytics over geo-located mobile social media. In Proceedings of the 20th International Database Engineering & Applications Symposium, Montreal, QC, Canada, 11–13 July 2016; pp. 62–69.
47. Bordogna, G.; Capelli, S.; Psaila, G. A big geo data query framework to correlate open data with social network geotagged posts. In Proceedings of the The Annual International Conference on Geographic Information Science, Boston, MA, USA, 7–10 November 2017; pp. 185–203.
48. Bordogna, G.; Capelli, S.; Ciriello, D.E.; Psaila, G. A cross-analysis framework for multi-source volunteered, crowdsourced, and authoritative geographic information: The case study of volunteered personal traces analysis against transport network data. *Geo-Spat. Inf. Sci.* **2018**, *21*, 257–271. [CrossRef]
49. Cortesi, N.; Gotti, K.; Psaila, G.; Burini, F.; Lwin, K.T.; Hossain, M. A network-based ranking approach to discover places visited by tourists from geo-located tweets. In Proceedings of the 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Malabe, Sri Lanka, 6–8 December 2017; pp. 1–8.
50. Pasquier, N.; Bastide, Y.; Taouil, R.; Lakhal, L. Discovering frequent closed itemsets for association rules. In Proceedings of the International Conference on Database Theory, Jerusalem, Israel, 10–12 January 1999; pp. 398–416.