

# TravelDiff: Visual Comparison Analytics for Massive Movement Patterns Derived from Twitter

Robert Krueger\*  
VIS, University of  
Stuttgart

Guodao Sun†  
Zhejiang University  
of Technology

Fabian Beck‡  
VISUS, University of  
Stuttgart

Ronghua Liang§  
Zhejiang University  
of Technology

Thomas Ertl¶  
VIS, University of  
Stuttgart

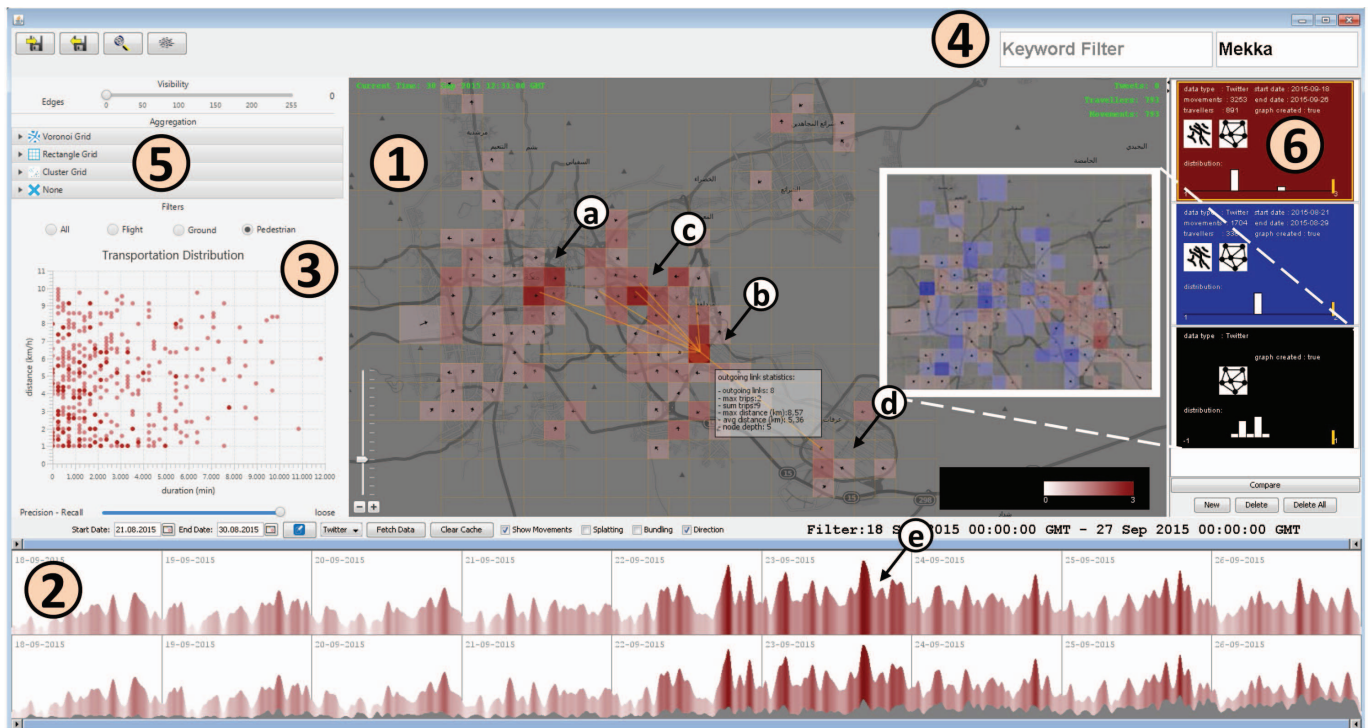


Figure 1: The visual analytics system for exploration and comparison of movements derived from geo-located microblog messages consists of multiple linked views: (1) map, (2) timeline, (3) velocity distribution & filtering, (4) searchboxes, (5) grid creation panel, (6) comparison interface. The map shows distribution and direction of pilgrim movements in Mecca, Sep. 2015. Patterns (labeled a–e) are described in a case study (Section 5.1).

## ABSTRACT

Geo-tagged microblog data covers billions of movement patterns on a global and local scale. Understanding these patterns could guide urban and traffic planning or help coping with disaster situations. We present a visual analytics system to investigate travel trajectories of people reconstructed from microblog messages. To analyze seasonal changes and events and to validate movement patterns against other data sources, we contribute highly interactive visual comparison methods that normalize and contrast trajectories as well as density maps within a single view. We also compute an adaptive hierarchical graph from the trajectories to abstract individual movements into higher-level structures. Specific challenges that we tackle are, among others, the spatio-temporal sparsity of the data, the volume of data varying by region, and a diverse mix of means of transportation. The applicability of our approach is presented in three case studies.

\*e-mail: robert.krueger@vis.uni-stuttgart.de

†e-mail: godoor.sun@gmail.com

‡e-mail: fabian.beck@visus.uni-stuttgart.de

§e-mail: rliang@zjut.edu.cn

¶e-mail: thomas.ertl@vis.uni-stuttgart.de

## 1 INTRODUCTION

The whole world is moving. Around 90,000 flights are scheduled each day, the number of cars surpassed one billion, and we travel around 130,000 kilometers in our lives by foot. Human movement patterns are valuable to better understand and improve our lives in many domains: City planners enhance living conditions by optimizing traffic flow, housing, shopping, and working areas [27]. Road networks and traffic control can be optimized with knowledge about driving behavior and route congestions [28]. For the management and prevention of crises, trajectory data is used to simulate and analyze crowded situations [37]. In epidemic research, movement data is a basis for propagation models [23].

When global movement is of interest, patterns are usually derived from flight schedules. On a lower scale, cell phone and public transportation data provide valuable information of urban movement such as daily commute [36]. Unfortunately, these datasets are often incomplete and of limited temporal and geographical resolution and scale. They can be expensive and hard to retrieve because they are monetized by large companies or cover privacy-sensitive data.

With the rise of smartphones and social media platforms, humans have become sensors themselves, contributing geo-located data points every day. On microblog platforms, many users send public messages with location information that can be used to retrieve a trajectory estimating the users' movement. Twitter users, for example, produce more than 200 million geo-located messages a month [43]. Based on this data, we can reconstruct movements of different scale,

from global air traffic to local ground transportation. The data is available for large timespans and often free to record and analyze.

Recently, a number of research projects have started to interpret movements reconstructed from microblogging platforms [13, 18, 19, 20, 25, 38, 42, 45] (see Section 2). While these approaches already reconstruct and classify trajectories, we still face many open research questions and challenges: Are the reconstructed trajectories reflecting real world patterns, comparable to explicit movement data for urban analysis? What similarities and differences exist when comparing such heterogeneous datasets and how can we make them visible? How to deal with varying spatial and temporal granularities, distribution, and volumes?

To address these questions, we contribute *TravelDiff*, a visual analytics system (Section 3) to explore and compare reconstructed trajectories from geo-located microblog messages. It combines visualization methods to depict the differences of explicit movement data and movements derived from tweets with interactive filters (Section 4). We propose different approaches like alpha blending and edge splatting that work on the raw data. Moreover, we present a hierarchical aggregation method that allows for a smooth transition from overview to detail. The applicability of our system is presented in Section 5. We first explore events based on temporal changes (Section 5.1). In particular, we focus on the pilgrimage to Mecca in September 2015 where we identify hotspots and travel directions. We then exemplarily evaluate how the reconstructed movement datasets reflect human mobility patterns by comparing the movements to daily city commute (Section 5.2) and to flight schedules (Section 5.3). We consider our approach as a step towards a better understanding of movement patterns derived from microblog messages and as an enabler for a comparative analysis of such data (Section 6).

## 2 RELATED WORK

The visualization of movement data is a broad research field and was surveyed by Andrienko and Andrienko [5]. They demonstrate how movement data can be analyzed in 2D or 3D map overlays as well as with timeline views and hierarchy representations. Our work focuses on the reconstruction, exploration, and comparison of movement data derived from social media (i.e., microblog services). Hence, we discuss specifically those visual analysis approaches that (1) reconstruct movement data from geo-located microblog messages or (2) support the comparison of trajectories.

### 2.1 Movements from Microblog Data

While the visual analysis of textual content from social media is a well-developed research field, movement extraction from social media recently gained more and more attention. Senseplace2 [33] and ScatterBlogs2 [14], and approaches by Fuchs et al. [25] allow for a location-based analysis of Twitter messages. Others even trace the geographic path and evolution of topics over time: Sakaki et al. [38] extracts earthquake related microblog messages and applies a Kalman filter to reconstruct an earthquake’s trajectory. Similarly, Senaratne et al [42] looks for hotspots of a certain topic and reconstruct a trajectory of a musician’s concert tour across the USA.

Besides content analysis there is new research that investigates human movement derived from geo-located microblog data. Chae et al. [18] reconstruct movements of Twitter users and map them to the road network to overcome data irregularities. They further present an expert-driven interactive analysis to identify outliers. Interactive filtering of movements derived from Twitter is also shown by Chen et al. [19]. In their work origin–destination (OD) patterns can be found with a *focus+context* technique. Chen et al. [20] also developed a full-fledged system for the analysis of movements from Sina Weibo, a Chinese microblogging platform. They address major challenges such as large data volumes, irregular sampling, data sparsity, and resulting uncertainty. A Gaussian-mixture model is used to categorize movements to different means of transportation. While our main focus is on data comparison, we also provide a categorization method to classify means of transportation. By contrast, our method is rule-based and allows for an interactive trade-off between precision and recall. With a more regional and application-specific focus, Blanford

et al. [13] analyze cross-border movement in Africa derived from Twitter and build graph models to aggregate the movements. Flow Sampler [22] also adapts a graph-based methods and infers movement pathways from point-based geo-located microblog data to explore movement flow. Recently, Landesberger et al. [45] contribute another visual approach to analyze temporal changes. While they also present a comparison graph, they do not provide a hierarchical structure that allows for a smooth transition between overview and detail. Also, they focus on Twitter analysis only and do not compare the patterns to other data.

### 2.2 Comparison of Trajectories and Graphs

In their taxonomy of visual comparison, Gleicher et al. [26] discern three types of comparison methods: *juxtaposition*, *overlay*, and *explicit encoding*. These general approaches are applicable to many visualization problems but have different advantages and limitations. For a visual comparison of trajectories, we could juxtapose small multiples, each representing a different version of the dataset [29]. Also, different groups of trajectories can be overlaid while discerning the groups by coloring [5] or 3D stacking [5, 44]. Overlay or juxtaposition approaches work well to reveal coarse differences. However, overlays get difficult to read if too many visual entities overlap and, in juxtaposed images, only large differences can be retrieved easily while identifying small changes becomes a *spot-the-difference game*. Since we want to compare rather subtle differences between variants of similar movement networks, we focus on explicit encoding: differences and commonalities are explicitly highlighted in order to guide the viewer to interesting patterns.

To gain overview of massive trajectory data, graphs provide the necessary level of abstraction by aggregating trajectories to transitions [5]. Various approaches for visual graph comparison have already been proposed [4, 46]. Also related are dynamic graph visualization techniques, which compare changing graphs over time [10]. Among the comparison approaches, some employ explicit encodings: In node-link diagrams, differences are usually encoded with colors [7, 8, 9, 40]. Also, adjacency matrix representations show differences between two graphs using color [11] or glyphs [32]. In matrix representations, however, the geographic information, which is important for our analysis scenario, is not preserved. In an interactive comparison approach, Beck et al. [12] compute union, intersection, and difference only on demand and thereby let the user control what should be visualized.

We apply a similar interactive approach but in a hierarchical manner, similar to [3], giving the user even more control over the merging and filtering process of trajectories. We offer both, the visual analysis of a single group of trajectories as well as the visual comparison of two groups of trajectories in a *diff* view.

## 3 SYSTEM DESIGN

Geo-located Twitter data and its reconstructed movements have certain peculiarities that require special treatment. We identified four specifics (S1-S4) and resulting requirements that go beyond common movement analysis practices:

- S1 **Implicit and unlabeled data.** Movements need to be reconstructed from geo-located microblog messages. This requires several preprocessing steps, including bot removal and categorization of means of transportation.
- S2 **Massive and unstructured.** From about seven million tweets a day, the reconstructed movement data is massive but often only consist of origin and destination points. Hence, movements do not follow an underlying network (e.g. roads). This can result in overplotting and cluttering. Clear visualizations, which preserve the density distribution are needed.
- S3 **Varying scale and resolution.** Analysts are interested in movements of different scale, from commutes to intercontinental flights. However, varying by region, data resolution is low. Detailed visualization of uncertain movement can be misleading. Adaptive data aggregation helps to solve the challenge.

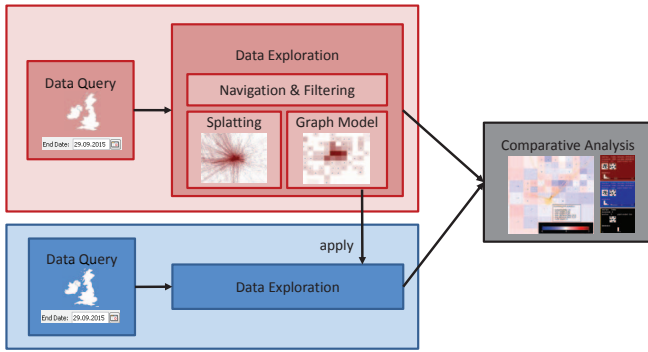


Figure 2: The abstract analysis process consists of an exploration part and a comparison part. Colors indicate an exploratory analysis for two different datasets, that precedes the comparative analysis.

**S4 Uneven distribution and unknown quantities.** Different data has different characteristics, such as resolution, distribution, and structure. Moreover, it is difficult to compare quantities derived from microblogs to real world quantities such as number of travelers. Normalization and comparison methods that handle different data distribution and outliers are needed.

With respect to these data specifics and resulting requirements, our analytics system (see Figure 1) orchestrates visual means to explore and compare large movement data of different transportation means, in different views. Using comparison visualizations, we contrast the data to other datasets or analyze seasonal changes and events in the microblog movements. All views are connected with brushing and linking. North and Shneiderman [35] showed in a user study that multiple coordinated views can improve user performance by 30–80% over detail-only and uncoordinated interfaces for most tasks.

Before we explain the design rationales (Section 4) based on the identified requirements, we want to give a brief overview of the analysis process, illustrated in Figure 2. The analyst starts the exploration process by selecting a region and time of interest. This is done with a polygon tool and a calendar widget. After data querying and loading, the movements can be explored. The analyst may first take a look at the map, located in the center (see Figure 1, 1), exploring spatial movement distribution and flow of messages. The map provides both detail visualizations for investigating the raw data and layers aggregating data for overview and advanced filtering. The abstracting structure—a hierarchically aggregated graph model—is configured in the grid configuration panel (5). The timeline view (2) allows for temporal navigation and analysis of time-dependent patterns. Depending on the analysis goals, the movements might be filtered by means of transportation; a scatterplot (3) shows the velocity distribution for the results. The system supports keyword filters and a geographic search to quickly jump to a specific location (4). For data comparison, a second dataset can be loaded, e.g., Twitter, flight, or taxi data. For a comparative analysis, the interface to the right (6) allows switching between the two datasets and a comparison visualization. Comparisons are either done on a raw data level or in the aggregated model.

## 4 APPROACH

*TravelDiff* integrates a blend of approaches that address the identified data specifics and requirements in order. Each of them (S1–S4) is described in detail in the following. We first discuss trajectory reconstruction and other preprocessing steps (Section 4.1). Secondly, to overcome occlusion and to address varying data distribution, we present a splatting-based visualization method (Section 4.2). We then introduce an aggregation technique that allows for abstraction and stepwise visual transition from overview to detail (Section 4.3), before we present visual comparison approaches (Section 4.4).

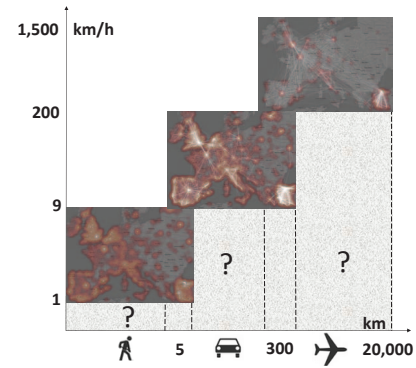


Figure 3: Rule-based categorization to assign the most likely means of transportation. Fuzziness can be interactively controlled with a precision&recall slider.

### 4.1 Preprocessing

To reconstruct movements from Twitter messages (S1), we apply a set of methods to sort and filter the data. We improve data quality by eliminating computer-generated messages and use a rule-based categorization to classify movements according to their means of transportation.

**Trajectory Composition** Around 1–2% of the messages from Twitter contain detailed geo-location consisting of latitude and longitude information. With respect to overall volumes of around 500 million tweets a day [2], it still sums up to around 200 million geo-located messages per month. Many users share messages frequently, i.e., several times a day or a week. From the location and temporal information, movement trajectories can be derived. We define a trajectory  $TR$  as follows.

$$TR = ((p_1, t_1), (p_2, t_2), \dots, (p_n, t_n)) \text{ with } t_i < t_{i+1} \text{ for } i = 1, 2, \dots, n-1.$$

$p_i$  is the geo-location of a message at time  $t_i$ .

There is a varying number of single messages that cannot be connected to any movement. The number of reconstructed trajectories highly depends on the selected area (region and size) and timeframe. On average, we are able to connect about 30% of the loaded messages to trajectories when the timeframe is at least a day.

**Bot Detection and Filtering** Twitter contains a large number of computer-generated (bot) messages. While we cannot guarantee to filter out all bot messages, we can significantly reduce their amount by applying content and movement-based filters.

- Content-based filters are widely used for bot removal in microblog data. We apply basic techniques, based on the observations by Lee et al. [30]. This includes a dictionary of keywords, such as *weather report* and *special offer*, defining a black-list to filter out news and ads. Also, we exclude user accounts with high word and sentence repetition. We plan to integrate more elaborated techniques in the future [21, 17].
- Besides content, certain spatio-temporal patterns relate to bots. We assume that human users tweet at maximum with an average daily frequency of two messages per hour. Higher frequencies (>48 tweets a day) usually indicate bots (~3–5%). Also, even fast passenger aircrafts no dot exceed a speed of 1,500 km/h. If the derived movement speed of subsequent tweets exceeds this threshold multiple times, we exclude the corresponding user, because it is presumably a bot.

**Movement Classification** Often, there are days between two consecutive messages of a user and it is unknown where the user has been in between or if she has arrived much earlier at the second location than the message was sent. Thus, classification of the used means of transportation is a non-trivial problem. To infer the most likely means, we apply different categorization rules (see Figure 3).



We split the data into three main categories (pedestrian, ground transportation, flight), since the data sparsity and resulting uncertainty is too high for a finer-grained classification. Still, it is challenging to define a clear boundary among different transportation categories. We apply a set of heuristics. Humans typically tend to walk at about 5 km/h [15], up to 9 km/h [34]. Similarly, for ground transportation we set the threshold to 300 km and above 5 km, the upper pedestrian walking distance. A fast passenger aircraft does not exceed a speed of 1,500 km/h and the maximum distance of a flight cannot exceed 20,000 km, as this is half the circumference of the earth and not possible without refuel. With these assumptions, additional data is further filtered out. However, depending on the region and timespan to be analyzed, this can lead to a small data sample that is susceptible to outliers. Hence, we provide interactive means to widen those strict filter settings and let the analyst choose a trade-off between precision (more reliable categorization) and recall (higher data volumes).

## 4.2 Raw Data Visualization

A straight-forward way to visualize the reconstructed movements (S2) would be to simply draw trajectories on the map. Unfortunately, this results in clutter and overplotting (see Figure 4, left).

A more scalable approach to visualize the movements is alpha blending (center). By decreasing the opacity of the trajectories, visual clutter can be reduced and trajectories with high density are clearly conveyed. However, alpha blending introduces a new problems. Firstly, the quantity of trajectories to be drawn largely depends on the selected area and timespan. Users need to adjust the opacity when drawing trajectories under different circumstances to avoid overplotting in dense areas. This issue may reduce efficiency of users' visual exploration. Secondly, encoding flexibility is limited. Alpha blending results in a linear density mapping. Hence, it is often not possible to perceive both, small and large density differences within a single visualization.

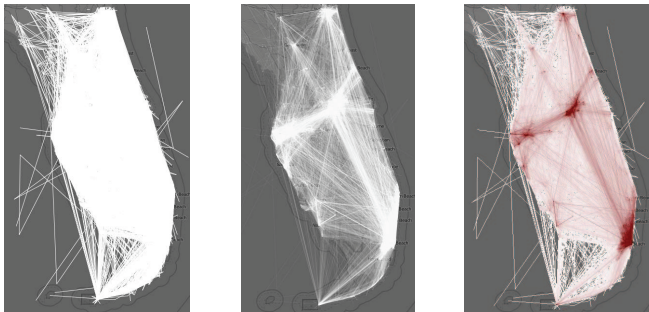


Figure 4: Drawing of large movement data. From left: full opacity, alpha-blending, edge splatting with normalization and color mapping.

To overcome these problems, we further support edge splatting (see Figure 4, right) [16]. Similar to Scheepens et al. [39], we transform the trajectories to a density distribution of pixels in a 2D matrix with the same pixel size and ratio as in the current viewport. More specifically, we iterate all trajectories and determine the intersecting pixels. Each pixel finally holds an amount of intersecting trajectories and we can easily obtain the minimum and maximum density in the distribution. We finally map the density of each pixel to color. We apply a non-linear mapping to approximate perceptual linearity [31] and to reduce sensitivity to outliers. Figure 9 shows edge splatting results for two large datasets. The calculated density values are mapped to different color schemes ranging from white (low density) to red or blue (high density).

## 4.3 Hierarchical Graph Model

For the exploration of movement volumes, edge splatting is an adequate technique. In addition to this, we also developed a graph-based method that especially addresses the data specifics of Twitter data (S2, S3). Firstly, the exact path between two geo-located messages is unknown. The more time is spent between two locations, the less certain it is that the user took a straight path (and uniform speed) to reach

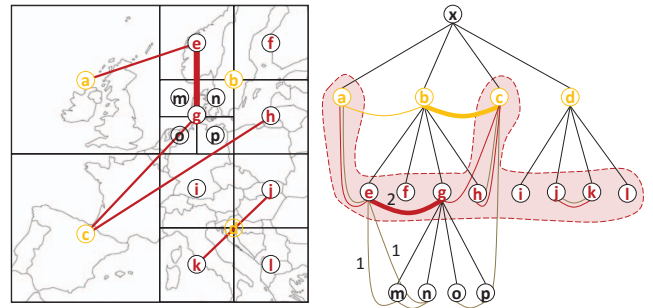


Figure 5: The hierarchical graph model and weight-based link propagation. Original links in brown (leaf level) are propagated to upper levels (red, yellow), allowing for more abstract/aggregated views.

the destination. With edge splatting, however, the direct path between two consecutive positions is visualized. Secondly, microblog data volumes vary strongly, depending on timespan and selected region. Peaks in popular locations skew the data distribution and the large number of trajectories from or to these peaks can occlude underlying patterns. While edge splatting mitigates such clutter and occlusion, it is still difficult to understand traffic flows, especially when many trajectories with different angles are displayed.

Aggregation of movement data can help to overcome the aforementioned issues and is often achieved by graph computation. Most graphs, however, have the drawback of a fixed, pre-computed level of aggregation. Our approach, on contrast, creates a hierarchical graph structure that allows for a stepwise data aggregation and abstraction from detail to overview. It hierarchically subdivides the geographic regions based on their data volumes. For this recursive subdivision, different structures could be applied. While political boundaries (e.g., countries, states) are a common way to bin geographic data, our current implementation relies on an adaptive rectangular grid (quadtree). The advantage of our method is that it bins the data depending on the natural distribution rather than creating hard cuttings on political borders. However, as country-based statistics are often of interest, we also want to support other shapes in the future.

To create the grid, we use all locations of Twitter messages for the selected region and time. The reason why we do not only use the movement points is that all messages better reflect the Twitter population and we reduce sparsity problems in regions with low message density. At first, the analyst defines a cut threshold. This means that each cell is split when the cut value is exceeded, leading to large cells in areas with low population and to a finer grid structure (smaller cells) in densely populated regions (see Figure 5, left). Depending on the underlying data and task, one can create flat or deep hierarchies with stronger or weaker aggregation per cell. Next, the movement data is used to compute directed links between the grid cells on leaf level. Each link holds the number of movements between two cells as weight. For example, a link from London to Paris can have a weight of a few hundred movements for a single day. Figure 5 illustrates the spatial properties of the grid on the left and its data structure on the right side. The original links on leaf level are shown in brown. Having a cutting threshold of one leads to the highest detail, because each data point is represented as a single cell on the leaf level. By contrast, looking at a higher level of the hierarchy provides more overview.

To achieve a smooth transition from detail to overview, we apply a link propagation algorithm that sums up the links weights to a higher tree level. Once calculated, the tree holds information for all levels and does not need to be computed again. The algorithm starts to propagate the links from the leaves to the first (red) aggregation level, while links disappear that point to direct siblings, and others get stronger because their weight sums up. Likewise, links for the next aggregation level are illustrated in yellow. Depending on the selected level, the respective nodes and their links are visualized in the map. Figure 5 illustrates the red link level. This also includes leaves with links that are shallower than the current selected level. Note that the illustration uses only a few links, while the quadtree can be large, holding thousands of nodes and links that have different directions.

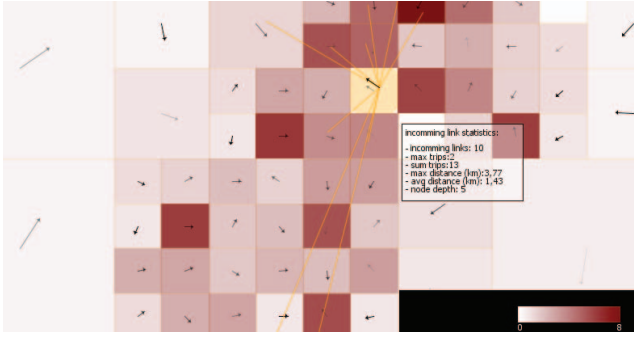


Figure 6: The grid structure is colored according to incoming or outgoing movement. Arrows indicate average movement direction. Standard deviation is encoded with varying opacity. A tooltip shows additional information and highlights the movements for the underlying cell.

Figure 10 shows results for microblog data in the USA and eastern Asia. Grid cells represent the nodes of the graph. They are colored according to the number of incoming or outgoing link weights, which leads to a grid-based heatmap encoding aggregated movement volumes. To take the spatial area of each cell into account, we normalize the number of links with respect to the cell size.

Besides drawing links, the gridded heatmaps can also be overlaid with arrows that indicate movement direction by showing the average angles of outgoing or incoming links, respectively, one per cell (see Figure 6). Increasing standard deviation of movement directions is encoded by increasing arrow transparency. While the link-based visualization shows main connections, the arrows have other benefits: they are free of clutter and reduce misinterpretation of routes traveled caused by sparse data resolution.

#### 4.4 Movement Comparison

With the techniques used to construct and visualize raw movements as well as aggregated ones, we can compare patterns on different levels of detail (S4). The statistics of the datasets loaded appear in the comparison interface to the right of the map view (see Figure 1). To help the analyst to distinguish between both datasets, the first dataset loaded is colored in red, the second one in blue.

The comparison of movement data is challenging. When comparing different datasets (e.g., Twitter and flight schedule data or Twitter data for different time periods), we need to normalize the data first. Directly comparing the datasets would lead to severe bias, because their density may not be in the same level of magnitude. We use a min-max normalization to transform the data into the range  $[0.0, 1.0]$ . Compared to the flight schedule and taxi route data, the spatial distribution of the movements extracted from Twitter can have large peaks, that would dominate the normalization and may lead to undesirable and incommensurable results. Often, the rest of the data distribution is hardly visible. To overcome this issue, we apply a cube root function that preserves differences in lower density range while it compresses the peaks to an acceptable level. Also, we allow analysts to further control the normalization.

A histogram first gives an overview on the data’s density distribution. Depending on the visualization in use, it shows the movement distribution in the edge splatting matrix or of link weights in the hierarchical graph structure. The analyst can then interactively shift the maxima densities to a lower range (see Figure 7). All values to the right of the marker will be downgraded to the value at the position of the marker (new maximum). Thus, a highly unbalanced distribution will be transformed to be more flat, and users become aware of the transformation. Together, all adjustments lead to the following equation with  $x'$  for the pixel value in case of splatting/link value for the graph and  $x_{max_{ad}}$  for the adjusted maximum.

$$x' = \alpha + \left( \min\left(1, \frac{x - x_{min}}{x_{max_{ad}} - x_{min}}\right) \right)^\beta * (1 - \alpha)$$

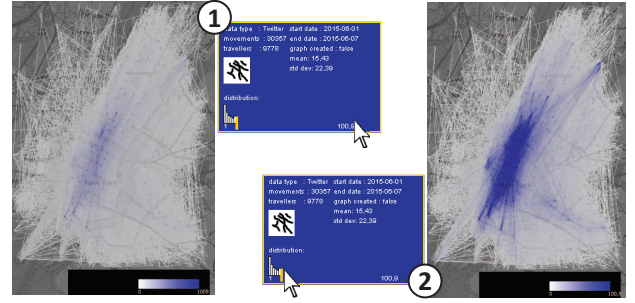


Figure 7: The density distribution (1) of one week Twitter data in Manhattan (see Section 5.2) is dominated by large peaks that can be compressed interactively (2) shifting the maximum to a lower value. Besides the histogram, overall movement count, mean  $\mu$ , and standard deviation  $\sigma$  per pixel give further information about the distribution.

$\alpha$  (presetting: 0.05) is a threshold that preserves visibility for areas with low density) and  $\beta$  (0.33) introduces non-linearity [31] that controls the degree of data smoothing.

After the normalization, users can run a comparison analysis using either edge splatting (detail level) or gridded heatmaps (with either links or arrow glyphs), which rely on the hierarchical graph structure. In edge splatting mode, we simply subtract the value in each pixel of one image from the other image. Thus, the pixel value in the subtracted image is in the range  $[-1.0, 1.0]$ . The example in Figure 9 shows edge splatting results for two datasets and their comparison. For the graph-based comparison, we also apply a subtraction. A precondition is that both graphs have the same layout, i.e., same quadtree structure with same depth and same number of nodes, otherwise the comparison will fail to find the right links to compare. We employ an *auto-filling* strategy to obtain the same hierarchical grid structure for different datasets: users first load a dataset and create a grid. The resulting heatmap can be seen in Figure 8, left. When a second dataset is loaded for comparison, the same hierarchical grid structure is constructed but filled with the movement data from the second dataset.

Afterwards, the comparison of the two hierarchical graphs is performed by subtracting their link weights (Figure 8, right). E.g., a value of 0.5 subtracted with 0.8 leads to a slightly negative value of  $-0.3$ . Links that only exist in one of the datasets result unaltered (e.g.,  $0.0 - 0.8 = -0.8$ ). Lastly, we apply a divergent color scheme using colors, ranging from blue over white (similar) to red.

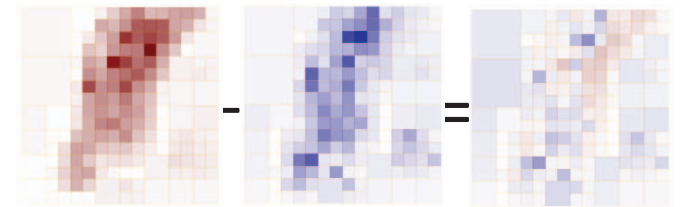


Figure 8: For the first dataset loaded (here, taxis in New York) a grid structure is created (left). The same grid is filled with the second dataset (Twitter data, center). The subtraction is shown on the right.

## 5 CASE STUDIES

We test the applicability of our approach in three case studies. First, we investigate differences in movement data derived from Twitter for different time periods by analyzing the pilgrimage to Mecca. For an exemplary evaluation of the extracted microblog movements, we compare them to other data that is explicitly collected to gain insights in human movement. On a global scale, we investigate similarities and differences to flight schedule data and on a local level, to taxi data, which reflects daily commute patterns.



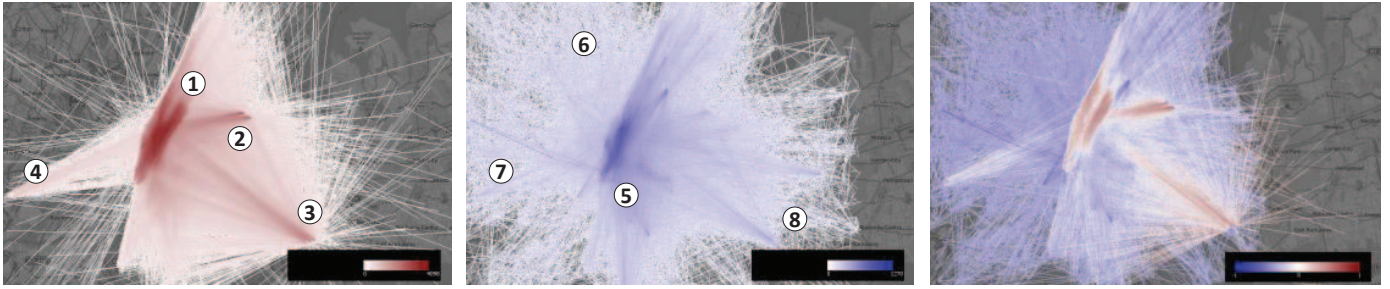


Figure 9: The taxi dataset (left) shows different peaks at New York’s airports (2,3,4) and in the inner city (1). Twitter data (center) has higher peaks and is not that smoothly distributed, but also gives information about movement in the north-west areas (6). The right image shows differences (red, blue) and some similarities (white), for example, the routes to and from Newark airport (cf. 4,7).

### 5.1 Pilgrimage to Mecca

We are interested if the movement data reflects larger events and seasonal changes. In this case study we compare different time periods of trajectories derived from microblog data in the area of Mecca, Saudi Arabia. Mecca is the central pilgrimage destination for Islam. During the *Hajj* (the pilgrimage) more than two million people travel to Mecca to circumambulate *Kabaa*. In September 2015 more than 700 of them died in a mass panic [1] near the *Jamaraat bridge*. We want to investigate if we can identify changes in movement volumes and directions derived from Twitter that could help to gain insights and eventually make the journey more secure. We first load the data for the 2015 *Hajj* and slightly extend the timeframe from 18th to 26th (the day the case study was carried out), to also cover arrivals and departures. Next, we filter for pedestrians only which reduces the dataset of 3,253 movements drastically to 194 movements. Hence, we accept to loose precision to increase recall and widen the filter restrictions (see Section 4.1). This leads to an amount of 793 movements for the selected area and time period. We have a first look using alpha blending and edge splatting. However, the data is not dense enough to produce a smoothly distributed image and we choose to make use of the hierarchical aggregation means (see Section 4.3) to reduce clutter from diverse movement directions. After interactively choosing a suitable level of detail using the hierarchy slider the grid-based heatmap shows clear patterns that are highlighted in Figure 1. Label (a) shows a peak at *Kabaa*, the cuboid building at the center of *Masjid al-Haram*, the sacred mosque. Label (b) points at the area of *Muzdalifah*, another stop of the pilgrimage, just like (d), which highlights the area around *Mount Arafat*. Arrows encode average travel directions. One can see that many arrows point from *Muzdalifah* to (c) *Jamaraat Brige*, a large pedestrian bridge build to carry hundred thousands of pilgrims. In this area the catastrophe happened between the 23th and 24th of September. This is also indicated by the temporal peek at the timeline (e). To compare this extraordinary situation to an average time period, we load Twitter data from August, as a second dataset. We apply the same grid structure to this dataset, making it comparable to the *Hajj* data. While the August dataset shows a peek around *Kabaa* only, the *Hajj* data clearly has its density peaks also at the pilgrimage places (see Figure 1, white box).

### 5.2 Taxi Traffic in New York

Transportation data derived from public buses, taxis [24, 47], and bike sharing systems [48] is often used to better understand urban mobility patterns. To determine if microblog data could be used instead, we compare one week of Twitter movements in New York with one week of taxi data. In New York, Twitter volumes are extremely dense and we also can choose from different traffic data sets that are available for the public. The dataset we use here is from the NYC Taxi & Limousine Commission and was recorded in June 2015. We chose a 5% sample of the first week in June, leading to 140,000 taxi trips. From Twitter we load data of the same timeframe and spatial area, leading to 50,000 trajectories.

In Figure 9 one can see the individual edge splattings for both datasets and their diversity splat. The visualization on the left shows the taxi dataset that has certain peaks (red) at the airports Newark (4),

John F. Kennedy (3), and La Guardia (2) that surround the New York area. The derived statistics ( $\mu = 27.1, \sigma = 73.6$ ) align with the visual impression: in the center of Manhattan many taxis are in use while in more remote areas the density of the data is mostly low, leading to a high standard deviation. This area locates main attractions and businesses. The movements derived from Twitter (center) share some of these peaks. One can, however, see that the color transitions are not as smooth as in the left image. We hypothesize that this is because most microblog movements show tourist routes to main attractions such as the Empire State Building and the City Hall (5). The JFK (8) and the Newark (7) airports also stand out. Compared to the taxi dataset, the number of Twitter movements is less but they also cover the north-west area. This leads to a lower mean and standard deviation ( $\mu = 11.18, \sigma = 24.47$ ). In the comparison view, we subtracted both edge splatting images with a min-max normalization and applied a cubic power function (see Section 4.4). The result is shown in the right image. The aforementioned peaks clearly stand out. The routes to the Newark airport are quite similar, as is the area around JFK. While the taxi data is most dense in the inner city, we only retrieve movement information for the western regions from the tweets. A negative mean of  $-0.14$  indicates that Twitter is slightly more prominent regarding the whole area.

### 5.3 Global Flight Schedules

Flight schedule data is often used to investigate global mobility patterns. We are interested how movements extracted from Twitter conform to real world transportation on a global level. For this study we use flight schedule data from the OpenFlight project (<http://openflights.org/>) that was collected in August 2014 and includes 8,107 airports and 67,663 air routes. For each of the 11,452 routes, the data contains origin and destination. The connectivity of this dataset is very strong as many airports allow flying to many destinations. To reduce clutter, we choose our hierarchical aggregation approach. We create an adaptive grid structure (shown in Figure 4.3) to aggregate the flight into a hierarchical graph. This graph is visualized with the grid-based heatmaps, presented in Section 4.3. The upper left image of Figure 10 shows results for the USA for a selected aggregation depth. The level of aggregation can be changed with a slider. Movement strength is encoded by the width of the links connecting different cells. Cell color encodes the accumulated weight of outgoing links. As the second dataset we sample one month Twitter data from the USA, also collected in August 2014. We apply the same grid structure (see Section 4.3) to get a comparable graph for the Twitter movements and filter the outcome to flights only, leading to 2,003 movements (Figure 10, upper right). One can see that visually similar connected links and movement volumes (relatively similar width), exist in both results (flights and tweets). Large airports around Los Angeles (1) and New York (3) are slightly busier in the Twitter data (upper right), as indicated by the width of the links of each corresponding grid cell. In addition, the color of each cell is also comparable for both datasets, indicating the relative number of scheduled flight departures and outgoing Twitter movements is equivalent in most areas. However, around the area of Atlanta (2) the connectivity to other areas is stronger and the relative number

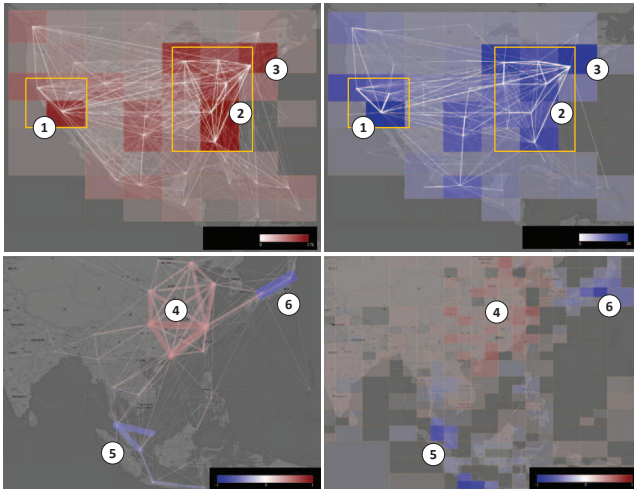


Figure 10: The upper two images show two graphs on the same hierarchy level: flight data (left) and Twitter data (right). Flight data is more prominent around Atlanta (2), Twitter data in California and New York (1, 3). The lower images show that flight data is being more prominent in the Chinese area (4). Twitter data has its peaks in Indonesia and Japan (5, 6). The left image shows connection differences while the right heatmap highlights different volumes of outgoing movement.

of flights is higher as in the microblog data (Atlanta airport is the world’s busiest airport). We also apply our comparison to eastern Asia. The lower images encode both datasets in a single view using a divergent color scheme for differences of outgoing links (left) and cells (gridded heatmap, right). For the aggregation level shown in the image, each cell contains on average 13.2 outgoing movements for the flight data and 3.1 for the Twitter data. The flight schedule data is more prominent in China (4), where Twitter is less popular (more people use Sina Weibo). Clear peaks in the Twitter data become visible in Indonesia, Malaysia (5), and Japan (6). Hovering over Malaysia highlights 10% of all flight related movements in the Twitter data, while in the flight schedule data, the relative amount is significantly lower (2%). Overall, this case study shows a high similarity for flight connections in the USA. In Asia similar connections could also be identified, but their travel volumes show significant differences.

## 6 DISCUSSION

The exploration and comparison of movements derived from social media is both, challenging and promising. Besides the three cases presented, we investigated other events and seasonal changes. For example, we visualized significant movement differences around the Mediterranean sea in the holiday season. Main events like exhibitions and conventions (ComicCon, CeBIT) show a clear difference in movement patterns of the surrounding area compared to other timespans. The same applies to sports events. To investigate movement volumes, directions, and origins could be especially interesting for event planning and marketing, but also for security purposes.

However, extraction and analysis of movement data from microblogs cover many challenges that we only partly address with our contributed techniques. One challenge that we identified is that movements derived from Twitter can be very unstructured in their directions. Because of their low resolution, they do not naturally follow any road network. While edge splatting nicely shows point-based movement distribution and allows for a visual comparison by image-based subtraction, it does not completely preserve angle information. This makes it especially difficult to see movement directions in areas where movement is diverse. Our proposed aggregation partly addresses this problem. We show a heatmap of different granularity. Based on an adaptive grid it visualizes the amount of movement volumes at their origins or destinations. We indicate main directions with arrow glyphs or spatially aggregate movement to links. One drawback of aggregation in general is information loss.

Another difficulty lies in the sparsity of the data (low resolution of trajectory points) and resulting uncertainty. The derived move-

ments contain origin and destination, but the precise route taken is unknown. Our aggregated visualization helps to correctly interpret the data by hiding individual trajectories. While a precision&recall slider improves the awareness of uncertainty, the visualizations do not yet take uncertainty into account. For example, one could handle more uncertain movement (e.g., long timespans between two messages) as less important and reduce its influence.

Lastly, highly varying data volumes in time and space and uneven data distribution with high peaks, for example, at famous places, are challenging. Varying volumes get even more present when comparing to other data sources. How is air travel comparable to a spatio-temporal sequence of geo-located text messages? A difficult question to answer because many details are unknown such as the number of people in the plane. Normalization can solve some of these problems. We apply a global (regarding the data loaded) min-max normalization to make data sources comparable in scale. We handle uneven distribution with interactive data smoothing (i.e., cubic root functions) that increases values in areas with lower densities and downgrades high peaks. We could not completely solve the problem of locally varying data volumes, e.g., by country. In our system, an analyst would simply narrow down the area of interest to a more local level and only load the data in this region to get a more local comparison. Another solution would be to apply a local normalization function, e.g., based on Twitter or population density in the neighborhood (cf. [6]). As a drawback, global differences may not be visible anymore.

Overall, we think the contributed techniques are a first main step towards a better understanding, and especially an enabler for comparison of movement data derived from microblogs.

## 7 CONCLUSION AND FUTURE WORK

We presented a system that embeds techniques for the exploration and comparison of movements derived from Twitter. In three case studies we showed the applicability of our approach for a seasonal comparison and a comparison to other global and local datasets.

In the future, we want to extend the adaptive rectangular grid to other shapes, e.g., to Voronoi cells or political boundaries, which take the underlying geographical structures into account. Link aggregation can be further supported with hierarchical edge bundling that also integrates link weights and direction [41]. A main task we identified is to compare data regarding temporal changes. We aim for a semi-automatic event and seasonal change detection. Detected changes could be encoded in a timeline to provide overview. While we focused on the exploration and comparison of historical data, our approach could also be applied to streaming data [14]. Lastly, we plan to connect the textual content to the movement patterns found.

While it is still questionable when movement data derived from microblog services can help in crisis situations, also other domains such as urban planning, event management, and security could benefit from additional perspectives and insights, especially in cases where no other data is available. We plan to carry out different expert studies to get input from these domains in the near future.

## ACKNOWLEDGEMENTS

This work was supported by the Horizon 2020 project *CIMPLEX*, grant no. 641191. and by the joint project Data-Driven Intelligent Transportation between China and Europe announced by the Ministry of Science and Technology of China, Zhejiang Provincial Natural Science Foundation (No. LR14F020002).

## REFERENCES

- [1] Hajj stampede near Mecca leaves over 700 dead. <http://www.nytimes.com/2015/09/25/world/middleeast/mecca-stampede.html>, New York Times, September, 2015.
- [2] Twitter company fact. <https://about.twitter.com/company>, June, 2015.
- [3] J. Abello and F. van Ham. Matrix Zoom: A visual interface to semi-external graphs. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 183–190, 2004.
- [4] K. Andrews, M. Wohlfahrt, and G. Wurzing. Visual graph comparison. In *Proceedings of the 13th Conference on Information Visualization*, pages 62–67. IEEE Computer Society, 2009.



- [5] N. Andrienko and G. Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, 12(1):3–24, 2012.
- [6] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. In *Proceedings of ACM Sigmod Record*, pages 49–60. ACM, 1999.
- [7] D. Archambault. Structural differences between two graphs through hierarchies. In *Proceedings of Graphics Interface*, pages 87–94. Canadian Information Processing Society, 2009.
- [8] D. Archambault, H. Purchase, and B. Pinaud. Difference map readability for dynamic graphs. In *Proceedings of Graph Drawing*, pages 50–61. Springer, 2011.
- [9] A. Barsky, T. Munzner, J. Gardy, and R. Kincaid. Cerebral: Visualizing multiple experimental conditions on a graph with biological context. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1253–1260, 2008.
- [10] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. A taxonomy and survey of dynamic graph visualization. *Computer Graphics Forum*, 2016.
- [11] F. Beck and S. Diehl. Visual comparison of software architectures. *Information Visualization*, 12(2):178–199, 2013.
- [12] F. Beck, R. Petkov, and S. Diehl. Visually exploring multi-dimensional code couplings. In *Proceedings of the International Workshop on Visualizing Software for Understanding & Analysis*, pages 1–8. IEEE, 2011.
- [13] J. I. Blanford, Z. Huang, A. Savelyev, and A. M. MacEachren. Geolocated tweets enhancing mobility maps and capturing cross-border movement. *PLoS one*, 10(6):e0129202, 2015.
- [14] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.
- [15] R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram. Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, 100(2):390–398, 2006.
- [16] M. Burch, C. Vehlou, F. Beck, S. Diehl, and D. Weiskopf. Parallel edge splatting for scalable dynamic graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2344–2353, 2011.
- [17] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):280–289, 2016.
- [18] J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert. Trajectory-based Visual Analytics for Anomalous Human Movement Analysis using Social Media. In *Proceedings of EuroVis Workshop on Visual Analytics*. The Eurographics Association, 2015.
- [19] S. Chen, C. Guo, X. Yuan, J. Zhang, and X. L. Zhang. MovementFinder: Visual analytics of origin-destination patterns from geo-tagged social media. In *Proceedings of Visual Analytics Science and Technology*, pages 239–240. IEEE, 2014.
- [20] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):270–279, 2016.
- [21] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.
- [22] A. Chua, E. Marcheggiani, L. Servillo, and A. V. Moere. FlowSampler: Visual analysis of urban flows in geolocated social media data. In *Proceedings of Social Informatics*, pages 5–17. Springer, 2015.
- [23] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, N. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.
- [24] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.
- [25] G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, and H. Stange. Tracing the german centennial flood in the stream of tweets: first lessons learned. In *Proceedings of the ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 31–38. ACM, 2013.
- [26] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.
- [27] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [28] F. D. Hobbs. *Traffic Planning and Engineering: Pergamon International Library of Science, Technology, Engineering and Social Studies*. Elsevier, 2013.
- [29] J. Kehrer, H. Piringer, W. Berger, and M. E. Gröller. A model for structure-based comparison of many categories in small-multiple displays. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2287–2296, 2013.
- [30] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 435–442. ACM, 2010.
- [31] Z. Liu, B. Jiang, and J. Heer. immens: Real-time visual querying of big data. *Computer Graphics Forum*, 32(3):421–430, 2013.
- [32] C. Ma, R. V. Kenyon, A. G. Forbes, T. Berger-Wolf, B. J. Slater, and D. A. Llano. Visualizing Dynamic Brain Networks Using an Animated Dual-Representation. In *Proceedings of Eurographics Conference on Visualization*, pages 73–77. The Eurographics Association, 2015.
- [33] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2: Geotwitter analytics support for situational awareness. In *Proceedings of Visual Analytics Science and Technology*, pages 181–190. IEEE, 2011.
- [34] A. Minetti. The three modes of terrestrial locomotion. *Biomechanics and biology of movement*, pages 67–78, 2000.
- [35] C. North and B. Shneiderman. Snap-together visualization: can users construct and operate coordinated visualizations? *International Journal of Human-Computer Studies*, 53(5):715–739, 2000.
- [36] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727, 2006.
- [37] C. Rogsch, W. Klingsch, A. Schadschneider, and M. Schreckenberg. *Pedestrian and evacuation dynamics 2008*. Springer, 2010.
- [38] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the international conference on World wide web*, pages 851–860. ACM, 2010.
- [39] R. Scheepens, N. Willems, H. van de Wetering, and J. J. Van Wijk. Interactive visualization of multivariate trajectory data with density maps. In *Pacific Visualization Symposium*, pages 147–154. IEEE, 2011.
- [40] A. Schipper, H. Fuhrmann, and R. von Hanxleden. Visual comparison of graphical models. In *Proceedings of the IEEE International Conference on Engineering of Complex Computer Systems*, pages 335–340. IEEE, 2009.
- [41] D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363, 2011.
- [42] H. Senaratne, A. Bröring, T. Schreck, and D. Lehle. Moving on Twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories. In *Proceedings of the ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 23–30. ACM, 2014.
- [43] D. Thom, H. Bosch, R. Krüger, and T. Ertl. Using large scale aggregated knowledge for social media location discovery. In *Proceedings of Hawaii International Conference on System Sciences*, pages 1464–1473. IEEE, 2014.
- [44] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2565–2574, 2012.
- [45] T. von Landesberger, F. Brodtkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. MobilityGraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):11–20, 2016.
- [46] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011.
- [47] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering. Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159–2168, 2013.
- [48] J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of london’s bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(4):239–251, 2011.