Constructing Semantic Interpretation of Routine and Anomalous Mobility Behaviors from Big Data

Georg Fuchs[1,3], Hendrik Stange[1], Dirk Hecker[1], Natalia Andrienko[1,2,3], Gennady Andrienko[1,2,3]
[1]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, St. Augustin, Germany
[2]City University London, London, UK
[3]Friedrich-Wilhelms University, Bonn, Germany

Abstract

Annually organized VAST Challenges provide a unique opportunity to analyze complex data with available ground truth. In 2014, one of the tasks was to interpret routine and anomalous patterns of human mobility based on big data: trajectories of cars and credit card transactions.

We describe a scalable visual analytics approach to solving this problem. Repeatedly visited personal and public places were extracted from trajectories by finding spatial clusters of stop points. Temporal patterns of people's presence in the places resulted from spatio-temporal aggregation of the data by the places and hourly intervals within the weekly cycle. Based on these patterns, we identified the meanings or purposes of the places: home, work, breakfast, lunch and dinner, etc. Meanings of some places could be refined using the credit card transaction data. By representing the place meanings as points on a 2D plane, we built an abstract semantic space and transformed the original trajectories to trajectories in the semantic space, i.e., performed ***semantic abstraction*** *of the data. Spatio-temporal aggregation of the transformed trajectories into flows between the semantic places and subsequent clustering of time intervals by the similarity of the flow situations allowed us to reveal and analyze the routine movement behaviors. To detect anomalies, we (a) investigated the visits to the places with unknown meanings, and (b) looked for unusual presence times or visit durations at different semantic places.*

The analysis is scalable since all tools and methods can be applied to much larger data. Moreover, the semantic data abstraction can serve as a tool for protecting the personal privacy.

1 Introduction

Huge amounts of data reflecting human mobility are constantly generated, including mobile phone use records, geographically referenced posts in social media (Twitter, Foursquare, Flickr, etc.), and GPS tracks. These data provide unprecedented opportunities for studying and understanding human mobility but require appropriate analysis methods, in particular, methods for semantic analysis that could infer and exploit meanings of places and purposes for attending places to enable understanding of people's everyday behaviors and life styles.

Combining multiple sources of mobility data is challenging. Despite traditional many V's of Big Data [10] (Volume, Velocity, Variety, Veracity, just to name a few), there exist specific complexities associated with the peculiarities of human mobility and corresponding data sets. Thus, different data sets have different structure, different quality, different spatial and temporal resolutions. Visual Analytics [12] creates opportunities for a synergy between human analyst and computer by providing appropriate visual interfaces to all stages of computational analysis, from data pre-processing and exploration to pattern search and model building. In the context of mobility analysis, visual analytics must address the specifics of space and time [3].

A common pattern of development in mobility analytics is the paradigm shift from syntactic [9] to semantic [11] analysis of movement data. Since mobility data by themselves are semantically poor, human interpretation, reasoning, and judgment are essential for giving sense and meaning to them. Purely computational methods only produce elementary results, e.g., trajectories with labeled segments. Often, semantic interpretation is based on a pre-defined set of places of interest (POI), also called areas of interest (AOI). This approach has limited applicability if POIs are not available or outdated. Moreover, POIs may be useful for identifying *public places* (visited by several people), but their applicability is limited for *personal places* (frequently visited by selected individuals). Additionally, the same POI may have different meanings for different people. For example, an apartment building may be a home place for its residents and a work place for service staff members.

In this paper we describe a visual analysis approach that facilitates human analyst-driven synthesis and semantic interpretation of human mobility behavior at different levels of abstraction, as appropriate for the analysis task at hand. It combines existing methods for the extraction, visual exploration and enrichment of raw mobility data with a novel concept of semantic spaces that allow analysis of routine as well as abnormal mobility behavior. We argue that the proposed transformation from geographic to semantic space creates new opportunities for analysis of mobility data – unlike existing mobility analysis approaches, semantic spaces allows to compare behaviors of a few to many individuals across different geographies and time periods.

This paper extends our contribution that received an award for outstanding scalable analysis [6] at VAST Challenge 2014 [8], mini-challenge 2. We demonstrate our approach to acquisition of semantically meaningful locations by combining two simulated data sources of that challenge, namely, trajectories of cars and credit card transactions. After extracting and interpreting personal and public places and assignment of the semantic labels to them, we transform the original trajectories into sequences of place visit records, each record containing the semantic label of the visited place and the start and end times of the visit. This transformation projects human mobility from *geographic* to *semantic space*. We demonstrate that the proposed transformation creates new opportunities for data analysis.

2 Problem Description and Example Data Sets

VAST challenges are open to participation by individuals and teams in industry, government, and academia. The challenge setup typically comprises several interrelated, large data sets together with a set of complex analysis tasks, to which the participants' submissions should showcase their visual analytics approach and provide well-founded answers. To accommodate the background story framing these analysis tasks, challenge data sets are either derived from real data with alterations, or generated artificially but with realistic artifacts such as missing values, precision limitations, and ambiguities just as could be expected from real use case data.

The 2014 IEEE VAST challenge's background story can be found on the archived challenge website [8]. The challenge consisted of three inter-related mini-challenges and an overall Grand Challenge. This paper presents our visual analysis approach targeting mini-challenge 2 (MC2), which involved geospatial, temporal, and transaction data analysis. In a nutshell, a company called GAStech provided company cars to its employees. Both personal and business uses were allowed. However, without the employees' knowledge, GAStech had installed trackers in the company vehicles. The devices periodically recorded the vehicles' geospatial positions when they were moving. The recorded tracks from a two week period were provided for the challenge. Additionally, credit and debit card transactions of the GAStech employees were available for the same period.

The GPS trajectories data set consists of 671,717 time-stamped positions of 40 distinct cars. The credit card transactions data set consists of 1,087 records of 35 individuals. Each record includes a card owner name, named location (but no geo-coordinates), date and time, and transaction amount. Both data sets include systematic and arbitrary mistakes such as shifted positions, wrong times, missing records etc.

The overall tasks for participants of MC2 was, first, to describe the routine behaviors of the GAStech employees, and, second, to identify suspicious patterns of behavior. The participants had to cope with uncertainties

that result from missing, conflicting, and imperfect data. The following sections briefly describe our approaches in addressing these questions.

3 Extraction and Interpretation of Places

We used an automated tool [7] that extracts repeatedly visited personal and public places by spatial clustering of points from trajectories. Relevant points can be previously selected by interactive filtering. We selected the points of stopping for at least one minute. The tool's work is based on finding groups of points fitting in circles with a chosen maximal radius and uniting close groups. We chose a sufficiently big maximal radius (100m) to account for the noise in the data. Personal places are extracted separately from the trajectory of each person. Public places visited by at least a given minimal number of distinct persons (two in our analysis) are extracted from all trajectories together. Figure 1 shows identified personal and public places.

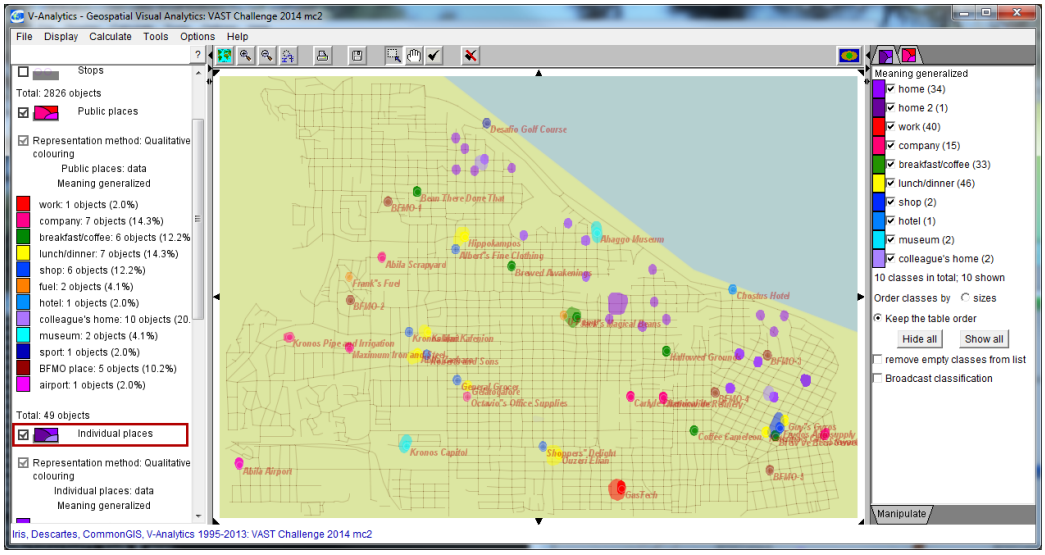

Figure 1: Identified personal and public places.

Through spatio-temporal aggregation of the trajectories [5], we obtained the visit counts for the extracted places by hourly intervals within the weekly time cycle. For the personal places, only the visits of the place owners were counted. We analyzed the temporal distributions of the place visits using 2D time histograms (Figure 2), where the rows correspond to 7 days the week, columns to 24 hours of the day, and marks in the cells represent aggregated visit counts for the whole set or subsets of places. By clustering the places according to similarity of their visit distributions, we found groups of places with prominent temporal patterns of visits, which could be attributed to certain categories of places or activities: home, work, breakfast or coffee, lunch, lunch and dinner. We combined card transaction data with the extracted stop points to (1) determine the geographic locations of the businesses and (2) among the places visited in the evenings and on the weekend, distinguish places for eating, shopping, sport, etc.

We could not determine the meanings of five public places visited mostly in hour 11 of the week days. The temporal pattern did not hint at any usual people's activity, and no card transaction records could be matched with the place visit times. We found that these places were attended by particular people and gave them a label "BFMO place", where BFMO consists of the initials of the last names of these people. These places and corresponding mobility patterns require further investigation.

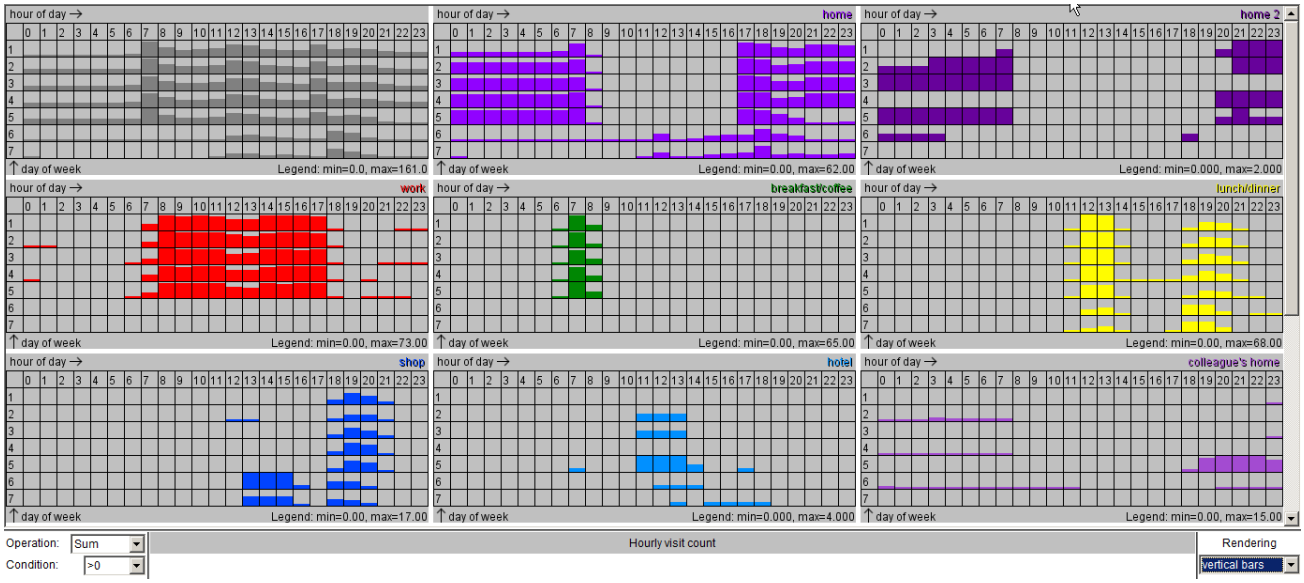

Figure 2: 2D time histograms represent the total counts of visits to different place categories by hourly intervals in the weekly cycle.

4 Analysis of Routine Behaviors in Semantic Space

After the assignment of the semantic labels to the places, we transformed the original trajectories into sequences of place visit records, each record containing the semantic label of the visited place and the start and end times of the visit. The intermediate trajectory points between the place visits were omitted.

We created an abstract semantic space where the semantic categories of places are represented as points on a 2D plane; we call them "semantic places". Then, we transformed the sequences of place visit records to trajectories in the semantic space. For this purpose, the place visit records were complemented with the coordinates of the semantic places (see Figure 3).

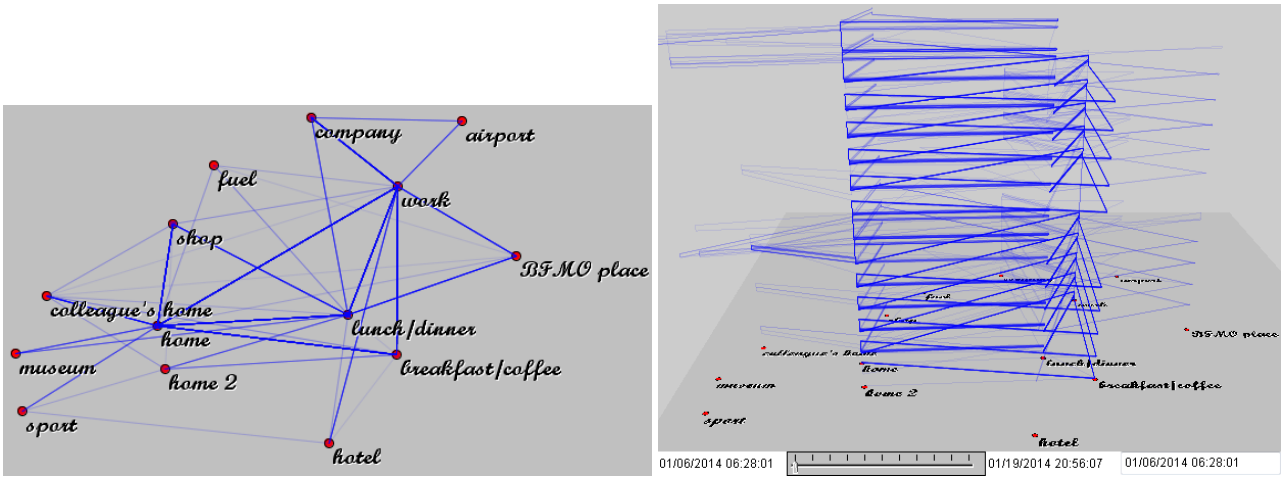

Figure 3: Trajectories in the semantic space: map (left) and space-time cube (right).

The data transformation in which real geographic coordinates are replaced by "locations" in an abstract semantic space is called *semantic abstraction*. Semantic abstraction is a tool for protecting personal location

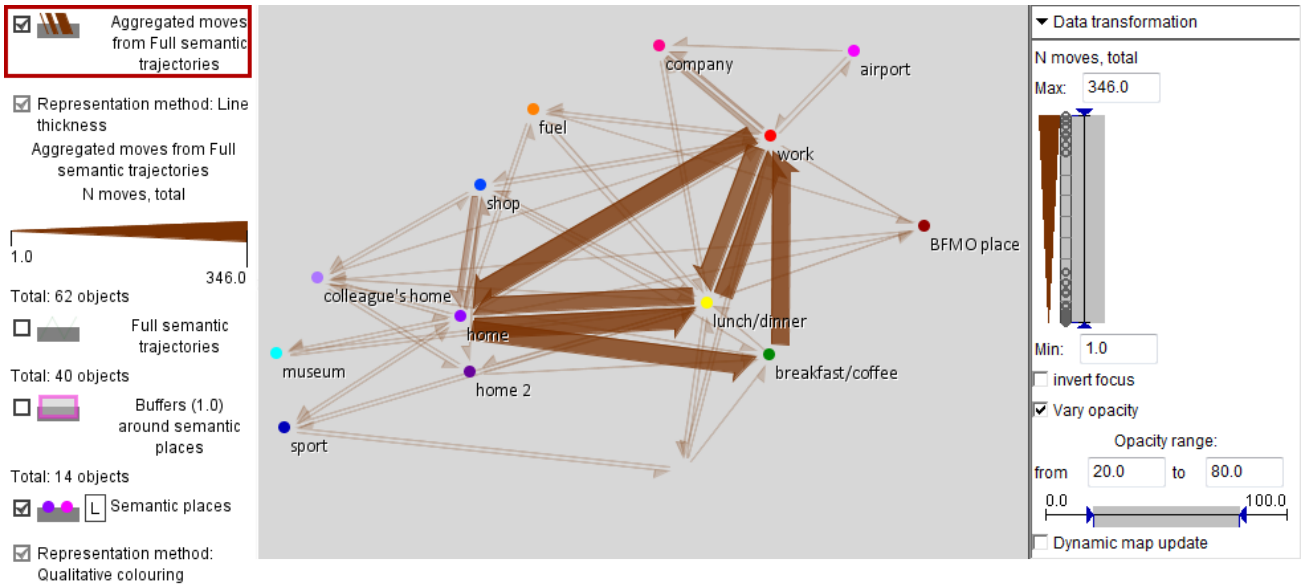

Figure 4: Summarized flows between semantic places. The widths of the flow symbols are proportional to the total counts of the moves between the respective types of places.

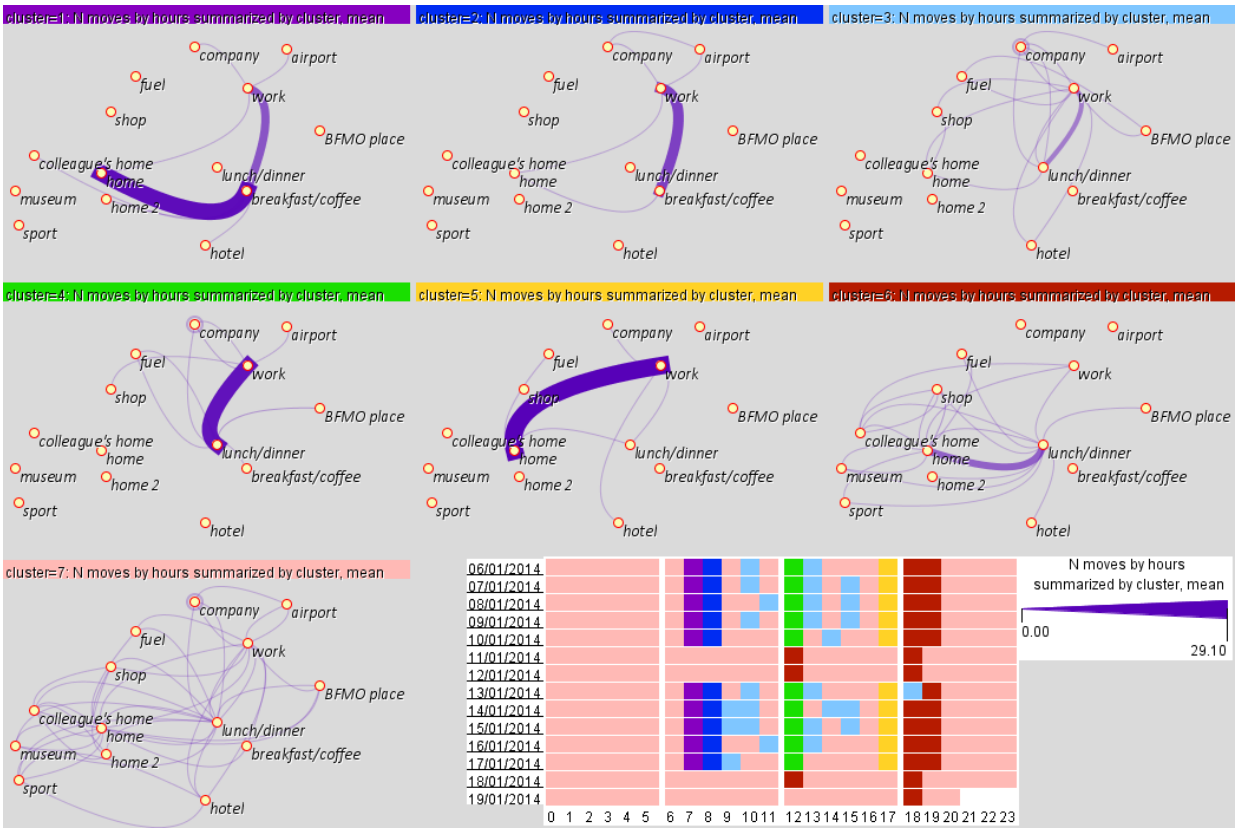

Figure 5: Clustering of hourly intervals by the similarity of the flows in the semantic space reveals high regularity of the movements.

privacy, since sensitive information about specific geographic locations visited by individuals is completely removed from the data.

Spatio-temporal aggregation can be applied to trajectories in abstract spaces in the same way as to trajectories in geographic space [2]. We aggregated the transformed trajectories into flows (aggregate moves) between the semantic places for the overall period and by hourly intervals. Figure 4 shows summarized flows between semantic places.

Then we clustered the intervals by similarity of the respective flow situations, i.e., vectors composed of the magnitudes of the flows for all ordered pairs of semantic places. We applied k-means clustering algorithm using Manhattan distance between the vectors as the similarity measure. In the calendar display (bottom center of Figure 5), pixels representing the hourly intervals are colored by their cluster membership. Periodic patterns with regard to the daily and weekly time cycles could be observed for different values of the parameter k (number of clusters). In the small multiple maps in Figure 5, the average hourly flows for the time clusters are represented by the widths of the flow symbols (curved lines with the curvature increasing towards the destination).

5 Detection and Analysis of Anomalies

By interacting with the map display of the semantic space and transformed trajectories, we selected the daily trajectories visiting BFMO places (cf. Section 3) as shown in Figure 6, and onserved that people went to these places from the work place. From BFMO places, the visitors almost always moved to lunch/dinner places and then returned to the work place. Hence, BFMO places were not visited for having lunch. By extracting corresponding place visit records we examined who was when in which place and found five cases when two or three people met in the same place. All four visitors of BFMO places were security employees.

For each semantic place, we analyzed the temporal distribution of the visits using a 2D time histogram similar to those shown in Figure 2 but with 14 rows corresponding to the consecutive days of the two week period of the data (as in the calendar display in Figure 5). We paid attention to place visits in unusual times. For each such case, we extracted (by spatio-temporal filtering) the place visit records including the visitors' names, exact visit times, and place names. In this way, we detected that four persons sometimes attended the homes of their colleagues in night time. We also detected night visits to work and some other anomalies.

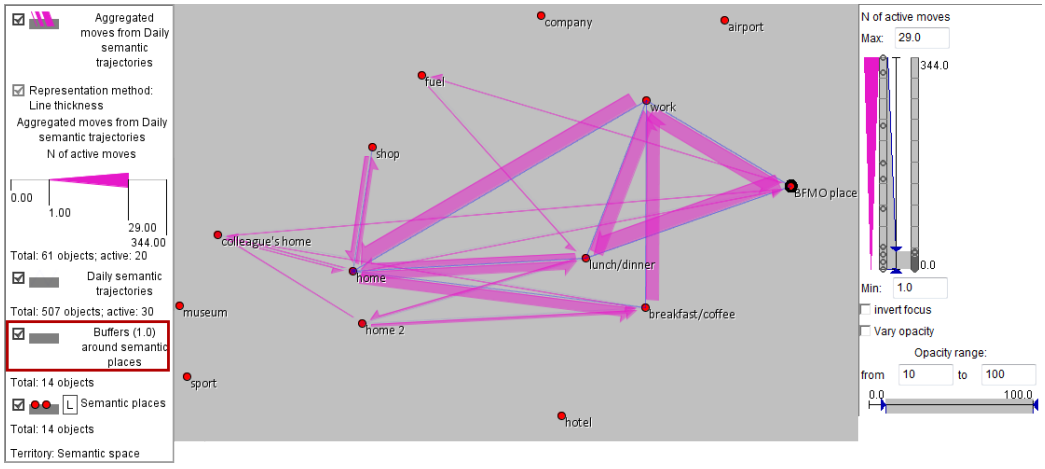

Figure 6: By filtering through the semantic space map, we have selected only those daily trajectories that include visits to BFMO places. There are 30 such trajectories, which are also shown in a summarized form as a set of flows. The map shows that the visitors typically went to the BFMO places from the work and after that went for lunch, which means that the BFMO places are not lunch places.

6 Social Network Analysis

From the trajectories, we have extracted all meetings of the people and excluded the meetings that occurred at work and the meetings of people living together at their homes. From the remaining meetings, we have computed distances between individuals based on the relative frequencies of their meetings.

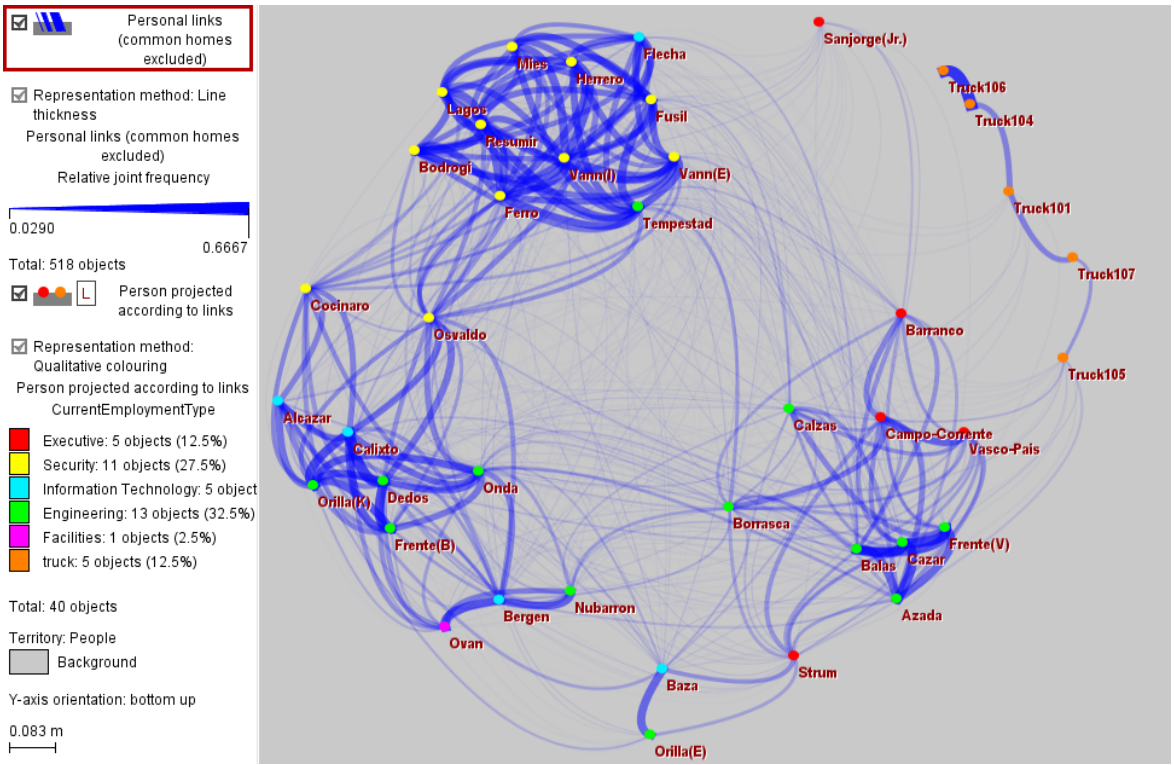

Figure 7: The space of inter-personal relationships.

The map display (Figure 7) shows the space of inter-personal relationships. The 2D projection has been obtained based on the pair-wise distances between the individuals. The dots represent the individuals and are colored according to their employment types. The curved connecting lines represent the strengths of the relationships between the individuals (i.e., the relative meeting frequencies) by proportional widths and opacities. We see a tight group of security employees (the group also includes two non-security persons). Two security persons, Cocinaro and Osvaldo, bridge this group with another tight group made by engineering and information technology employees. Another group of engineers is relatively separated from the latter group and from the security group but has strong links to executive staff.

7 Conclusion

All methods and tools that we used for our analysis are scalable with regard to the number of individuals, number of places, and length of the time period covered by the data. We mostly used aggregated views, which could also be applied to much larger data. Detailed data (place visit records) were accessed only for analyzing anomalies. The analysis greatly relied on computational data processing: stop and place extraction, data aggregation, and clustering. These operations are also scalable. Apart from the examination of the anomalies, the analysis was done in a way respecting personal privacy, without accessing personal data. Our approach demonstrates that

Visual Analytics may be not only harmful for personal privacy [4], but also has potential to create opportunities for privacy-preserving analysis of human mobility [1].

The page limit does not allow us to provide here a detailed bibliography on known methods and approaches to place extraction from movement data and to describe in detail all algorithms used in our approach. Interested readers are pointed at our recently published paper [7] that includes a comprehensive bibliography and provides all necessary algorithmic details. In that paper, the approach was used for reconstructing mobility patterns of residents of San-Diego based on geo-located twitter messages and land use map layers.

References

[1] G. Andrienko and N. Andrienko. Privacy Issues in Geospatial Visual Analytics. In G. Gartner and F. Ortag, editors, *Advances in Location-Based Services*, Lecture Notes in Geoinformation and Cartography, pages 239–246. Springer Berlin Heidelberg, 2012.

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer Verlag, 2013.

[3] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600, 2010.

[4] G. Andrienko, N. Andrienko, D. Keim, A. M. MacEachren, and S. Wrobel. Challenging problems of geospatial visual analytics. *Journal of Visual Languages & Computing*, 22(4):251–256, 2011.

[5] N. Andrienko and G. Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):205–19, 2011.

[6] N. Andrienko, G. Andrienko, and G. Fuchs. Analysis of Mobility Behaviors in Geographic and Semantic Spaces. In *VAST Challenge @ IEEE VAST 2014*, 2014. Award for Outstanding Scalable Analysis.

[7] N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski. Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces. *Information Visualization*, ?(?):???, 2015. (accepted).

[8] K. Cook, G. Grinstein, and M. Whiting. IEEE VAST Challenge 2014 – The Kronos Incident. http://vacommunity.org/VAST+Challenge+2014, November 2014. Last accessed 2015-03-29.

[9] F. Giannotti and D. Pedreschi. *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media, 2008.

[10] A. McAfee and E. Brynjolfsson. Big data: the management revolution. *Harvard business review*, 90:60–68, October 2012.

[11] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Computing Surveys*, 45(4):42, 2013.

[12] J. J. Thomas and K. A. Cook. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.