Human-in-the-Loop Integration of Complex and Noisy Data

Daniël M. Bot[1,*], Jannes Peeters[1,*], Danai Kafetzaki[1,2], Jan Aerts[1]
jelmer.bot@uhasselt.be, jannes.peeters@uhasselt.be, danai.kafetzaki@uhasselt.be, jan.aerts@uhasselt.be
1: I-Biostat, Data Science Institute, Hasselt University, Haseslt, Belgium. 2: Sagalassos Archaeological Research Project, KU Leuven, Leuven, Belgium.
*: These authors contributed equally to this work.

Mini Challenge 2

In mini-challenge 2†, contestants were asked to analyse transaction and GPS tracking data to support law enforcement in investigating the disappearance of GASTech employees. This poster presents our methodology for solving this mini challenge.

1. Transactions data

The challenge contained two types of transactions: (1) credit card transactions accurate to the minute and (2) loyalty card transactions accurate to the day. There was no direct relation between loyalty cards and credit cards, so we used two metrics to match them: (1) The correlation between vectors containing the total amount of money spent by a card at each business on each day. (2) The Jaccard index between the credit cards' and loyalty cards' transactions, considering the date, business, and price of transactions. Matches were reviewed manually using a bipartite graph indicating the metric values (Fig. 1) and a detail view of the individual transactions (Fig. 2).

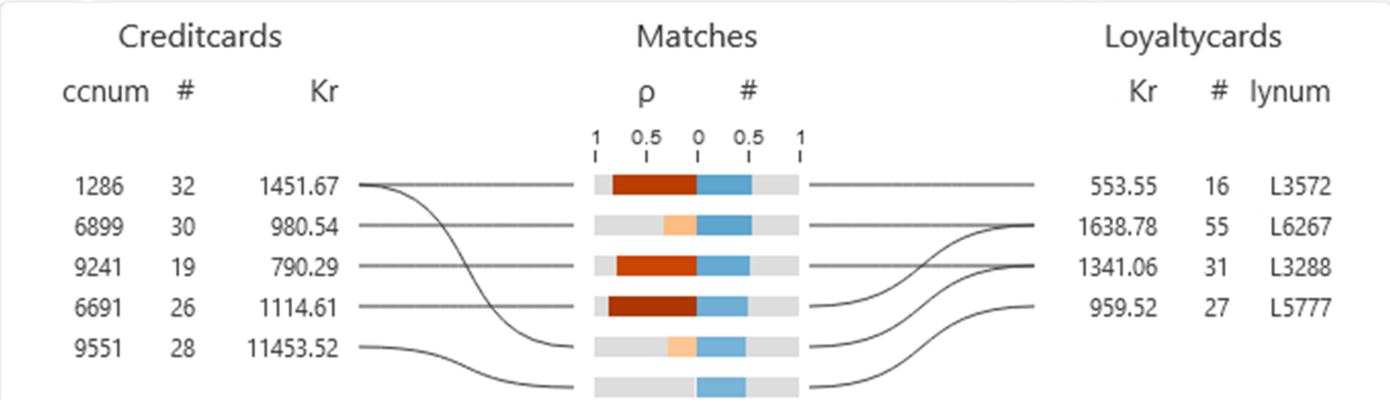

Figure 1: Partial bipartite graph with credit card to loyalty card matches, indicating the correlation and Jaccard Index metrics.

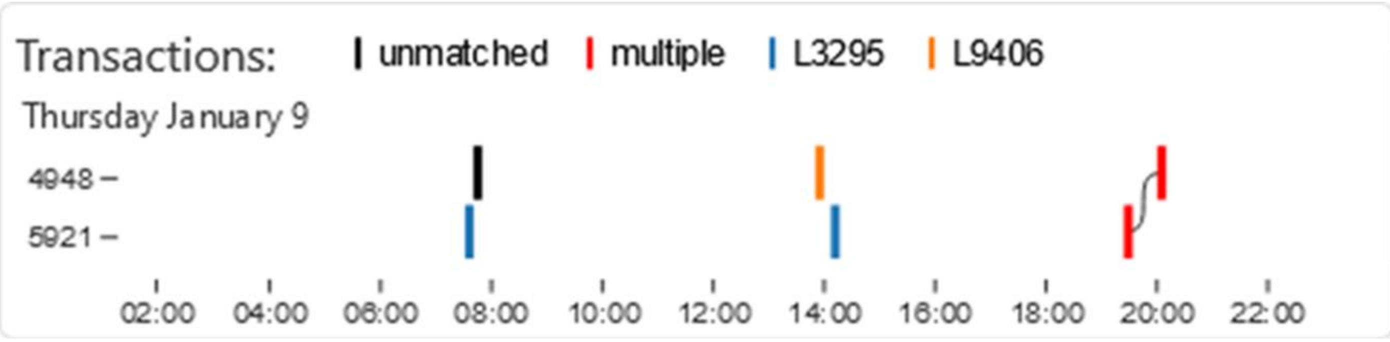

Figure 2: Partial transaction view, showing individual credit card transactions over time coloured by the loyalty card with matching transactions. Conflicts are shown in red.

2. GPS trace pre-processing

The GPS data contained traces of car positions. We looked at the periods in between subsequent GPS samples. A simple decision tree was used to classify these time periods as moving (duration < 30s and average speed > 20km/h), or stationary (not moving and distance travelled < 220m), or missing (otherwise). Several cars were found to have periods of missing GPS data (Fig. 3).

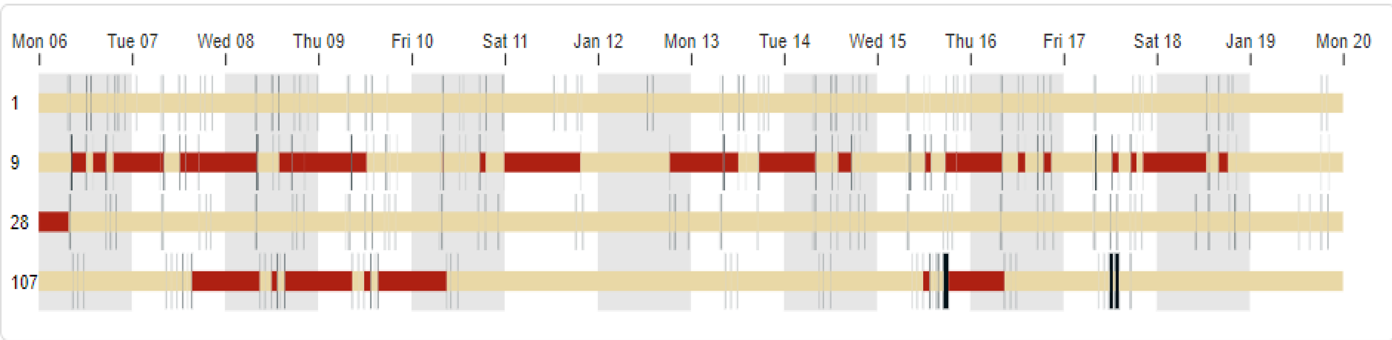

Figure 3: Timelines of GPS activity for cars 1, 9, 28 and 107. Motion is shown in black, stationary periods in yellow, and periods of missing GPS data in red.

Then, we manually annotated *points of interest* (POI) on a map (Fig. 4), marking all positions where cars had been stationary. The POIs were given a type based on the average number of cars present throughout a day.

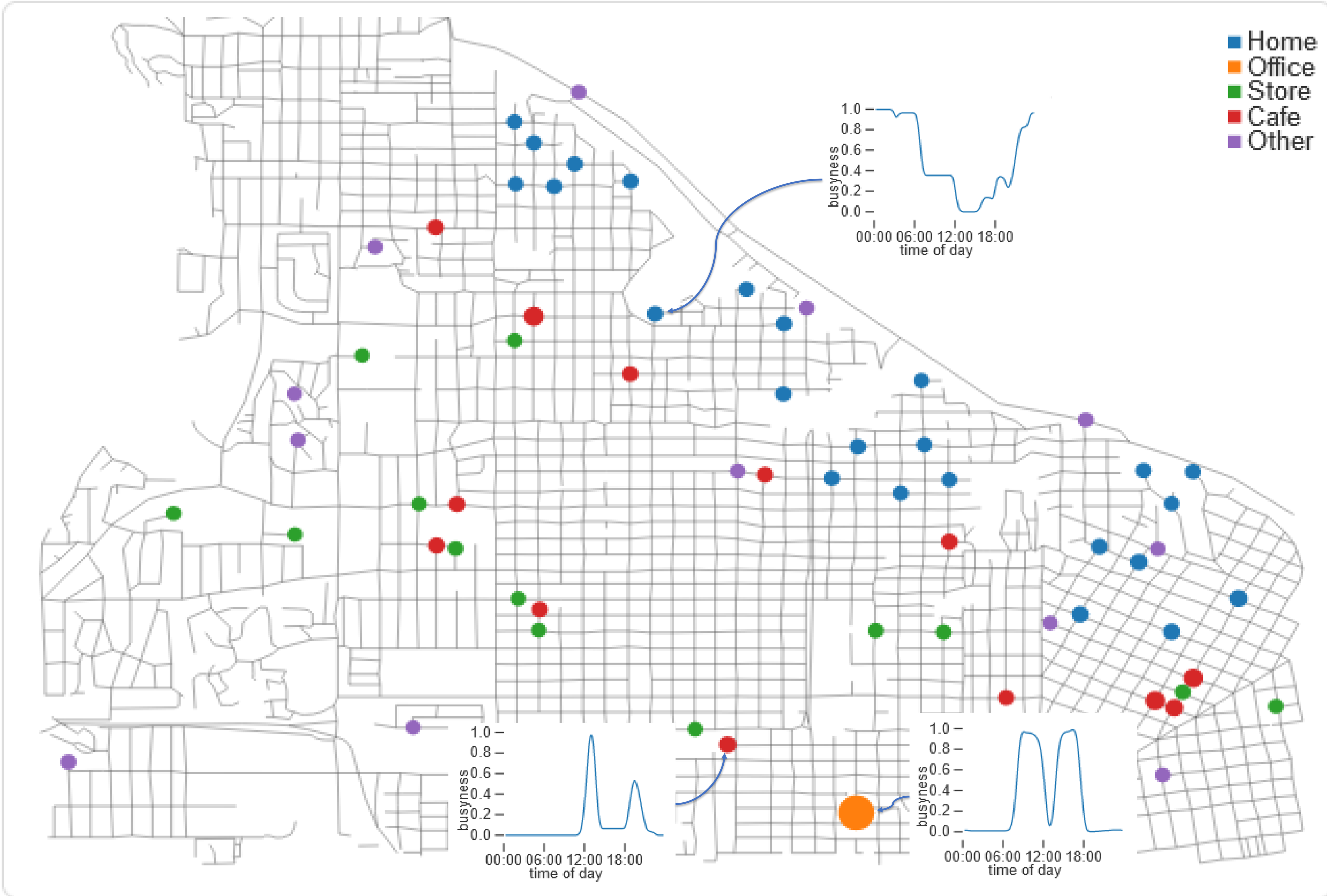

Figure 4: Street map of Abila with *points of interest* (POI) marking all positions where cars have been stationary. POIs are coloured by type, based on the average number of cars present throughout a day, the matching locations on the tourist map, and the business with transactions matched to the POI.

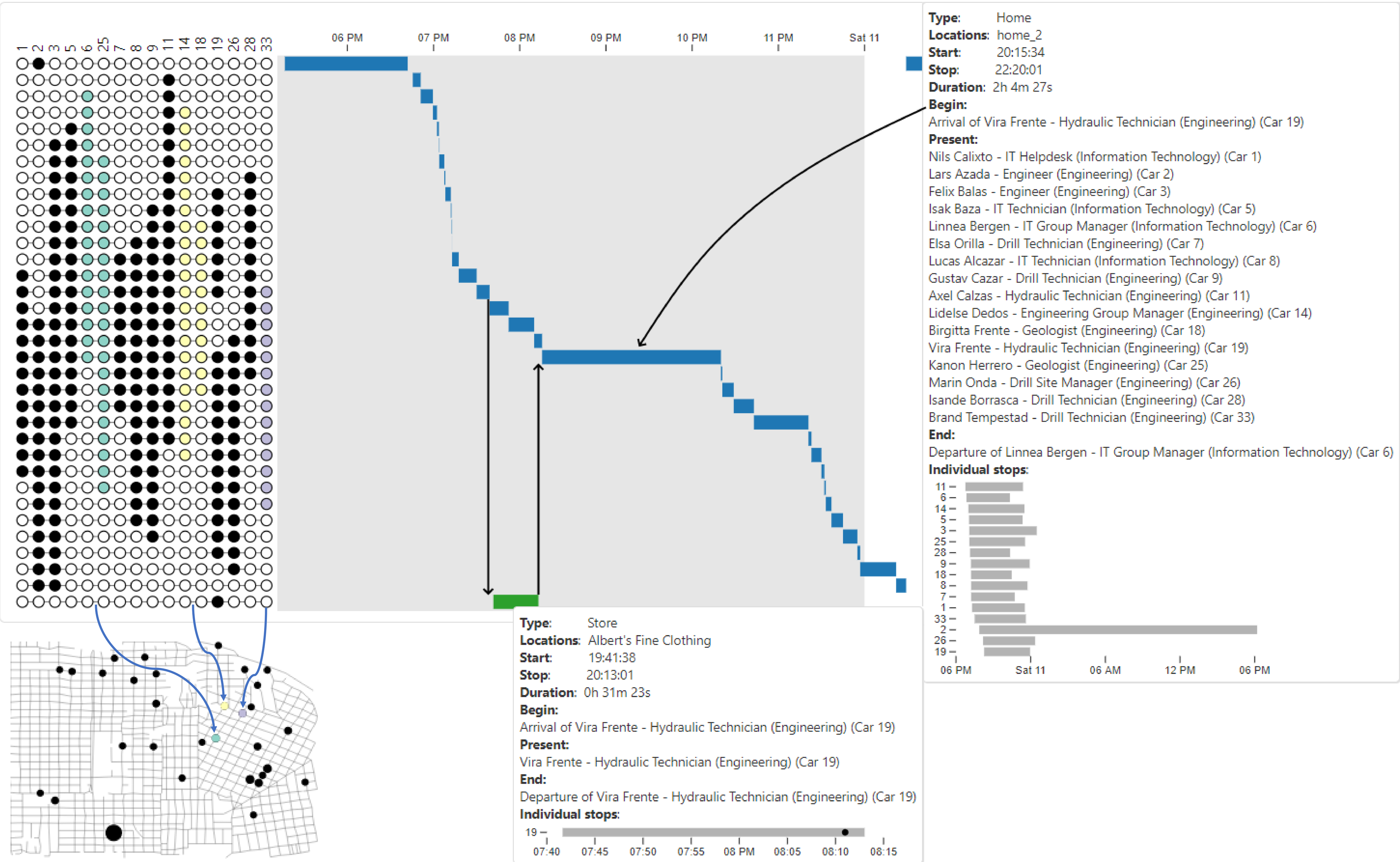

Figure 5: UpSet visualization (Lex et al. 2014) showing a (surprise) party for Lars Azada on Friday the 10th. On the left, each column of circles represents a car. The circles are filled in when the car is present in a row's unique combination of cars. Circles are coloured by household for households with multiple employees. On the right, a timeline shows when meetings occured for each row, coloured by the POI type. Tooltips indicate the people present and the events that started and ended the meeting. In addition, the full stationary periods of the individual cars are shown, with their matched transactions.

3. Matching cars to transactions

We estimated how consistently a credit card's transactions occured while a car was stationary using two metrics: (1) the correlation between binary vectors indicating whether a car was stationary or a transaction occured in 5-minute segments. (2) the precision and recall of a car's stationary periods as predictor for a credit card's transaction, where the transaction had to occur while the car was stationary and the transaction's business had to match the car's GPS position. Both metrics ignored stationary periods at homes and the office, as we did not expect transactions to occur there. Initially, we did not yet know where each business is located; these constraints were introduced as transactions were manually matched to stationary periods. For this task, we showed the matches as a bipartite graph (similar to Fig. 1) and provided detail on the individual stationary periods and transactions (Fig. 6).

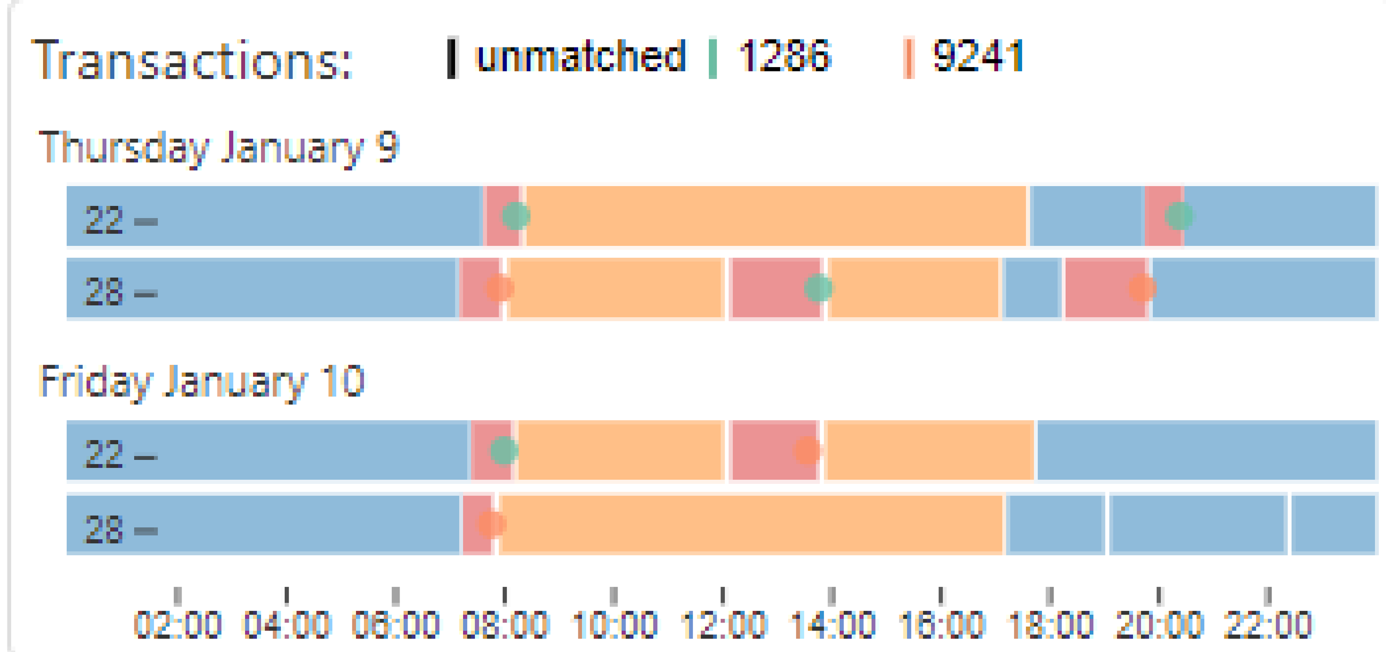

Figure 6. Partial view showing stationary periods over time as rectangles coloured by POI type and transactions coloured by credit card.

4. Meetings and suspicous behaviour

Meetings between employees were analysed using an UpSet visualization (Lex et al. 2014) to investigate suspicious behaviour and to find unofficial relationships (Fig. 5). The visualization shows a grid of circles, where columns correspond to cars and rows to a unique combination of cars present simultaneously. A timeline indicated the type of location at which meetings occurred. Using this visualisation we identified five households where multiple employees cohabit, a potential relationship between Hennie Osvaldo and Birgitta Frente, a potential relationship between Elsa Orilla and Brand Tempestad, a (surprise) party for Lars Azada on Friday the 10th, and a suspicious pattern of security personnel guarding executives' homes at night.

Acknowledgements

This work was supported in part by Hasselt University BOF grant ADMIRE [BOF21GP17], Flemish Government programme "Onderzoeksprogramma Artificiele Intelligentie (AI)", the Research Fund of the University of Leuven [C14/17/025] and the Research Foundation Flanders [G088319N].

References

A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics, 20(12): 1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248

†: https://vast-challenge.github.io/2021/MC2.html