

Geosemantic Network-of-Interest Construction Using Social Media Data

Sophia Karagiorgou¹, Dieter Pfoser², and Dimitrios Skoutas³

¹ School of Rural and Surveying Engineering
National Technical University of Athens

`sokaragi@mail.ntua.gr`

² Department of Geography and GeoInformation Science
George Mason University

`dpfoser@gmu.edu`

³ Institute for the Management of Information Systems
R.C. ATHENA

`dskoutas@imis.athena-innovation.gr`

Abstract. Within the last years, an ever increasing amount of data from mobile and navigation devices (e.g. web check-in, vehicle tracking data, etc), as well as social media (e.g. Twitter) are becoming available, presenting and enabling new research challenges and applications. To unveil persistent and meaningful knowledge from user-generated location-based “stories”, this work proposes a novel methodology that converts Twitter check-in data into a mixed geo-semantic network-of-interest (NOI). It does so by introducing a novel network construction algorithm on segmented input data based on different mobility types. This produces network layers by means of behavioral and geometric trajectories, which are then combined into a single network. This segmentation addresses also the challenges imposed by noisy, low-sampling rate trajectories. An experimental evaluation assesses the quality of the algorithms by constructing the a network from trajectories based on Twitter check-in data for London and New York. Our results show that this method is robust and provides accurate and interesting results.

1 Introduction

An important resource in today’s mapping efforts, especially for use in mobile navigation devices, is an accurate collection of point-of-interest (POI) data. However, by only considering isolated locations in current datasets, the essential aspect of how these POIs are connected is overlooked. The objective of this work is to take the concept of POIs to the next level by computing *Networks of Interest* (NOIs) that encode different types of connectivity between POIs and capture human behavior while visiting these POIs. This new concept of NOIs has a wide array of application potential, including traffic planning, geomarketing, urban planning, and the creation of sophisticated location-based services, including personalized travel guides and recommendation systems. Currently, the only datasets that consider connectivity of locations are road networks, which

connect intersection nodes by means of road links. This is on a purely geometric basis. POIs, however, encode both geometric and semantic information, and it is not obvious how to create meaningful links and networks between them. We propose to capture, both, *geometric* and *semantic* information in one NOI by analyzing social media in the form of spatial check-in data. Existing road maps and POIs encode mostly geometric information and consist of street maps, but may also include subway maps, bus maps, and hiking trail maps. To complement this dataset, *geometric trajectories* consist of geo-referenced trajectory data, such as GPS tracking data obtained from people moving on a road network. This type of data is assumed to have a relatively high sampling rate. Typical examples include vehicle tracking data sampled every 10 or 30 seconds. Such datasets are constructed using *map construction* (cf. [1], [2] for surveys).

In this work, we will use *behavioral trajectories* as a datasource. They are obtained from social media in the form of spatial check-in data, such as geospatial tweets of users. Similar to GPS tracking, the user contributes a *position sample* by checking in at a specific location. Compared to geometric trajectories, such check-in data result in very low-sampling rate trajectories that when collected for many users provide for a less dense, but semantically richer “movement network” layer. The main challenge arises from the fact that trajectories composed from Twitter check-in data differ technically and semantically from raw GPS-based type of trajectories. Unlike trajectories obtained from GPS devices in typical tracking applications, check-in data are typically quite sparse since individuals tend to publish their positions only at specific occasions. However, we advocate that by combining and analyzing time and location of check-in data, it is possible to construct event-based trajectories, which can then be used to analyze user mobility and to extract visiting patterns of places. The expectation towards behavioral trajectories is that by integrating them into a Network of Interest, the resulting dataset will go beyond a homogeneous transportation network and will provide us with a means *to construct an actual depiction of human interest and motion dependent on user context and independent of transportation means*. As early maps were traces of people’s movements in the world, i.e., view representations of people’s experiences, NOIs try to fuse different qualities of such trace datasets obtained through intentional (e.g., social media, Web logs) or unintentional efforts (e.g., routes from their daily commutes, check-in data) to provide for a *consequent modern map equivalent*.

Specifically, in this paper we address the challenge of extracting a geo-semantic NOI from noisy, low-sampled Twitter check-in data. To do so, we introduce a new NOI construction algorithm that segments the input dataset based on sampling rate and movement characteristics and then infers the respective network layers. To fuse the semantic and geometric network layer into a NOI, we introduce a semantics-based algorithm that takes position samples (check-ins) to create network hubs. A detailed experimental evaluation uses two real-world datasets of check-in data and discussed the NOI construction results in terms of quality and significance.

The remainder of this paper is organized as follows. Section 2 reviews related work on spatiotemporal inference techniques. Sections 3 and 4 present our algorithms for trajectories segmentation and re-association to build the NOI in a layered fashion. In Section 5, we evaluate the quality of the NOI construction method. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

Various approaches have been proposed for using user-generated geospatial content to extract useful knowledge, such as identifying travel sequences, interesting routes or socio-economic patterns. In the following, we present a review of the literature using a categorization of the approaches according to the type of problem solved.

Several methods focus on *sub-sequence extraction (routes) from moving objects trajectories* by mining spatiotemporal movement patterns in tracking data. Kisilevich et al. [3] present an automatic approach for mining semantically annotated travel sequences using geo-tagged photos by searching for sequence patterns of any length. In [4], Chen et al. extract important routes between two locations by observing the traveling behaviors of many users. Although, they mine a transfer network of important routes, they accept that the distance between any two consecutive points in a trajectory does not exceed $100m$, which becomes unrealistic. Zheng et al. [5] use online photos from Flickr and Panoramio to analyze people’s travel patterns at a tour destination. They extract important routes, but no transportation network. Asakura et al. [6] investigate the topological characteristics of travel data, but they focus on identifying a simple index of clustering tourist’s behavior. Mckercher and Lau [7] identify movement patterns and styles of tourists within an urban destination. Our approach analyzes, both, traffic patterns and topological characteristics of travel routes, while most existing work focuses on traffic patterns only. Choudhury et. al [8] explore the construction of travel itineraries from geo-tagged photos. In contrast, in our approach the itinerary is defined to be the spatiotemporal movement trajectory of much finer granularity.

There also exist various methods based on *trajectory clustering*. The majority of the proposed algorithms such as k -means [9], BIRCH [10] and DBSCAN [11] work strictly with point data and do not take the temporal aspect into consideration. Several approaches match some sequences by allowing some elements to be unmatched as in the Longest Common Sub Sequence (LCSS) similarity measure [12]. However, our goal in this work is rather to apply a trajectory clustering approach and also take into consideration the temporal aspect of the data. Similarity measures for trajectories that take the time and derived parameters, such as speed and direction, into account have been proposed in [13]. This approach is close to ours w.r.t. the examined aspects of temporal dimension, however, our method applies clustering techniques in order to infer the connectivity of a NOI. In a previous work [14], we derived a connected road network embedded in

vehicle trajectories, while in [15] we inferred a hierarchical road network based on different movement types. The current approach differs in that it deals with uncertain social media check-in data by taking into account both the spatial and the temporal dimension of the data. The objective is to derive a NOI.

Characterized by its spatial and temporal dimension, Twitter check-in data can be regarded as one kind of spatiotemporal data, which also connects this study to the knowledge extraction-based techniques of the spatiotemporal data mining domain. Crandall et. al [16] investigate ways to organize a large collection (~ 35 million) of geo-tagged photos and determine important locations of photos, such as cities, landmarks or sites, from visual, textual and temporal features. Kalogerakis et. al [17] estimate the geo-locations of a sequence of photos. Similarly, Rattenbury et. al [18] and Yanai et. al [19] analyzed the spatiotemporal distribution of photo tags to reveal the inter-relation between word concepts (photo tags), geographical locations and events. Girardin et al. [20] extract the presence and movements of tourists from cell phone network data and the geo-referenced photos they generate. Similarly, [21] proposes a clustering algorithm of places and events using collections of geo-tagged photos. These approaches efficiently deliver focal spatial data extractions from diverse data sources, while the aim of this work is to also extract *how this data is connected (links)*. In [22], Kling studies urban dynamics based on user generated data from Twitter and Foursquare using a probabilistic model. However, these dynamics have not been translated to a (transportation) graph structure. All these works target the extraction of some kind of knowledge and patterns from photos or geo-referenced sources with textual and spatiotemporal metadata, while we focus on mining transportation and mobility patterns from check-in data such as tweets.

Overall, what sets this work aside is that *social media data is used as a tracking data source*. We use it not to only extract features or knowledge patterns of human activities, but a complete Network of Interest.

3 NOI Layer Construction

As explained in Section 1, our goal is to extract a Network of Interest that captures interesting information about user movement behaviors based on user check-in data from social media (in particular, Twitter). User check-in data are tuples of the form $U = \langle u, x, y, t \rangle$, denoting that the user u was at location (x, y) at time t . These data are organized into trajectories, which represent the sequence of locations a user has visited. Typically, multiple trajectories are produced for each user by splitting the whole sequence of check-ins, e.g., on a daily basis. Hence, each resulting trajectory is an ordered list of spatiotemporal points $T = \{p_0, \dots, p_n\}$ with $p_i = \langle x_i, y_i, t_i \rangle$ and $x_i, y_i \in R, t_i \in R^+$ for $i = 0, 1, \dots, n$ and $t_0 < t_1 < t_2 < \dots < t_n$.

The goal is to construct a NOI that reveals the *movement behavior* of users. This Network of Interest is a directed graph $G = (V, E)$, where the vertices V indicate important locations and the edges E important links between them, according to the observed user movements. In particular, we are interested in

two aspects of the Network of Interest. The *geometric NOI aspect* provides a representation of how users actually move across various locations, thus preserving the actual geometry of the movement. The *semantic NOI aspect* represents the qualitative aspect of the network by identifying significant locations and links between them. In our approach, we treat these two aspects as different layers of the same Network of Interest. In the following, we describe the steps for construing these layers and fusing them to produce the final Network of Interest.

3.1 Segmentation of Trajectories

Behavioral trajectories, as in our case the Twitter check-in data, contain data to construct both the geometric and the semantic layer of a Network of Interest. Conceptually, users tweet when they stroll around in the city as well as when they commute in the morning. While all these tweets will result in behavioral trajectories, *some of them depict actual movement paths*, while others simply are tweets sent throughout the day. In what follows, we try to separate our input data into two subsets and to extract the trajectories corresponding to the respective layer.

A main challenge when inferring a movement network from check-in data is that these data are very heterogeneous in terms of their sampling rate, often being very sparse. However, even the sparse subsets of the data are helpful in identifying significant locations, whereas the denser subsets can be used to capture more fine grained patterns of user movement. For this purpose, we analyze the trajectories and group them into subsets with different temporal characteristics. In our approach, we treat these two aspects by applying (i) *mean speed* threshold to capture the user movement under an urban transportation mode and by applying (ii) a *sampling rate* threshold to identify “abstract” and “concrete” movement. This allows us to treat each subset separately later on in the network construction phase. The “abstract” type of movement corresponds to the *semantic NOI aspect* and the “concrete” corresponds to the *geometric NOI aspect*. Users with frequent check-ins, i.e., a high sampling rate, provide us with the means to derive a geometric NOI layer, while low sampling rates only allow us to reason about abstract movement, i.e, derive a semantic NOI layer.

Notice that typically the same individual, within one daily trajectory may have moved with different sampling rates. In this case, the trajectory needs to be segmented according to the frequency of user check-ins. A naïve process for achieving this separation is the following. First, a duration and a speed (length divided by duration) is recorded for each segment of a trajectory. Each segment is assigned a corresponding duration type of movement. Focusing on urban transportation, we use a mean speed to filter out trajectories and then the duration between samples to determine “abstract” and “concrete” movement. Figure 1 shows the trajectories classified to different sampling rates using the example of Twitter data for London. Using a heatmap coloring schema, concrete and abstract movement are shown in blue and red, respectively.

The process is outlined in Algorithm 1. For each line segment L_j of each trajectory T , we compute a duration and mean speed value (Algorithm 1, Lines

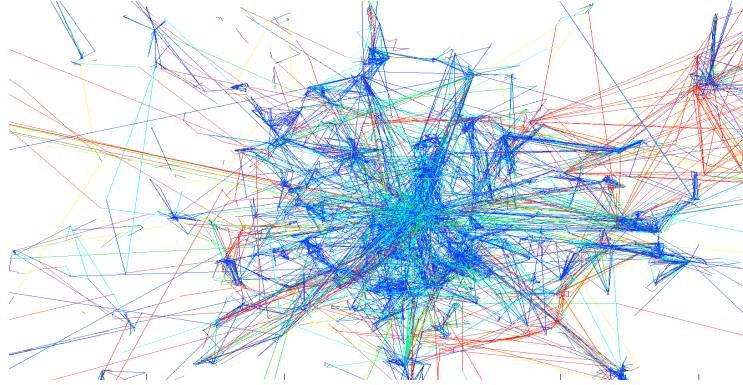


Fig. 1: Segmentation of Trajectories (“slow”: blue, “fast”: red)

6-7), and the segment is then assigned to the corresponding segmented set of trajectories T_G , T_S according to the min and max time interval (Lines 9-13). The algorithm produces segmented sets of trajectories (Lines 10 and 13) based on the corresponding time interval attributes.

Algorithm 1: Segmentation of Trajectories

Input: A set of trajectories T
Output: Two sets of segmented trajectories T_G , T_S

```

1 begin
2   /*Trajectories segmentation according to time intervals*/
3    $V_{max} \triangleright$  maximum mean speed
4   foreach ( $T_i \in T$ ) do
5     foreach ( $L_j \in T_i$ ) do
6        $\bar{t}(L_j) \leftarrow \delta t(P[i-1], P[i]) \triangleright$  Time interval
7        $\bar{v}(L_j) \leftarrow \frac{\delta x(P[i-1], P[i])}{\delta t(P[i-1], P[i])} \triangleright$  Mean speed
8       if  $\bar{v}(L_j) \leq V_{max}$  then
9         if  $\bar{t}(L_j) \leq T_{min}$  then
10          |  $T_G \leftarrow L_j$ 
11          end
12        else if  $\bar{t}(L_j) \geq T_{min}$  and  $\bar{t}(L_j) \leq T_{max}$  then
13          |  $T_S \leftarrow L_j$ 
14          end
15        end
16      end
17    end
18 end

```

3.2 Geometric Layer Construction

To construct the geometric NOI layer we use frequently sampled trajectories. The sampling rate threshold was established through experimentation. In the examples of Section 5, the sampling rate threshold was set to *5mins*. I.e., for the construction of the geometric layer the duration in between position samples of trajectory dataset is less than *5mins* (cf. Table 1).

The geometric NOI layer construction approach follows a modified map construction approach (e.g., [14,15]) by (i) initially clustering position samples to derive network nodes, (ii) linking nodes by using the trajectory data and (iii) refining the link geometry.

To derive network nodes we employ the DBSCAN clustering algorithm [11] a distance threshold and a minimum number of samples threshold parameter. We revisit the segmented trajectories to identify how the network nodes are connected by creating links. The links represent clustered trajectories as two nodes can be connected by different trajectories. For each link (i) a *weight* is derived representing the number of the trajectories comprising the link and also (ii) a *length* representing the Euclidean distance between the nodes that constitute the link. In addition to this, we apply a reduction step to simplify the constructed network. The intuition is that due to varying sampling rates, links between nodes might exhibit redundancy. This reduction step eliminates redundant links but substituting longer links with links of more detailed geometries. We reconstruct links of longer duration by using links of shorter duration if their geometries are similar. We achieve this by using the degree of constructed nodes. Starting with nodes of a higher degree of incoming links, i.e., significant nodes, for such a node, we sort all incident links based on descending duration order. We then reconstruct those, which temporally and spatially cover other links that can be reached in less time. Figure 2a gives an example by showing in dark gray links before reduction and in light gray a portion of the OSM transportation network, while Figure 2b shows in dark gray the reduced links. Part of the larger geometry has been substituted with a more detailed geometry.

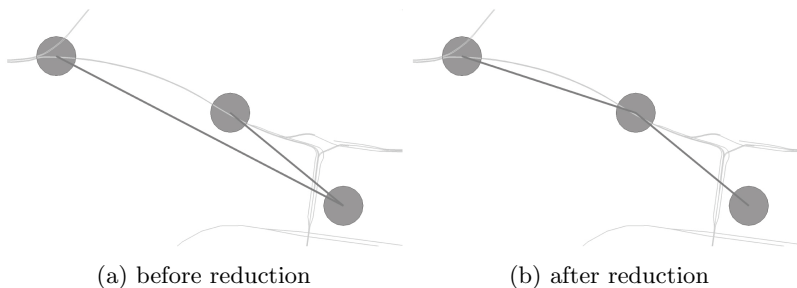


Fig. 2: Network reduction example (constructed network is shown in dark gray and the road network in light gray)

3.3 Semantic Layer Construction

To construct the Semantic NOI layer, we rely on trajectories exhibiting low sampling rates, i.e., potentially cover large distances in between position samples making it difficult to reconstruct the actual movement. By initially applying the DBSCAN clustering algorithm with a distance threshold and a minimum number of samples (see Table 1 for parameter details) we extract a set of nodes that

correspond to the hubs of the semantic layer. Performing a linear scan of the trajectories reveals the respective portions that connect the sets of nodes. For each link sample (i) a *weight* is derived representing the number of the trajectories comprising a link. At this step, we do not apply any reduction method as the geometries of the semantic layer are less accurate. Overall, this layer allows us to extract a network with less spatial accuracy but of greater semantic value.

4 NOI Construction and Layer Fusion

The final part of the NOI construction process consists (i) the extraction of hubs, i.e., significant locations that user frequently visits, and (ii) the fusion of the layers, i.e., the geometric and the semantic layer to produce the integrated network.

4.1 Network Hubs

Next, we describe how we identify hubs, which are POIs that users frequently depart from and arrive at. In particular, specific indicators for hubs are (i) number of constituting position samples (ii) stemming from many different users (iii) over extended periods of time. To extract hubs, we consider the entire trajectory dataset for this process, as hubs are meant to align the geometric and semantic NOI layers.

The Network Hubs Inference algorithm takes as input the combined trajectories used in geometric and semantic layer construction (Algorithm 2, Line 9) and determines the k -NNs of each position sample (Line 12), which are subsequently filtered according to the number of users and the period of time covered (Lines 13-15). On these filtered position samples, we apply the DBSCAN clustering algorithm using a distance threshold and a minimum number of samples (Line 16). The centroids of the resulting clusters are the candidate hubs (Line 17). A final filtering step is applied as follows. For each candidate hub, we also record two properties. A *weight* for the hub is derived as the total number of nodes the hub was derived from, i.e., the size of the corresponding cluster. In addition, we record the *degree* of each hub, i.e., the number of incoming and outgoing edges in the cluster. A candidate hub is included in the output if both the following two conditions hold: (a) both the in-degree and out-degree are above a specified threshold and (b) the in-degree and out-degree do not differ significantly (threshold determined by experimentation). These conditions are used to ensure that the identified hubs correspond to places where a sufficiently large number of users frequently depart from and arrive at (Lines 23-24).

Algorithm 2: Hub Inference	Algorithm 3: NOI Fusion
<p>Input: A set of segmented trajectories T_G, T_S</p> <p>Output: Network Hubs</p> <pre> 1 begin 2 /*Clustering position samples of segmented trajectories to compute network hubs*/ 3 $H^* \leftarrow \emptyset \triangleright$ Candidate Hubs 4 $H \leftarrow \emptyset \triangleright$ Hubs 5 $d_{max} \triangleright$ proximity threshold 6 $u_{min} \triangleright$ min. number of users 7 $h_{min} \triangleright$ min. number of time periods 8 $deg_{in}, deg_{out}, deg_{min}, \epsilon$ 9 \triangleright position samples from combined trajectories 10 $P \leftarrow \text{UNION}(T_G, T_S)$ 11 \triangleright Samples \rightarrow Hubs 12 foreach ($P[i]$) do 13 $\nu_i \leftarrow \text{FINDNN}(P[i], d_{max})$ 14 $u_p \leftarrow \text{COUNTUSERS}(\nu_i)$ 15 $h_p \leftarrow \text{COUNTHOURS}(\nu_i)$ 16 if ($u_p \geq u_{min}$) and ($h_p \geq h_{min}$) 17 then 18 $C \leftarrow \text{DBSCAN}(\nu_i, d_{max}) \triangleright$ Clusters 19 $H^* \leftarrow \text{CENTROID}(C) \triangleright$ Hub candidates 20 end 21 end 22 foreach $H^*[i]$ do 23 $deg_{in} \leftarrow \text{GETINDEG}(H^*[i])$ 24 $deg_{out} \leftarrow \text{GETOUTDEG}(H^*[i])$ 25 if $deg_{in} \geq deg_{min}$ and 26 $deg_{out} \geq deg_{min}$ and 27 $\left \frac{deg_{in}}{deg_{out}} - 1 \right \leq \epsilon$ then 28 $H \leftarrow H^*[i]$ 29 end 30 end 31 end </pre>	<p>Input: Networks to be conflated S, G</p> <p>Output: Network of Interest</p> <pre> 1 begin 2 /*Network layers fusion to extract the final map*/ 3 \triangleright edges and nodes of Semantic and Geometric layers 4 $E_S \leftarrow \text{EDGES}(S)$, $N_S \leftarrow \text{NODES}(S)$ 5 $E_G \leftarrow \text{EDGES}(G)$, $N_G \leftarrow \text{NODES}(G)$ 6 $H \triangleright$ Hubs 7 $H_G \triangleright$ hubs \cap geometric nodes 8 $H_S \triangleright$ hubs \cap semantic nodes 9 $H_O \triangleright H - H_G - H_S$ 10 \triangleright Node alignment 11 foreach $H[i]$ do 12 \triangleright finding Nearest Neighbors 13 $H_G \leftarrow (H[i], \text{NN}(H[i], N_G))$ 14 $H_S \leftarrow (H[i], \text{NN}(H[i], N_S))$ 15 end 16 \triangleright Node alignment 17 foreach $H_G[i]$ do 18 $H_O \leftarrow (H_G[i], 1\text{-NN}(H_G[i], H_S))$ 19 \triangleright Node insertion to semantic layer 20 foreach ($H_G[i] \notin H_O$) do 21 $E_i = \text{ON}(E_S, H_G[i])$ 22 if $E_i \neq \text{NULL}$ then 23 $H_S.add(H_G[i])$ 24 $E_S.delete(E_i)$ 25 end 26 end 27 \triangleright Link insertion 28 foreach ($H_G[i] \notin H_S$) do 29 $H_S.add(H_G[i]) \triangleright$ remaining nodes 30 foreach ($E_G[i] \notin E_S$) do 31 $E_S.add(E_G[i]) \triangleright$ remaining links 32 end 33 end 34 end </pre>

4.2 Layer Fusion

The final part of the process comprises the fusion of the geometric and semantic NOI layers. We construct the NOI by starting with the semantic layer and merging the geometric layer onto it. The intuition for this is that the semantic layer corresponds to a geometrically abstract but semantically richer user movement that contains relevant transportation hubs. The geometric layer corresponds to a less semantic but more accurate depiction of movement, i.e., fills in the gaps of the semantic layer. The fusion of these layers should result in a comprehensive movement network.

The fusion task involves (i) finding hub correspondences among the different network layers and (ii) introducing new links of geometric to the semantic layer for the uncommon portions of the NOI.

Using both layers and the hubs, we try to identify common nodes by spatial proximity (Algorithm 3, Lines 11-13). Any node from the geometric layer that has not been introduced yet since it is not connected to the semantic layer will be added (Lines 22-23). The next step involves introducing new links for uncommon portions of the layered network. Here links of the geometric layer are introduced by adding them to the semantic layer (Lines 28-30). Typically this accounts for the cases of adding complete (local) network portions.

A result of applying this conflation algorithm to network layers is shown in Figure 3. The hub correspondences in the Semantic and the Geometric layer result in the Fused Network, which constitutes the Network of Interest. In this figure, characteristic hubs are circled in each layer and are connected to illustrate the layer fusion process.

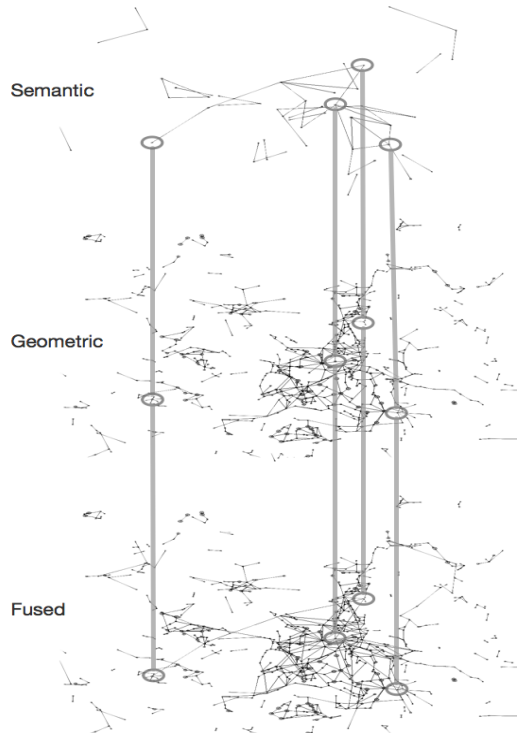


Fig. 3: London - Fused Network

5 Experimental Evaluation

An assessment of the quality of a Network of Interest is a challenging task as there is no ground truth data. In the case of map-construction algorithms, an existing road network can be used. However, a NOI represents a geo-semantic construction containing aspects of both, regular transportation networks (roads, public transport, etc) but also the overall movement sentiment of users in a city. For the following evaluation, we will use a combination of existing POI datasets and (public) transportation networks to assess the constructed NOIs. Before giving details of the experimental results and constructed NOIs, we first describe the characteristics of the datasets used and our overall evaluation methodology.

5.1 Experimental Setup

We conduct experiments on two real-world datasets comprising geocoded Tweets retrieved for London and New York City over a period of 60 days. To focus on trajectories of active users, we kept only the trajectories of the top 200 geo-tweeters

Algorithm	Value
Segmentation of Trajectories	
Mean Speed	10km/h
Time Interval	5, 60min
Geometric NOI	
Distance Threshold	100m
Minimum Number of Samples	2
Semantic NOI	
Distance Threshold	300m
Minimum Number of Samples	2
Extraction of Hubs	
Minimum Number of Samples	10
Minimum Number of Users	2
Minimum Number of Time Periods	10
Distance Threshold	300m
Layer Fusion	
Distance Threshold	50m

Table 1: Parameter summary

for each respective city. Moreover, we only consider trajectories consisting of least 5 position samples. Figure 1 visualizes the movements of 200 Twitter users during the course of a single day in London. Notice that some very prominent areas, such as highways, can be distinguished visually even before any processing of the data takes place. However, our approach enables an automated extraction of a more refined and enriched Network of Interest.

Through experimentation, we established the parameters for the various steps of the algorithm as summarized in Table 1. For our evaluation, we consider as ground truth the corresponding public transportation network obtained by OSM [23]. What follows is a brief description of the trajectories collected from the geocoded tweets, as well as the networks obtained from OSM for the two cities of London and New York.

In London, the actual public transportation network consists of 27,021 links (edges) and 47,575 nodes and has a length of 21,287km. It covers an area of 420km × 118km. The check-in data covers a great portion of this network and has a total length of 256,400km (Figure 1). It consists of 463 trajectories with an average length of 7.2km. The average sampling rate is 35min, while the average speed is 92km/h. In New York, the actual public transportation network consists of 84,367 links (edges) and 75,070 nodes and has a length of 9,846km. It covers an area of 105km × 85km. The check-in data consists of 37,962 trajectories, with an average length of 1.9km and total length of 214,090km. The average sampling rate is 55min, while the average speed is 76km/h.

5.2 Visual Comparison

A first and quick overview of the quality of the inferred Network of Interest can be obtained by *visual inspection*, i.e., by overlaying it on the reference network and looking for similarities and differences. In this way one can assess how well

the constructed network lines up with the actual network as in our case a public transportation network.

Figure 4 visualizes the NOIs of the cities of London (Figure 4a) and New York (Figure 4b). In each case, the constructed network is visualized using black lines, while the ground-truth network is shown using light gray lines. As evident, especially for the case of New York, the constructed NOI lines up with the transportation network and identifies major hubs.

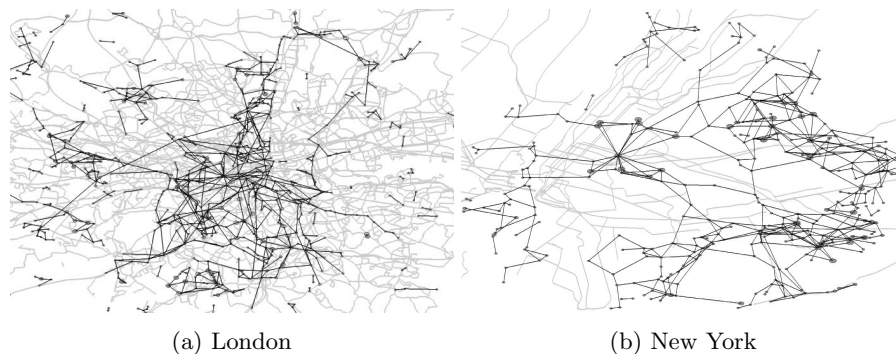


Fig. 4: Networks of Interest

5.3 Quantitative Evaluation

For a more systematic and quantitative assessment of NOIs, we devise two means, (i) comparing the constructed NOI to the geometry of a respective transportation network and (ii) comparing the nodes of our NOI with a POI dataset to discover semantics in terms of their type. This approach allows us to assess the similarity with respect to the ground-truth network and draw conclusions with respect to not only the spatial accuracy of the result, but also the semantics of the nodes.

To *compare networks*, given the constructed and ground-truth public transport network, we select all the nodes of the constructed network and identify corresponding nodes in the ground-truth network by means of nearest-neighbor queries. Using the OSM public transport data, we select for every hub of the Network of Interest the nearest node in the OSM data. If the inferred nodes are close to the actual transportation network nodes, then the constructed NOI closely relates to the transportation network.

To discover the *type of transportation node* a hub represents, e.g., bus, metro, tram and railway, we again use OSM data. We apply reverse geocoding (identify POIs based on coordinates) to the NOI hubs to identify actual POIs corresponding to these locations. We determine which of the constructed nodes corresponds to nodes in a transportation network by comparing the description obtained by reverse geocoding to the nodes of the ground-truth network. The results are

	Nearest Neighbor Statistics		Reverse Geocoding Statistics	
	Found	Total	Ratio %	
London	1389	1562	89	964 1562 62
New York	1423	1649	86	873 1649 53

Table 2: Evaluation summary

summarized in Figure 5, which shows the degree of a node, i.e., the number of incoming and outgoing links. Nodes identified as transportation nodes have higher degrees (≥ 20) when compared to other nodes. In this case we use the degree as an indicator for the importance of the node and the fact that high-degree nodes were identified as transportation nodes allows us to reason about the type of network we constructed.

Essentially, the nearest-neighbor queries evaluate the spatial accuracy of the NOI, while the reverse geocoding evaluates the semantics of the node extraction. The higher the score of the correctly extracted nodes, the higher also the quality of the network. As shown in Table 2, transportation nodes are inferred with high accuracy. Indeed, 89% of the extracted hubs in London and 86% in New York are located close to transportation nodes in the OSM ground truth network. In the case of the reverse geocoding test, the ratios are a bit lower due to the fact that the reverse geocoding service returns only POIs that are located exactly or very closely at the queried coordinates.

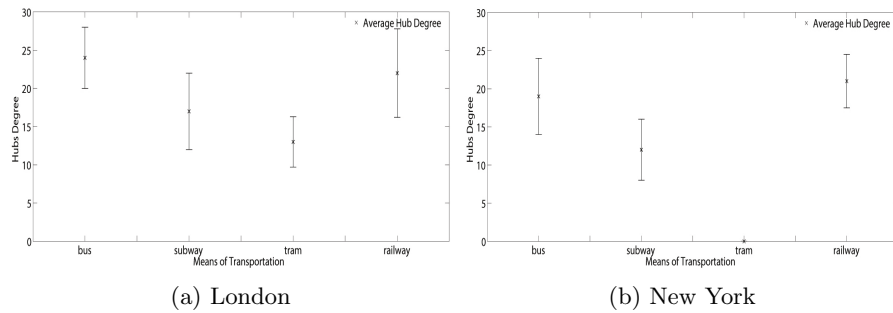


Fig. 5: Hubs Statistics

An overall sentiment of our experimentation could be that the network construction process results in a Network of Interest that captures certain aspects of a public transportation network. A core problem in such experimentation is that using social media check-in data as a tracking data source and construction a network has the inherent challenge that no actual ground-truth data is available to assess the quality of the result. Using in our case a public transportation network allows us to show some similarities, however, the constructed NOI could not be completely mapped (explained) by it as it represents a more complex

network whose characteristics cannot be captured by a single existing network dataset. These concerns are also issues we want to address in future work.

6 Conclusions

Social media data has been used in a wide range of data mining applications. This work however is the first to consider social media as a tracking data source and constructs movement information in the form of *Networks of Interest* (NOI) from geocoded Twitter feeds. The NOI construction algorithm is based on segmenting the Twitter check-in data and constructing two separate network layers. The layers are a geometric and a semantic layer representing the low-level geometric and a more high-level, semantic aspect of the NOI, respectively. Subsequently, we identify network hubs, which are used for fusing the layers to a complete Network of Interest. Performing an experimental evaluation using two large-scale datasets, the algorithm produces a NOI of considerable accuracy. The resulting NOIs capture portions of the ground-truth transportation network geometry, both, in terms of spatial accuracy and semantic precision.

The directions for future work are on refining the NOI construction process and using it for larger datasets to extract more complex NOIs. Here, we will also have the opportunity to identify temporal aspects of the NOIs, e.g., transportation routes to and from a city, temporal variations, as well as characteristics of the NOI graph itself (connected components). We are also in the process of applying the proposed methods to mobile phone tracking data, as such a dataset would be “in between” GPS tracking data and check-in data, both, in terms of positional accuracy and sampling rate.

Acknowledgment

This work was supported by the EU FP7 Marie Curie Initial Training Network GEOCROWD (FP7-PEOPLE-2010-ITN-264994).

References

1. Ahmed, M., Karagiorgou, S., Pfoser, D., Wenk, C.: A comparison and evaluation of map construction algorithms. Under submission (2013)
2. Biagioni, J., Eriksson, J.: Map inference in the face of noise and disparity. In: Proc. of the 20th ACM SIGSPATIAL GIS conf. (2012) 79–88
3. Kisilevich, S., Keim, D.A., Rokach, L.: A novel approach to mining travel sequences using collections of geo-tagged photos. In: Proc. of the 13th AGILE Int’l conf. on Geographic Information Science. (2010) 163–182
4. Chen, Z., Shen, H.T., Zhou, X.: Discovering popular routes from trajectories. In: Proc. of the 27th Int’l conf. on Data Engineering. (2011) 900–911
5. Zheng, Y.T., Zha, Z.J., Chua, T.S.: Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology* **3**(3) (2012) 56:1–56:18

6. Asakura, Y., Iryo, T.: Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transportation Research Part A: Policy and Practice* **41**(7) (2007) 684 – 690
7. Mckercher, B., Lau, G.: Movement patterns of tourists within a destination. *Tourism Geographies* **10**(3) (2008) 355–374
8. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Constructing travel itineraries from tagged geo-temporal breadcrumbs. In: *Proc. of the 19th Int'l conf. on World Wide Web.* (2010) 1083–1084
9. Lloyd, S.: Least squares quantization in pcm. *IEEE Transactions on Information Theory* **28**(2) (2006) 129–137
10. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In: *Proc. of the 1996 ACM SIGMOD Int'l conf. on Management of Data.* (1996) 103–114
11. Ester, M., Kriegel, H.P., S, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proc. of the 2nd Int'l conf. on Knowledge Discovery and Data Mining.* (1996) 226–231
12. Bollobás, B., Das, G., Gunopulos, D., Mannila, H.: Time-series similarity problems and well-separated geometric sets. In: *Proc. of the 13th Annual Symposium on Computational Geometry.* (1997) 454–456
13. Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsis, I., Andrienko, G., Theodoridis, Y.: Similarity search in trajectory databases. In: *Proc. of the 14th Int'l Symposium on Temporal Representation and Reasoning.* (2007) 129–140
14. Karagiorgou, S., Pfoser, D.: On vehicle tracking data-based road network generation. In: *Proc. of the 20th ACM SIGSPATIAL GIS conf.* (2012) 89–98
15. Karagiorgou, S., Pfoser, D., Skoutas, D.: Segmentation-based road network construction. In: *Proc. of the 21th ACM SIGSPATIAL GIS conf.* (2013) 470–473
16. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: *Proc. of the 18th Int'l conf. on World Wide Web.* (2009) 761–770
17. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A.A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: *Proc. of the 11th Int'l conf. on Computer Vision.* (2009) 253–260
18. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: *Proc. of the 30th Annual Int'l ACM SIGIR conf. on Research and Development in Information Retrieval.* (2007) 103–110
19. Yanai, K., Kawakubo, H., Qiu, B.: A visual analysis of the relationship between word concepts and geographical locations. In: *Proc. of the ACM Int'l conf. on Image and Video Retrieval.* (2009) 13:1–13:8
20. Girardin, F., Calabrese, F., Fiore, F.D., Ratti, C., Blat, J.: Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing Magazine* **7** (2008) 36–43
21. Kisilevich, S., Mansmann, F., Keim, D.: P-dbscan: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: *Proc. of the 1st Int'l conf. and Exhibition on Computing for Geospatial Research and Application.* (2010) 38:1–38:4
22. Kling, F., Pozdnoukhov, A.: When a city tells a story: Urban topic analysis. In: *Proc. of the 20th Int'l conf. on Advances in Geographic Information Systems.* (2012) 482–485
23. OpenStreetMap Foundation: Openstreetmap: User-generated street maps (2013)