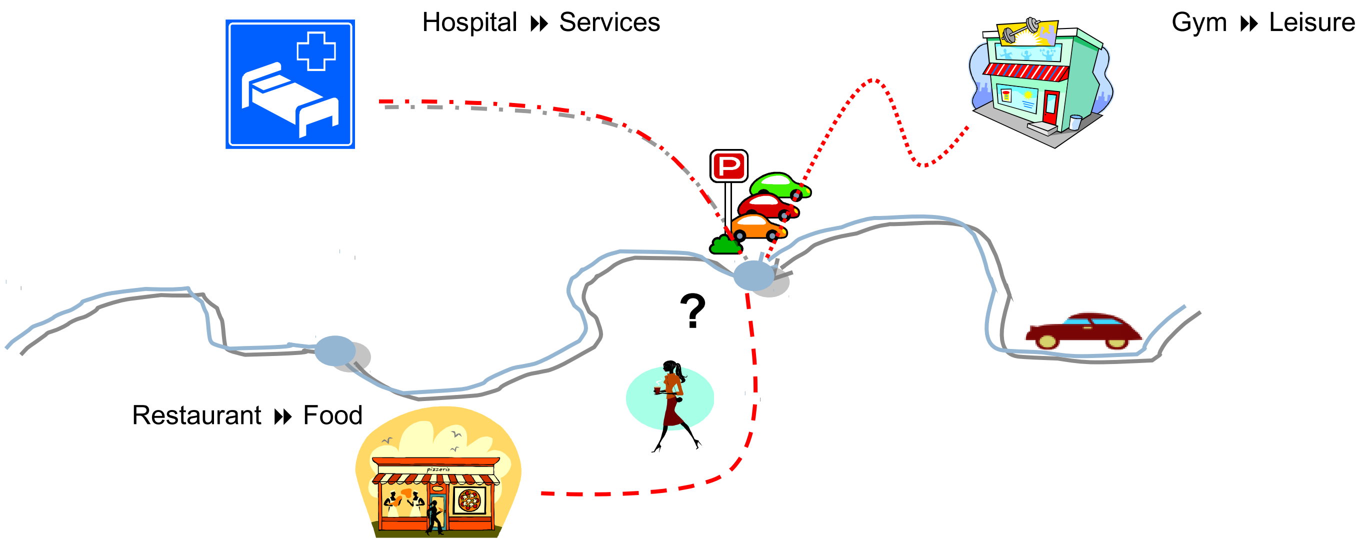
Classification in Mobility Data Mining

Activity Recognition – Semantic Enrichment

Recognition through Points-of-Interest

Given a dataset of GPS tracks of private vehicles, we annotate trajectories with the most probable activities performed by the user.

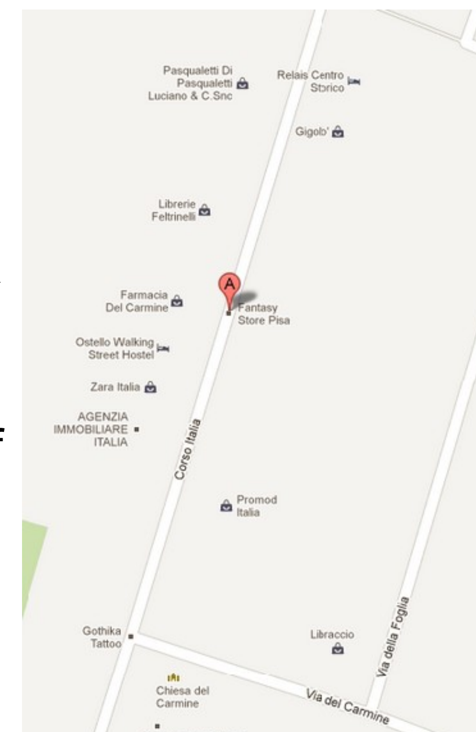


The method associates the list of possible <u>POIs</u> (with corresponding probabilities) visited by a user moving by car when he stops.

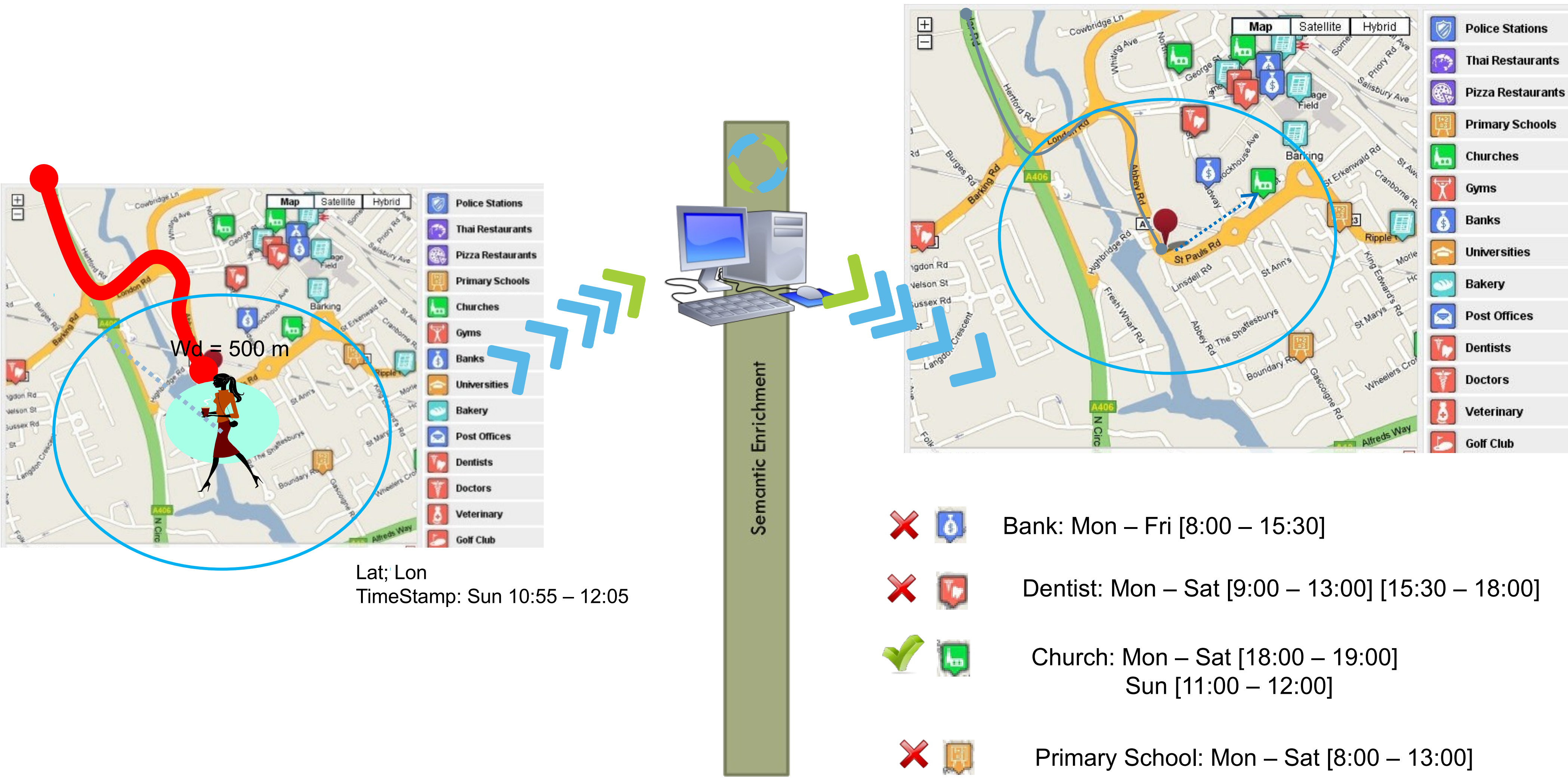
A mapping between POIs categories and Transportation Engineering activities is necessary.

The enrichment process

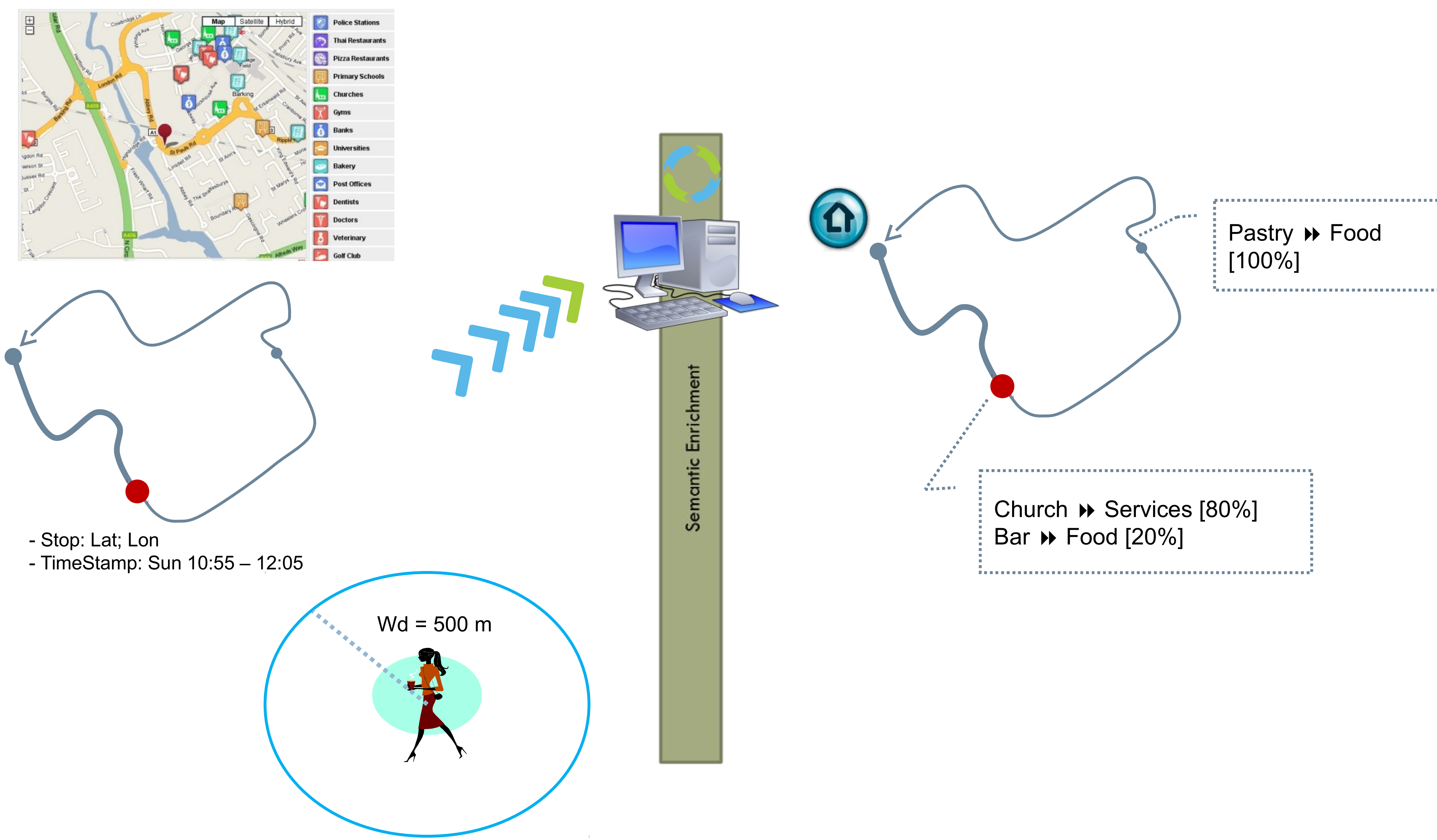
- **POI collection**: Collected in an automatic way, e.g. from Google Places.
- **Association POI – Activity**: Each POI is associated to a ``activity". For example Restaurant → Eating/Food, Library → Education, etc.
- **Basic elements/characteristics:**
  - C(POI) = {category, opening hour, location}
  - C(Trajectory) = {duration of the stop, stop location, time of the day}
  - C(User) = {max walking distance}
- **Computation of the probability to visit a POI/ to make an activity:** For each POI, the probability of ``being visited" is a function of the POI, the trajectory and the user features.
- **Annotated trajectory**: The list of possible activities is then associated to a Stop based on the corresponding probability of visiting POIs



Input & Output



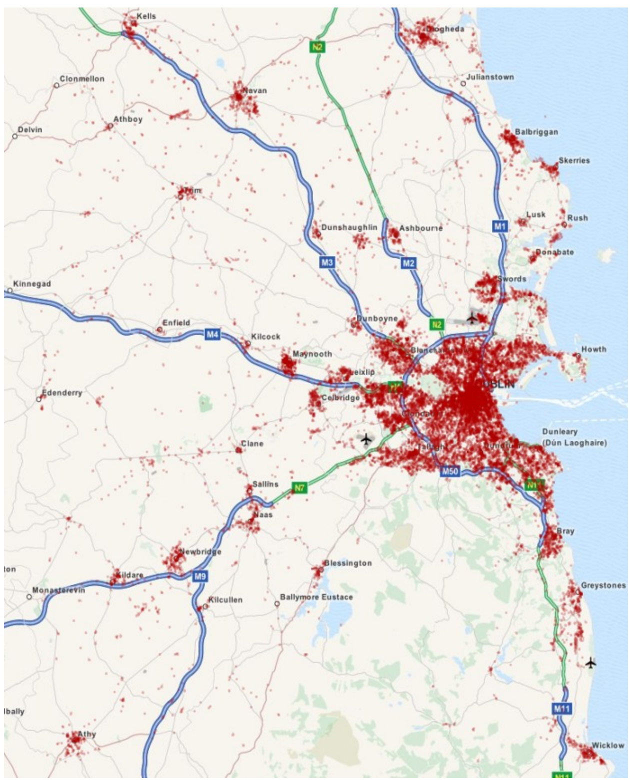
Input & Output



Inferring Activities from social data

Extraction of personal places from Twitter trajectories in Dublin area

The points of each trajectory taken separately were grouped into spatial clusters of maximal radius 150m. For groups with at least 5 points, convex hulls have been built and spatial buffers of small width (5m) around them.
1,461,582 points belong to the clusters (89% of 1,637,346); 24,935 personal places have been extracted.



Examples of extracted places

N	N?	min	q1	med	q3	max	ave	stdd
5180	0	1.0	2.0	3.0	5.0	120.0	4.8	6.3

Statistical distribution of the number of places per person

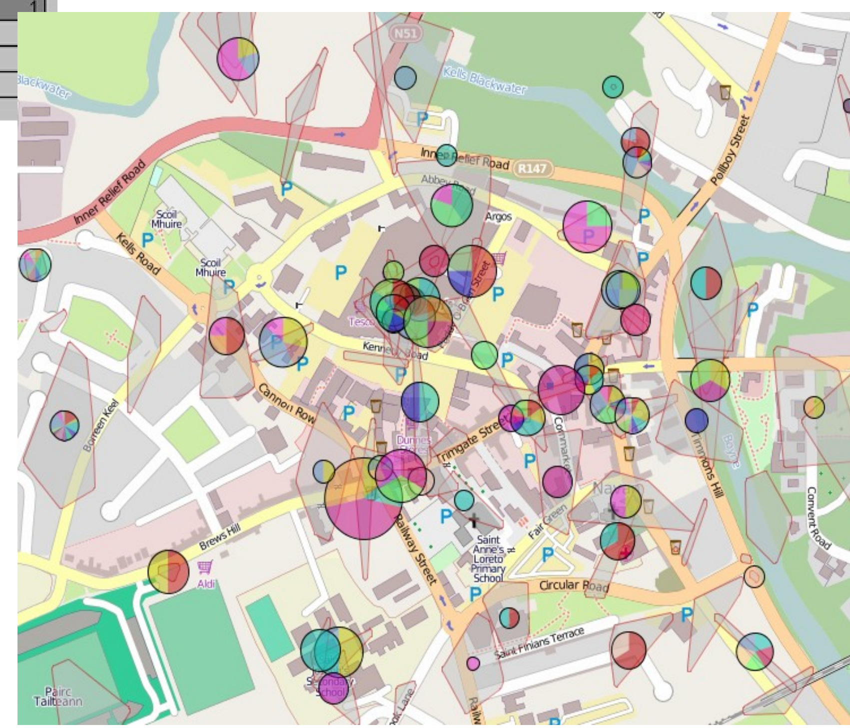
Recognition of the message topics, generation of topical feature vectors, and summarization by the personal places

Topics have been assigned to 208,391 messages (14.3% of the 1,461,582 points belonging to the personal places)

Message	Features	topic=family: Occurrences of topic	topic=home: Occurrences of topic	topic=education: Occurrences of topic	topic=work: Occurrences of topic	t...
@joe_lennon I usually	education	0	0	1	0	
@joe_lennon together	education	0	0	1	0	
@jas_103 deadly; don	work	0	0	0	1	
Just got home and see	home	0	1	0	0	
So excited about my ne	sweets	0	0	0	0	
@lamtcdizzy I haven't b	shopping	0	0	0	0	
Get in from my night ou	family;home;work	1	1	0	1	
Home again at 6pm! N	home	0	1	0		
Bussing it home for th			1	0		
Ah shite. It's been a p			0	0		
@ronanhutchinson be	education	0	0	1		

Get in from my night out; my dad gets home from work two minutes later. Great timing :)

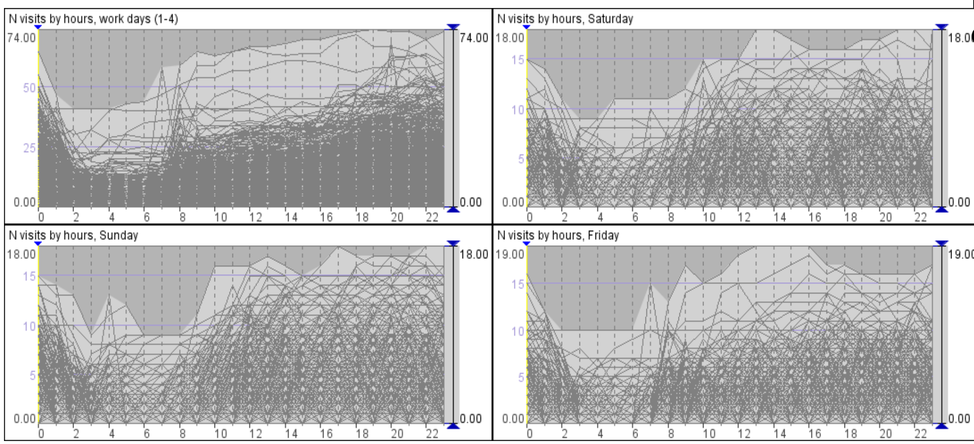
...



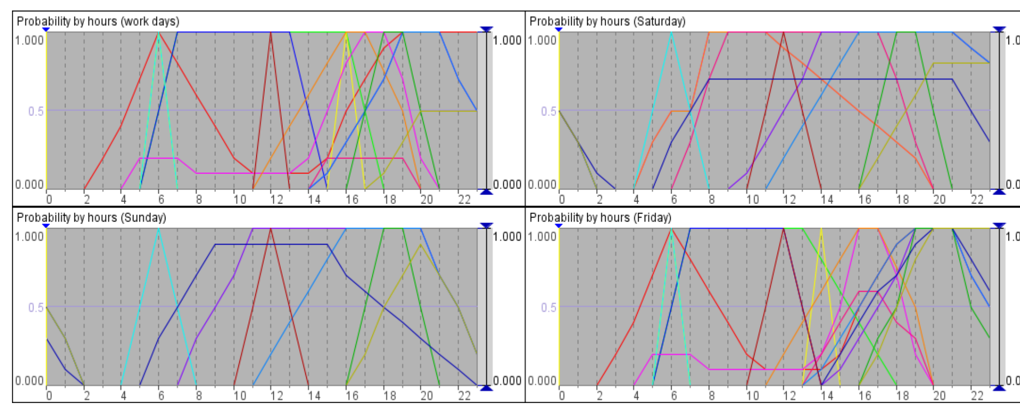
1) Some places did not get topic summaries (about 20% of the places)
2) In many places the topics are very much mixed
3) The topics are not necessarily representative of the place type (e.g., topics near a supermarket: family, education, work, cafe, shopping, services, health care, friends, game, private event, food, sweets, coffee)

Obtaining daily time series of place visits and comparison with exemplary temporal profiles

The daily time series of place visits have been obtained through aggregation of daily trajectories using only relevant places for each trajectory. The aggregation was done separately for the work days from Monday to Thursday, and for Saturday, Sunday, and Friday.

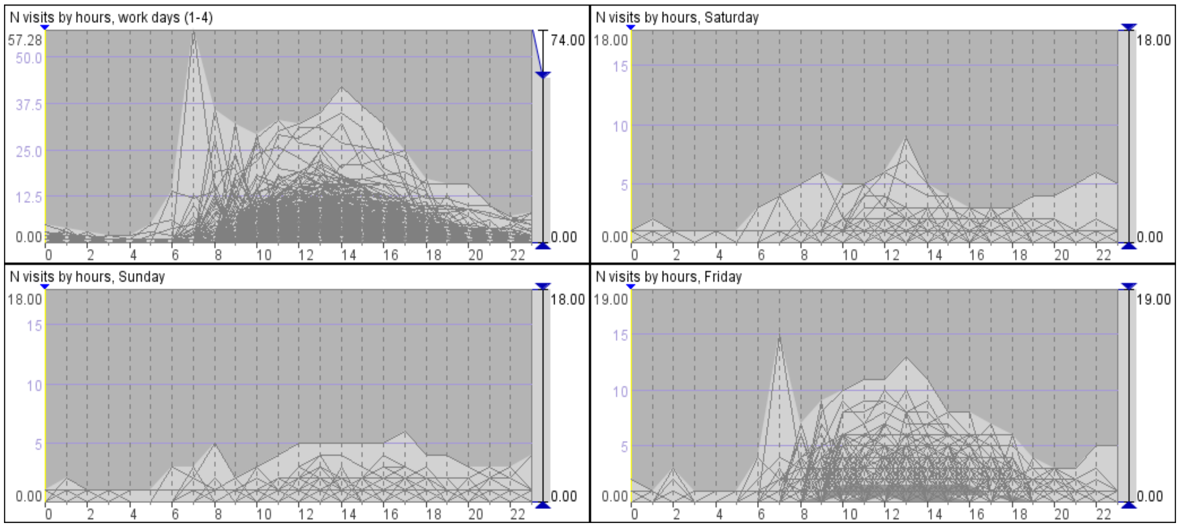


Exemplary temporal profiles of different activities or visits to different types of places



The time series of place visits are compared to the exemplary time profiles by means of the Dynamic Time Warping (DTW) distance function. Resulting scores: from 0 (no similarity) to 1 (very high similarity).
15,950 places (64% of all) have no similarity to any of the exemplary time patterns. 4,732 places (19%) have the maximal similarity score of 0.8 or higher; 4,179 of them (16.8% of all) were visited in 6 or more days.

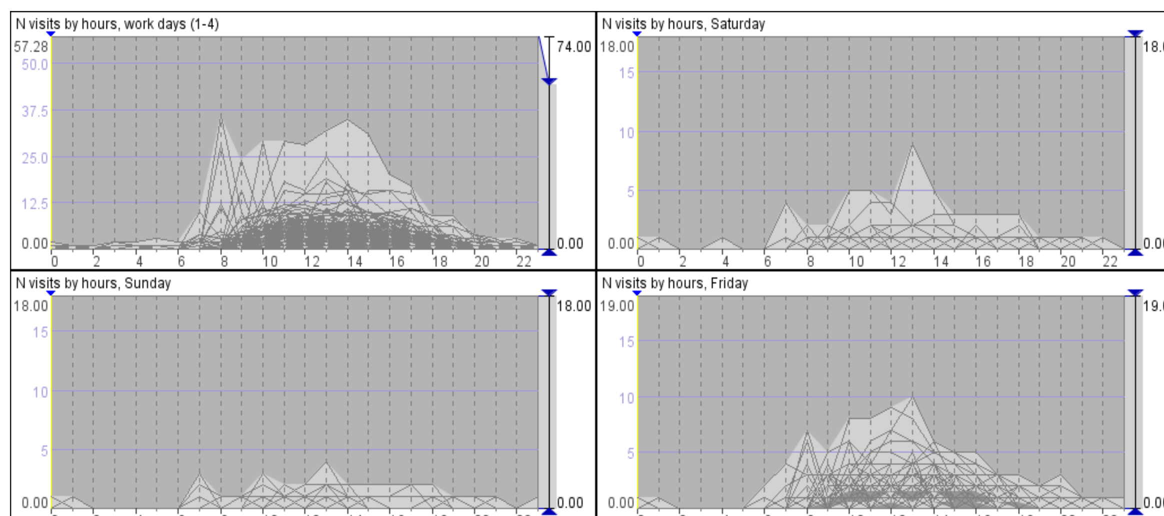
Time series with high similarity to “work” (>=0.8)



The time series similarity scores have been combined with the relative frequencies of the topics using a combination matrix

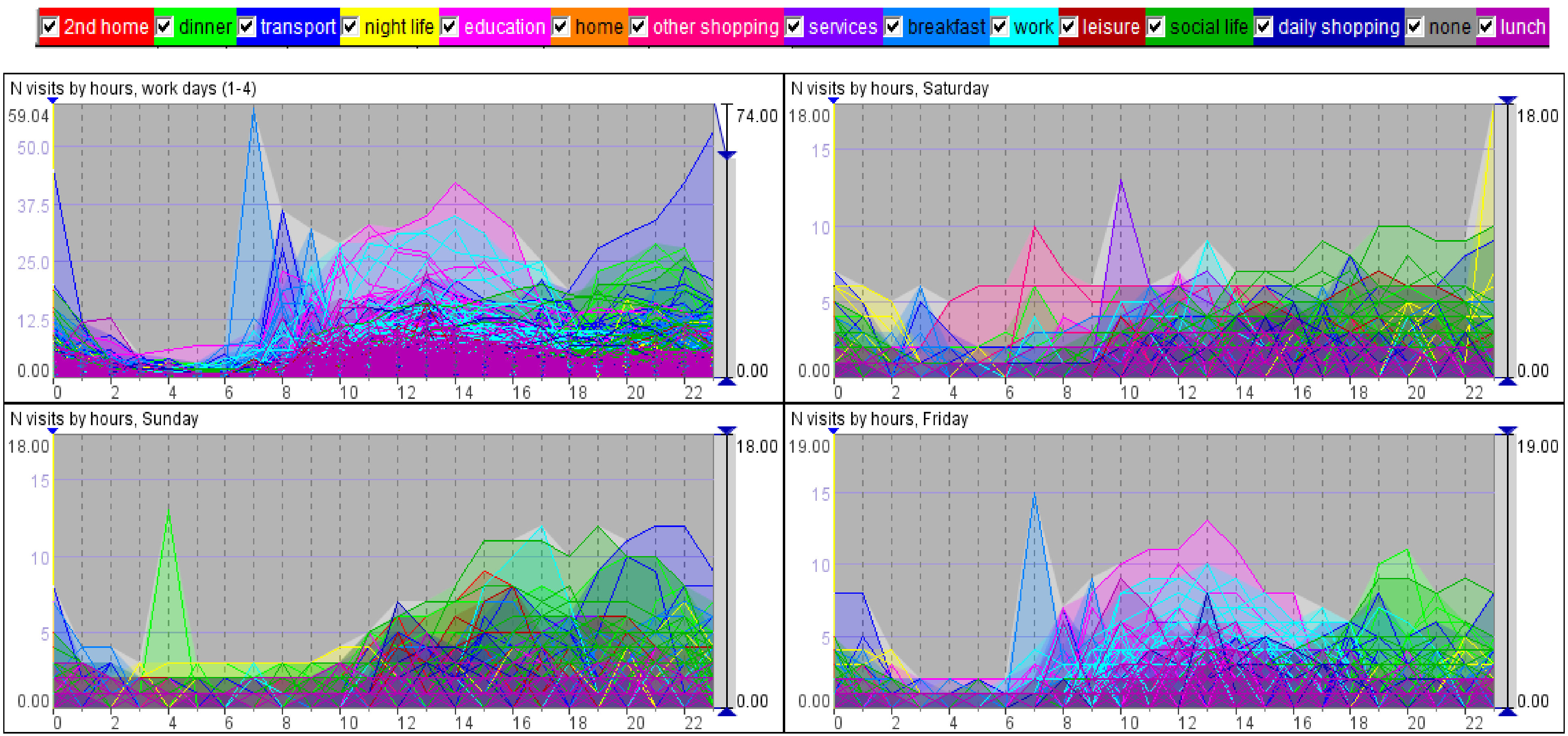
	home	work	education	transport	daily shopping	services	other shopping	leisure	social life	breakfast	lunch	dinner	2nd home	night life
family	1	0	0	0	0	0	0	0	1	0	0	0	1	-0.5
home	1	0	0	0	0	0	0	0	0	0	0	0	1	-0.5
education	0	0	1	1	0	0	0	-1	-1	-1	-1	-1	0	-0.5
work	0	1	0	1	0	0	0	-1	-1	-1	-1	-1	0	-0.5
local transport	0	0	0	1	0	0	0	0	0	-1	-1	-1	0	-0.5
regional transport	0	0	0	1	0	0	0	0	0	-1	-1	-1	0	-0.5
far transport	-1	-1	-1	1	-1	-1	-1	0	0	-1	-1	-1	0	-0.5
eatery	-1	-1	-1	-1	-1	-1	-1	0	0	0	1	0.5	-1	0
restaurant	-1	-1	-1	-1	-1	-1	-1	0	0.5	-1	0.6	1	-1	0.5
pub	-1	-1	-1	-1	-1	-1	-1	0.5	0.5	-1	-1	0.7	-1	1
cafe	-1	-1	-1	-1	-1	-1	-1	0.5	0.5	0.3	0.3	0.3	-1	0.2
shopping	-1	-1	-1	-0.5	1	0	1	0	0	0	0	0	-1	-0.5
daily shopping	-1	-1	-1	-0.5	1	0	0	0	0	0	0	0	-1	-0.5
other shopping	-1	-1	-1	-0.5	0	0	1	0	0	0	0	0	-1	-0.5
services	-1	-1	-1	-0.5	0	1	0	0	0	-1	-1	-1	-1	-0.5
health care	-0.5	-0.5	-0.5	-0.5	-0.5	0	-0.5	-0.5	-0.5	-1	-1	-1	-0.5	-0.5
fitness	0	0	0	0	0	0	0	1	0	-1	-1	-1	0	-1
wellness	0	0	0	0	0	0	0	1	0	-1	-1	-1	0	-1
nature	-0.5	0	0	0	0	0	0	1	0	0	0	0	-0.5	-1
culture	-0.5	-1	-1	-1	-1	-1	0	1	0	-1	-1	-1	-0.5	-1
sports	0	-1	-1	-1	-1	-1	-1	1	0	-1	-1	-1	0	-1
friends	0	0	0.5	0	0	0	0	0.5	1	0	0	0	0	0.5
game	0	-1	-1	-1	-1	-1	-1	1	0	-1	-1	-1	0	0
public event	-1	-1	-1	-1	-1	-1	-1	1	0	0	0	0	-1	0
private event	0	-1	-1	-1	-1	-1	-1	0	1	-1	-1	0	0	0.5
alcohol	0	-0.5	-0.5	-0.5	-0.5	-0.5	0	0.5	1	-0.5	0	0.5	0	1
food	0.5	-0.5	0	0	1	0	0	0	0.5	1	1	1	0.5	0.5
sweets	0	0	0	0	0	0	0	0.5	0.5	0.5	0	1	0	0.5
coffee	0	0	0	0	0	0	0	0	0.5	1	1	0	0	0
tea	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0

1,520 places (6.1% of all). These places have also high similarity to “education”, “transport”, and “lunch”.



In 233 places out of the initial 1,520 (15%, 0.9% of all places) the similarity to the “work” profile has been reinforced based on the topic frequencies.

Classification of the places according to the highest combined score (minimum 0.8)



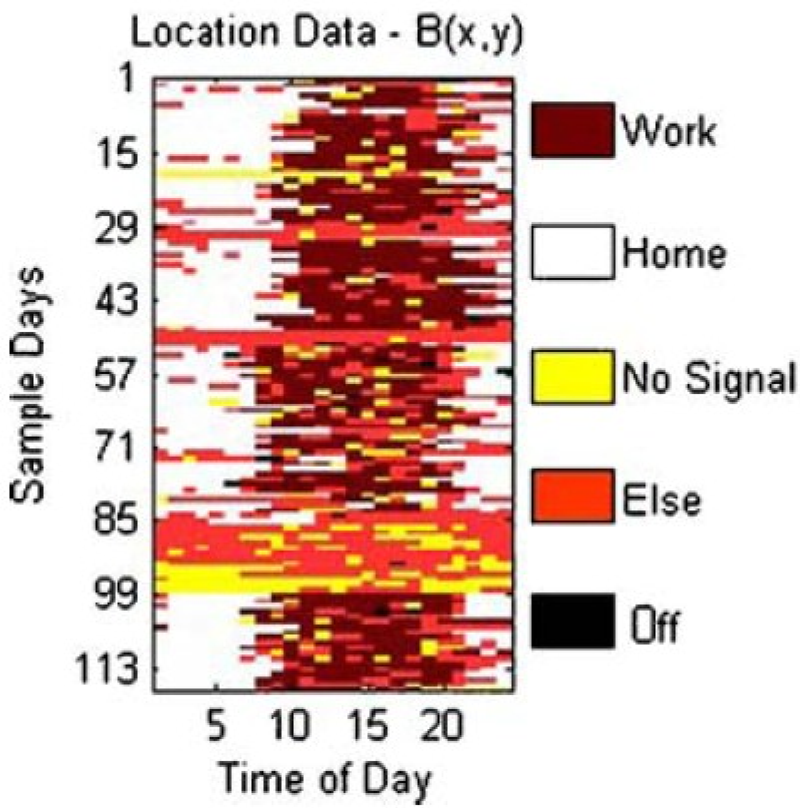
20,247 places (81.2%) are not classified; 4,688 (18.8%) are classified, of them 4,048 (16.2%) were visited in at least 6 days

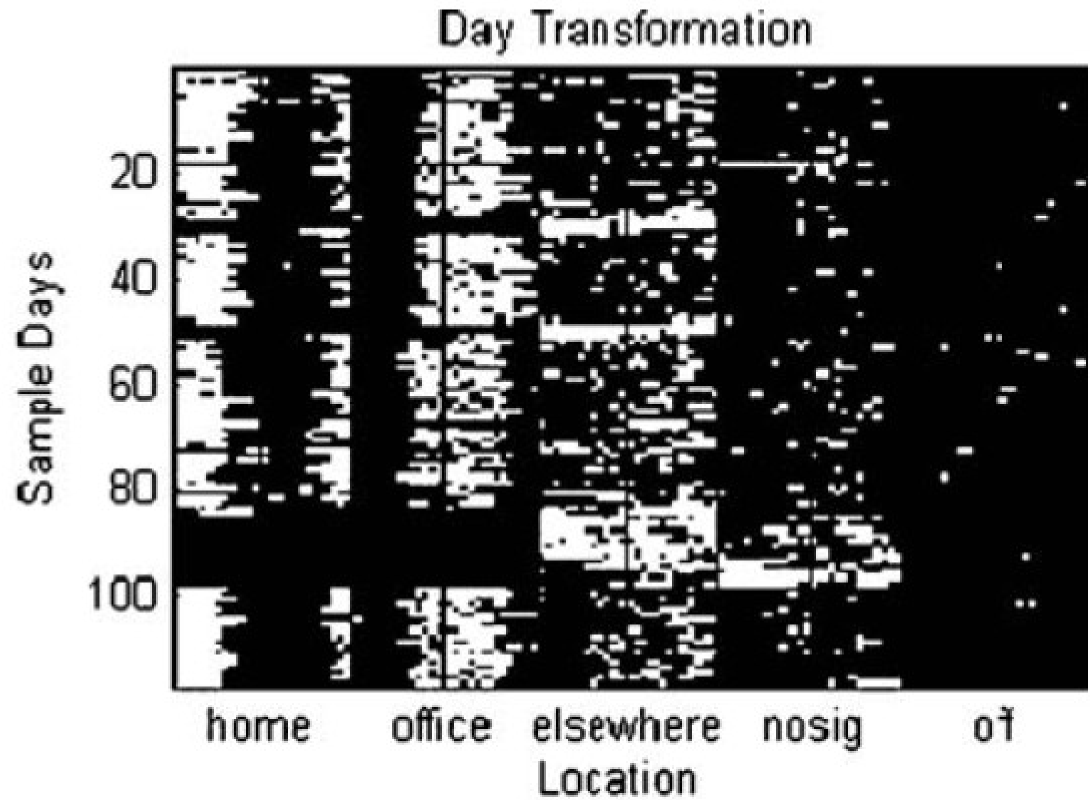
Activity Recognition – Inductive approach

Eigen-behaviours

Input

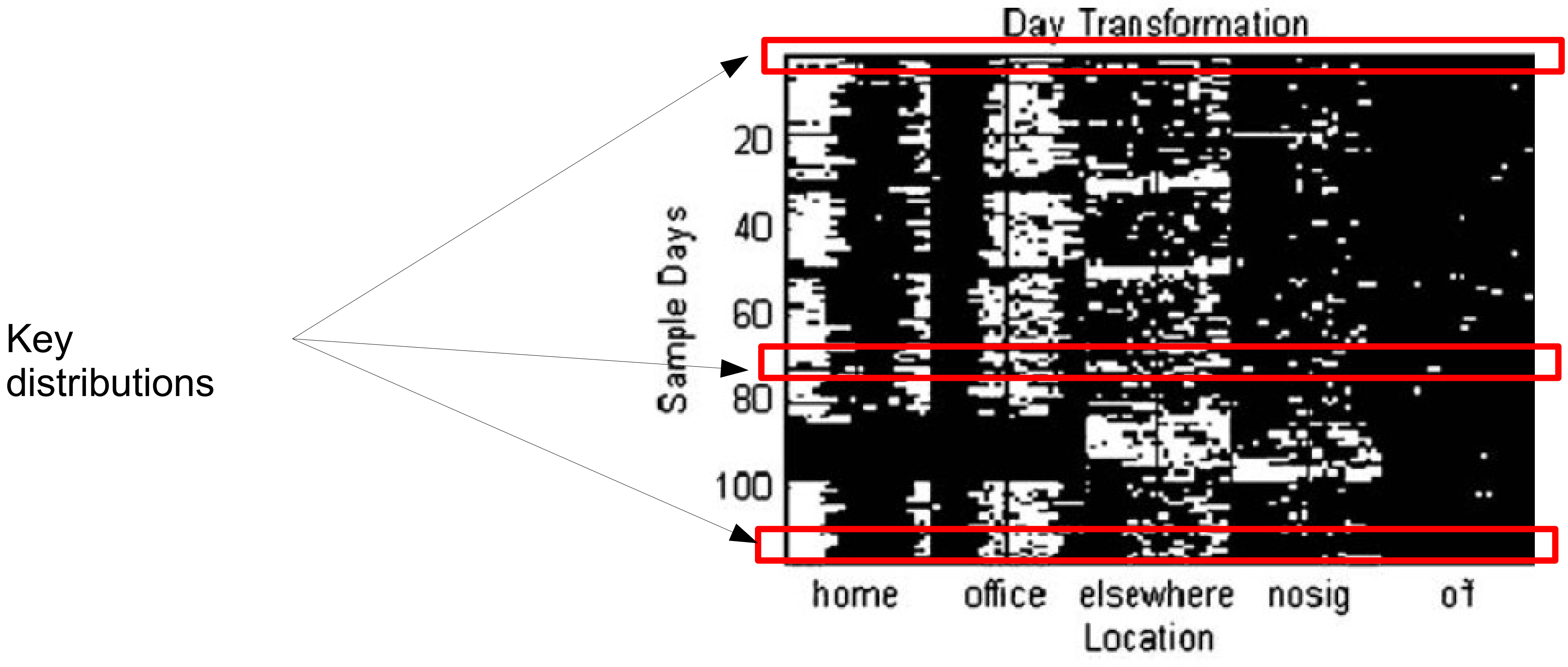
- Left: subject's behavior over the course of 113 days for five situations / activities
- Right: same data represented as a binary matrix of 113 days (D) by 120 (H, which is 24 multiplied by the five possible situations)





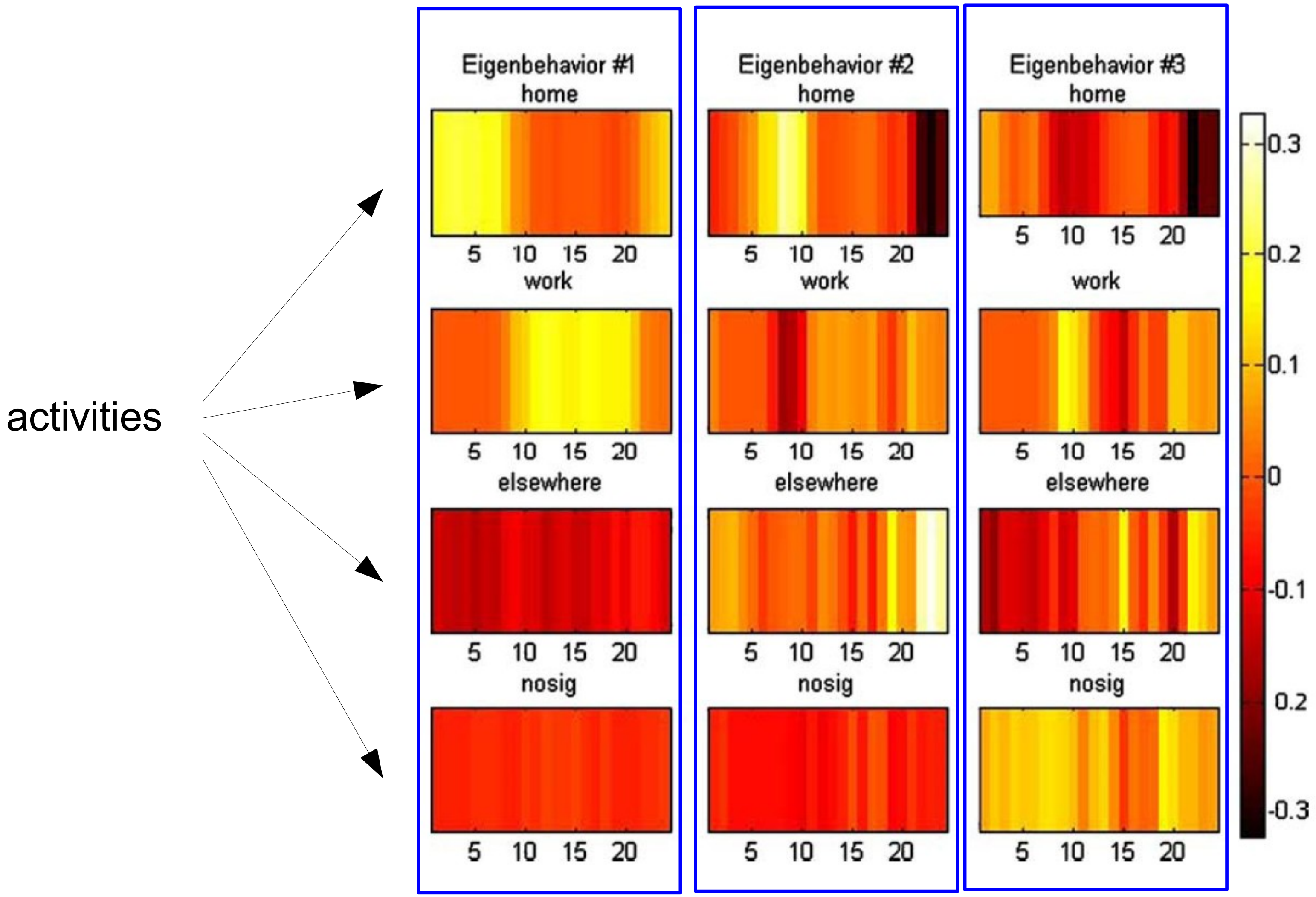
Eigen-behaviours
Method

- Are there key activity distributions from which to infer all others through linear combination?
- Same idea as PCA

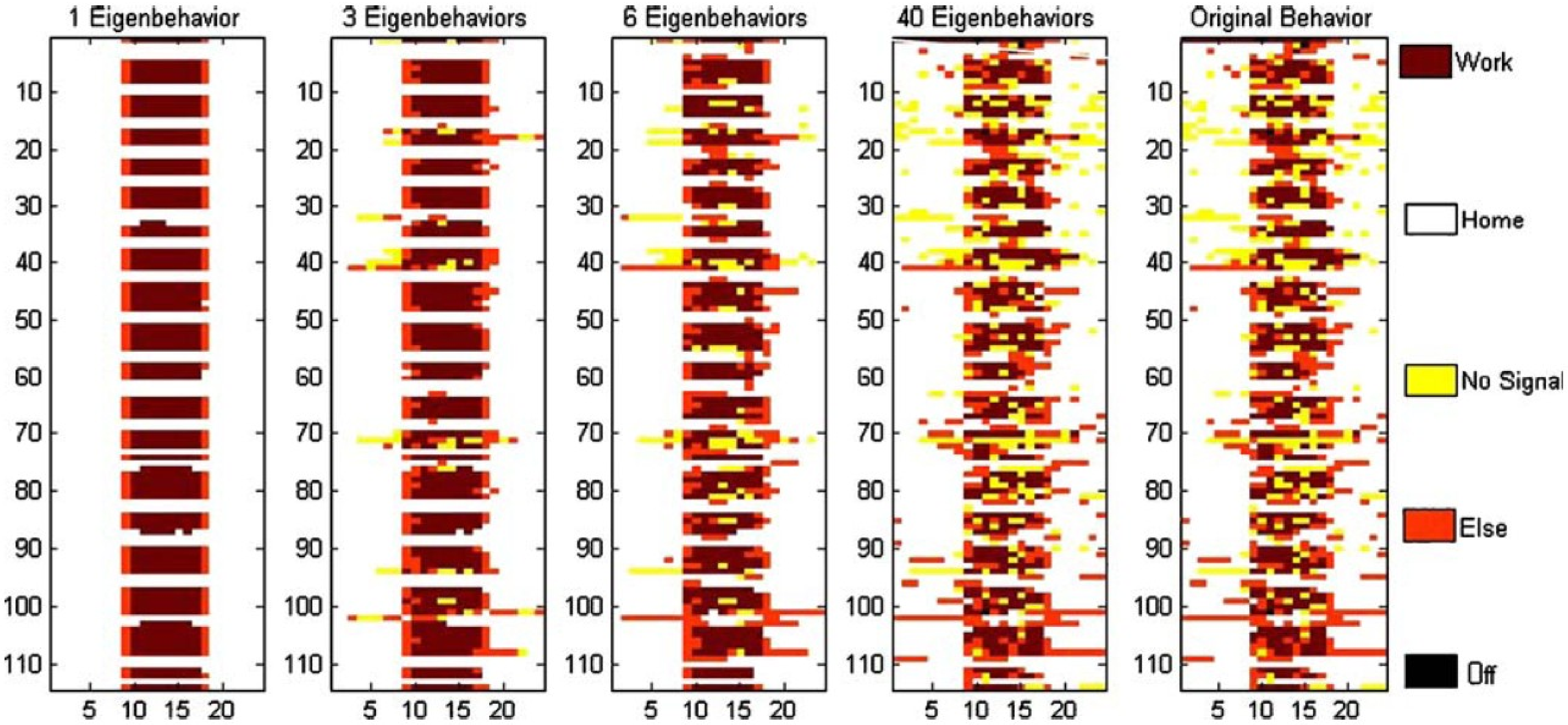


Eigen-behaviours
Output

- Set of 3 representative eigen-behaviours
- Each user's activity can be rewritten as their linear combination

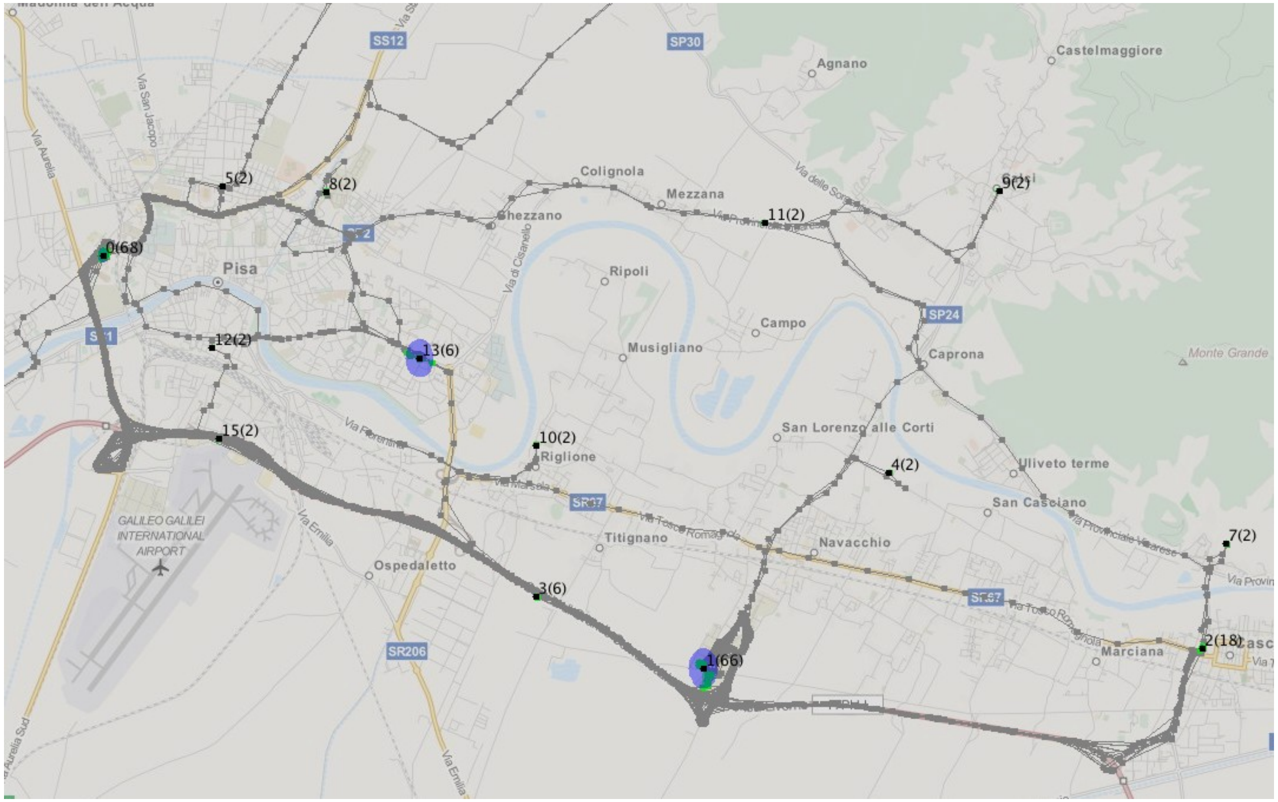


Eigen-behaviours
Example



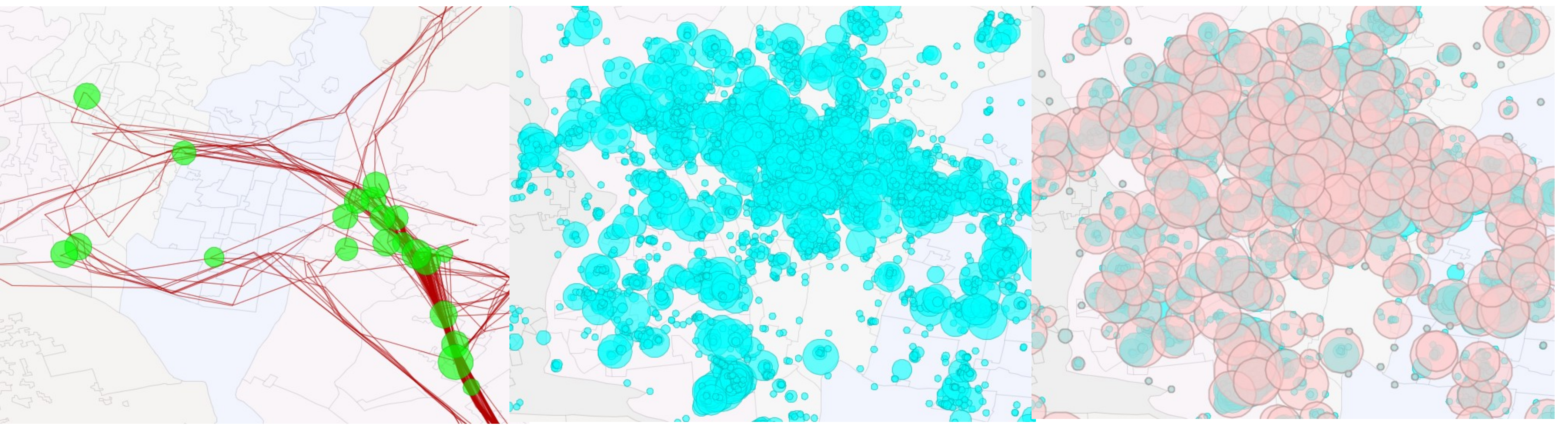
Individual Mobility Networks

How to synthesize Individual Mobility?

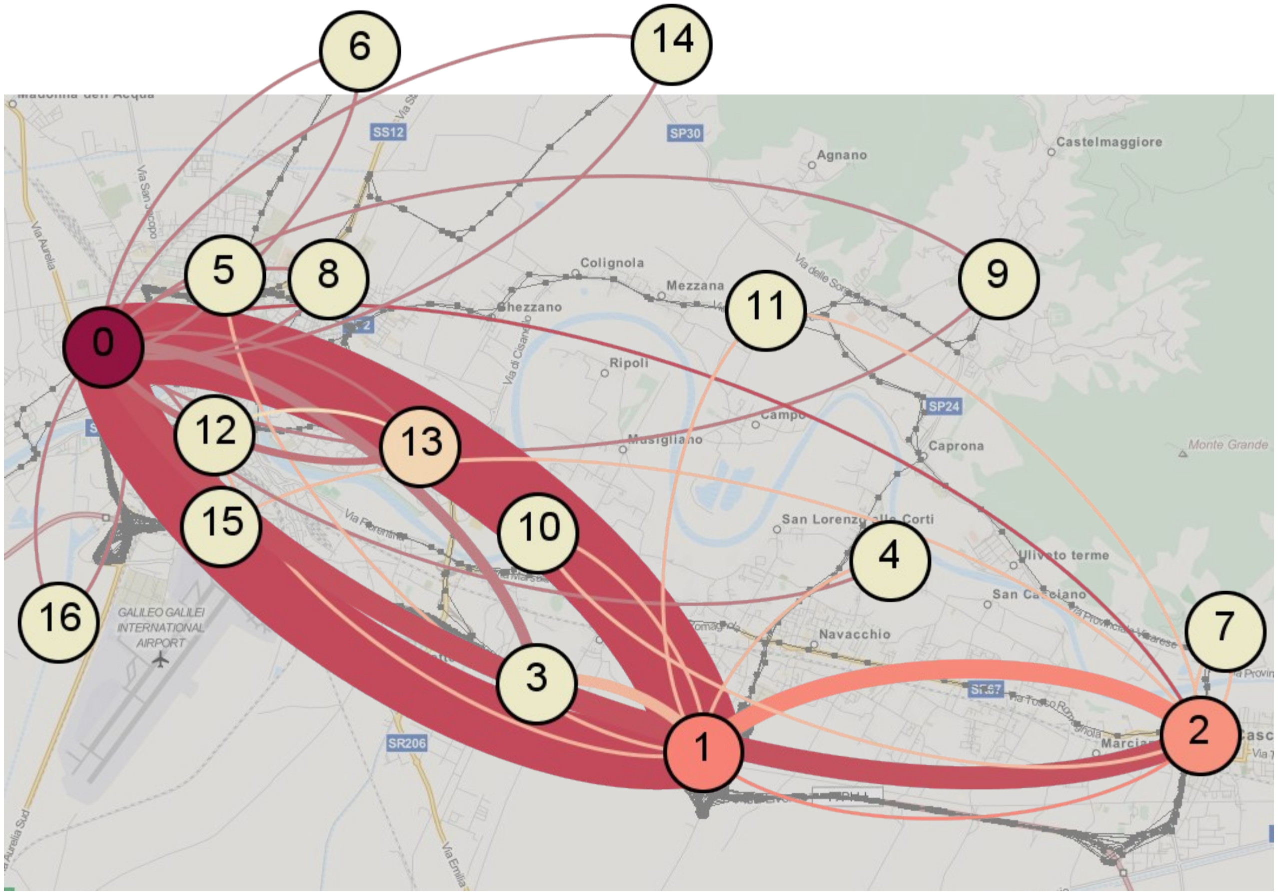


Mobility Data Mining methods automatically extract relevant episodes: locations and movements.

Rank individual preferred locations

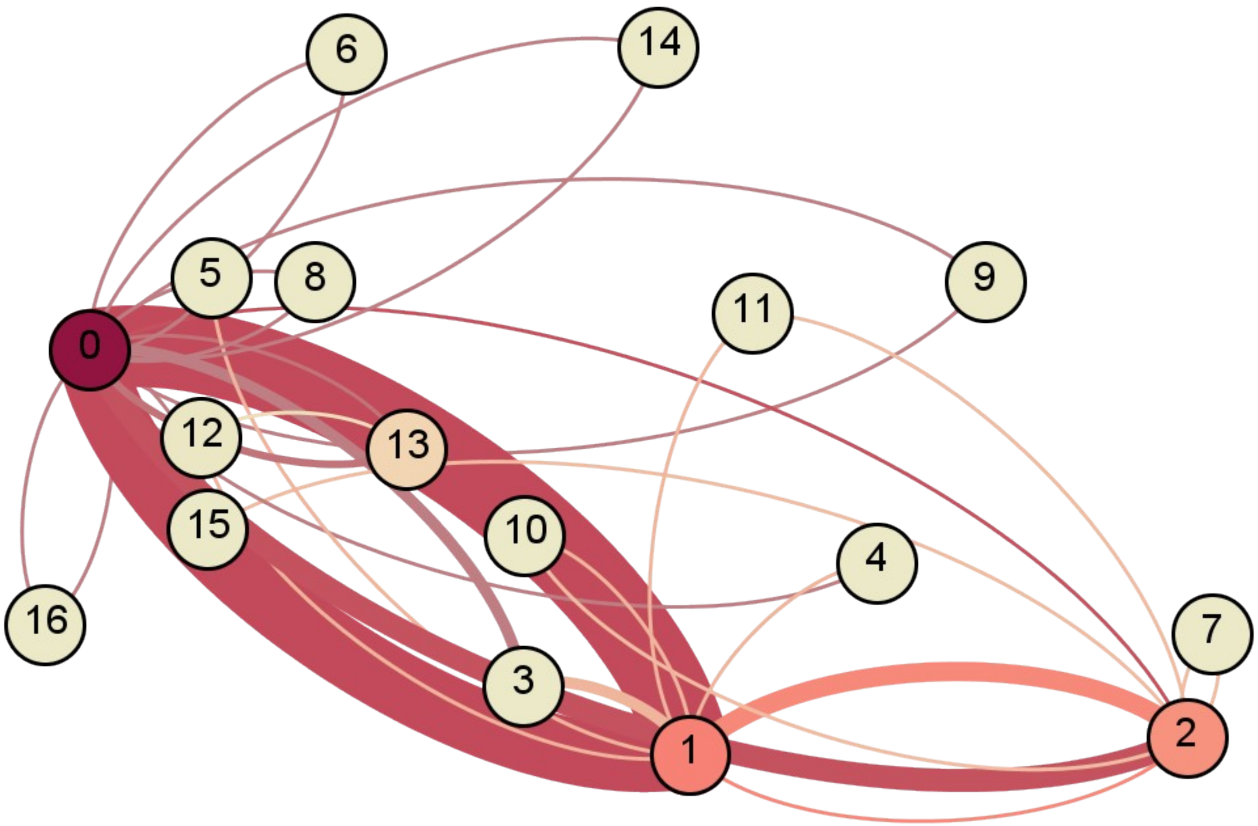


How to synthesize Individual Mobility?



Graph abstraction based on locations (nodes) and movements (edges)

How to synthesize Individual Mobility?

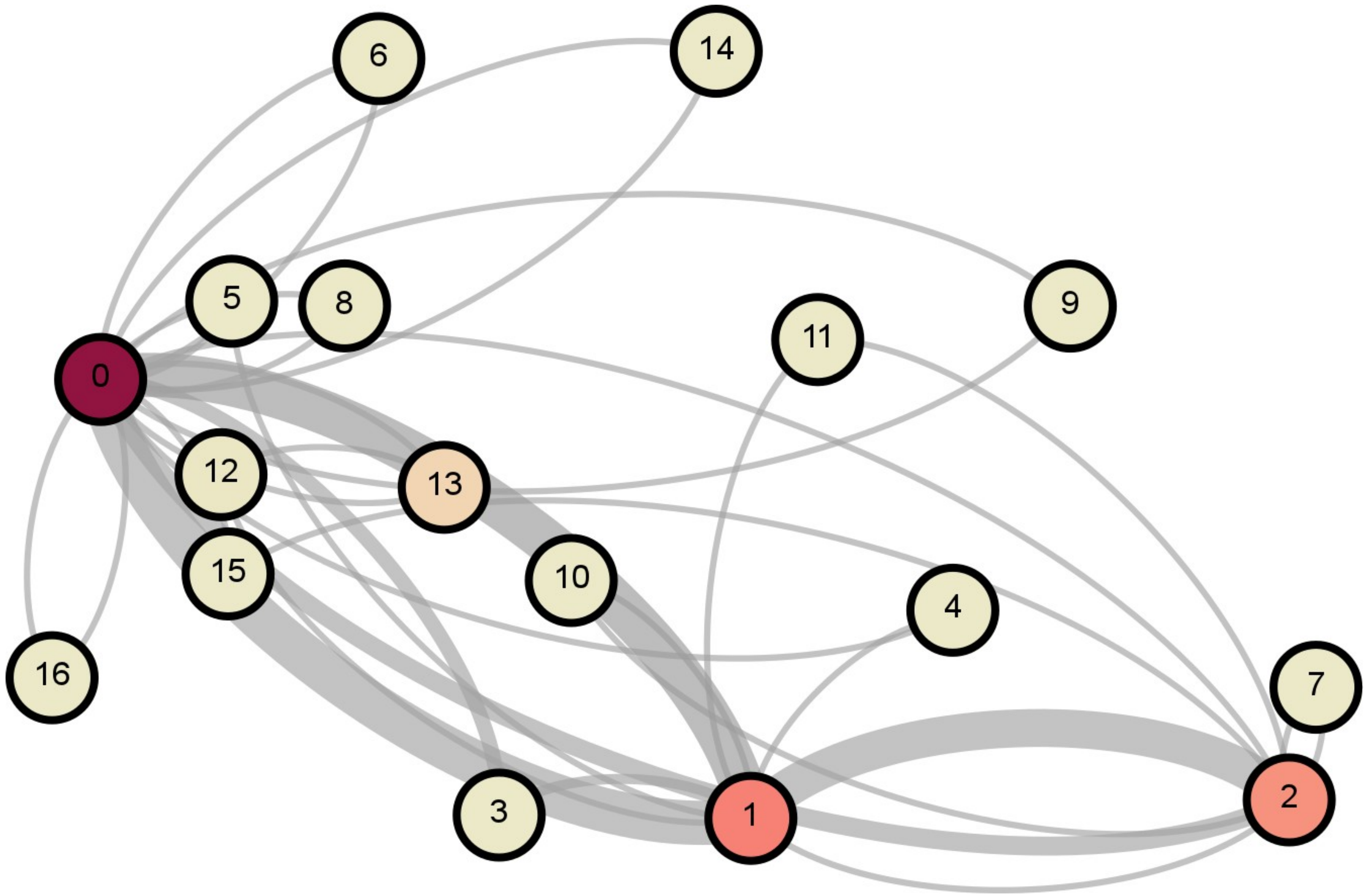


High level representation

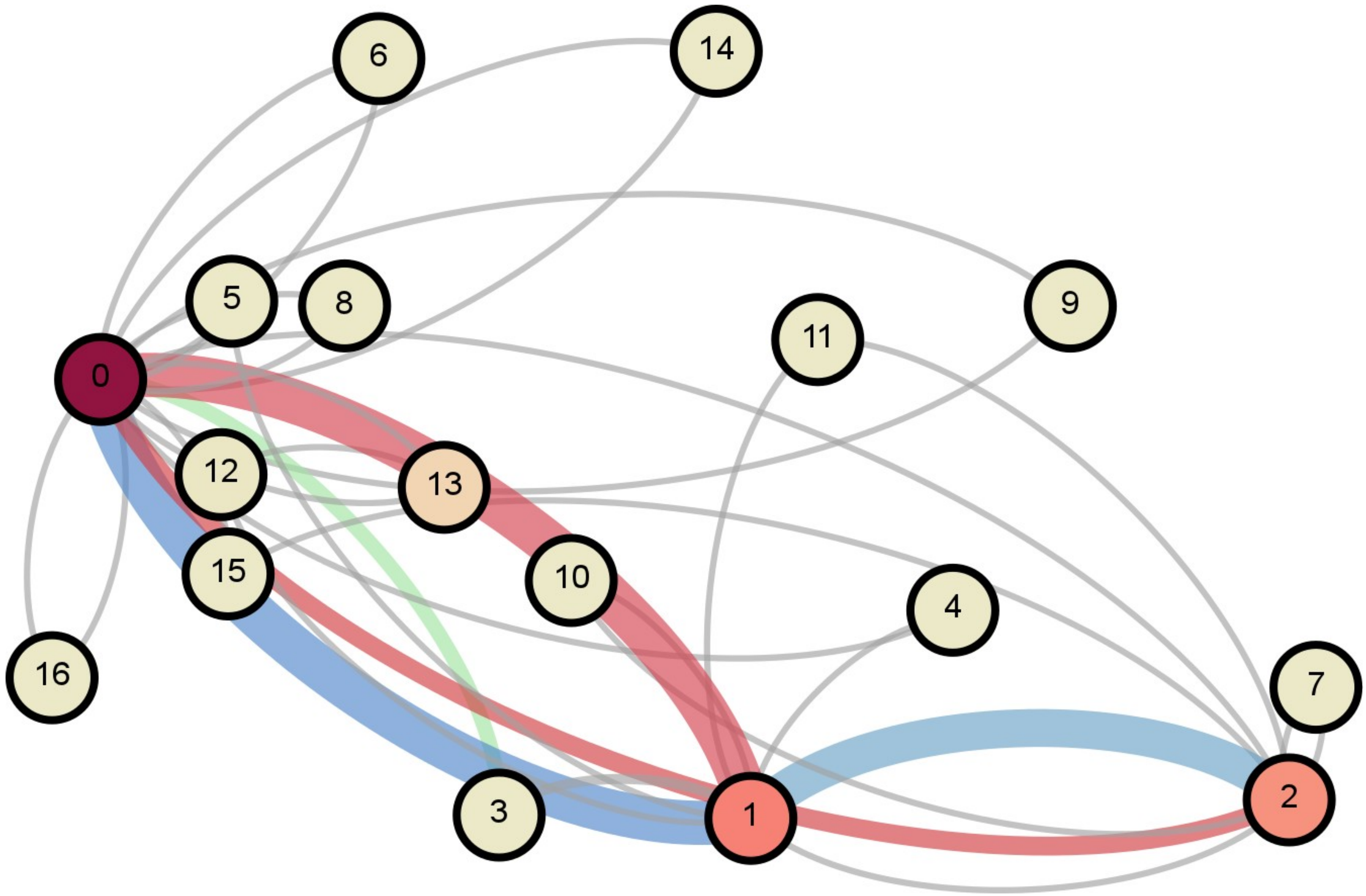
Aggregation of sensitive data

Abstraction from real geography

From raw movement…



… to annotated data



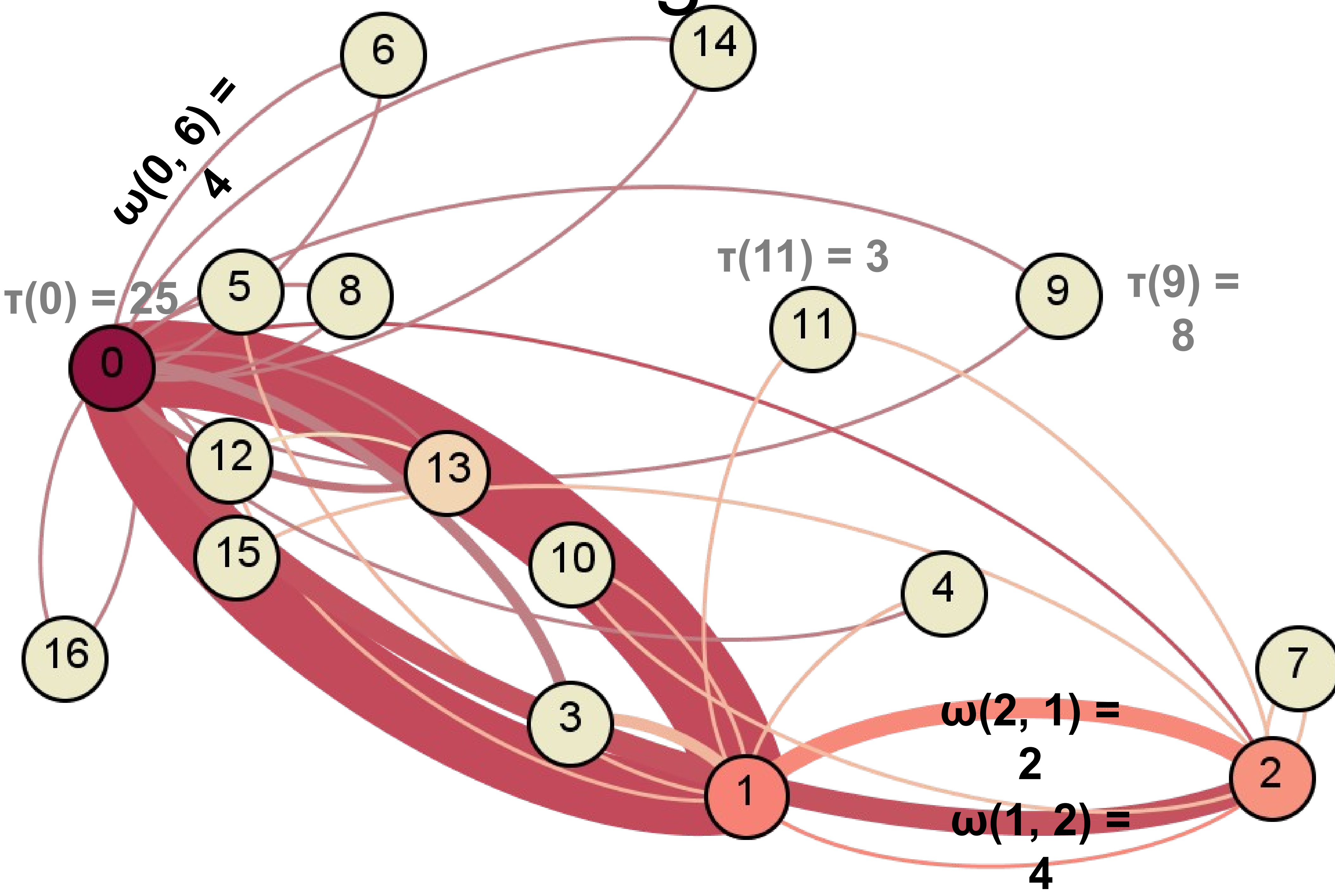
The ABC classifier

1) Build from data an
   Individual Mobility Network (IMN)

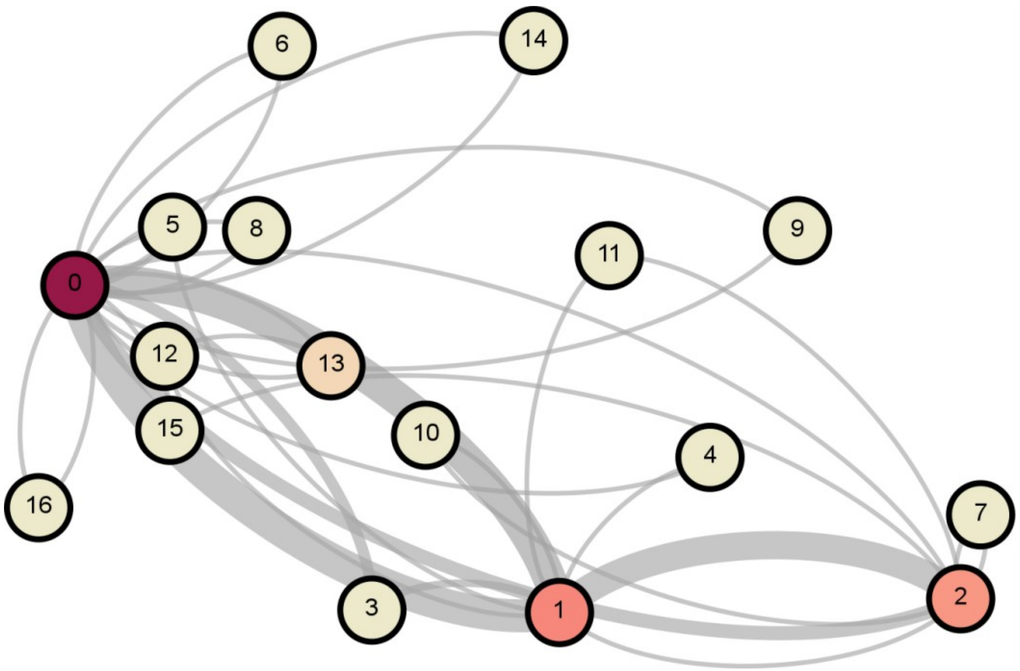
2) Extract structural features from the IMN

3) Use a cascading classification with label propagation (ABC classifier)

Extracting the IMN



Extracting the IMN

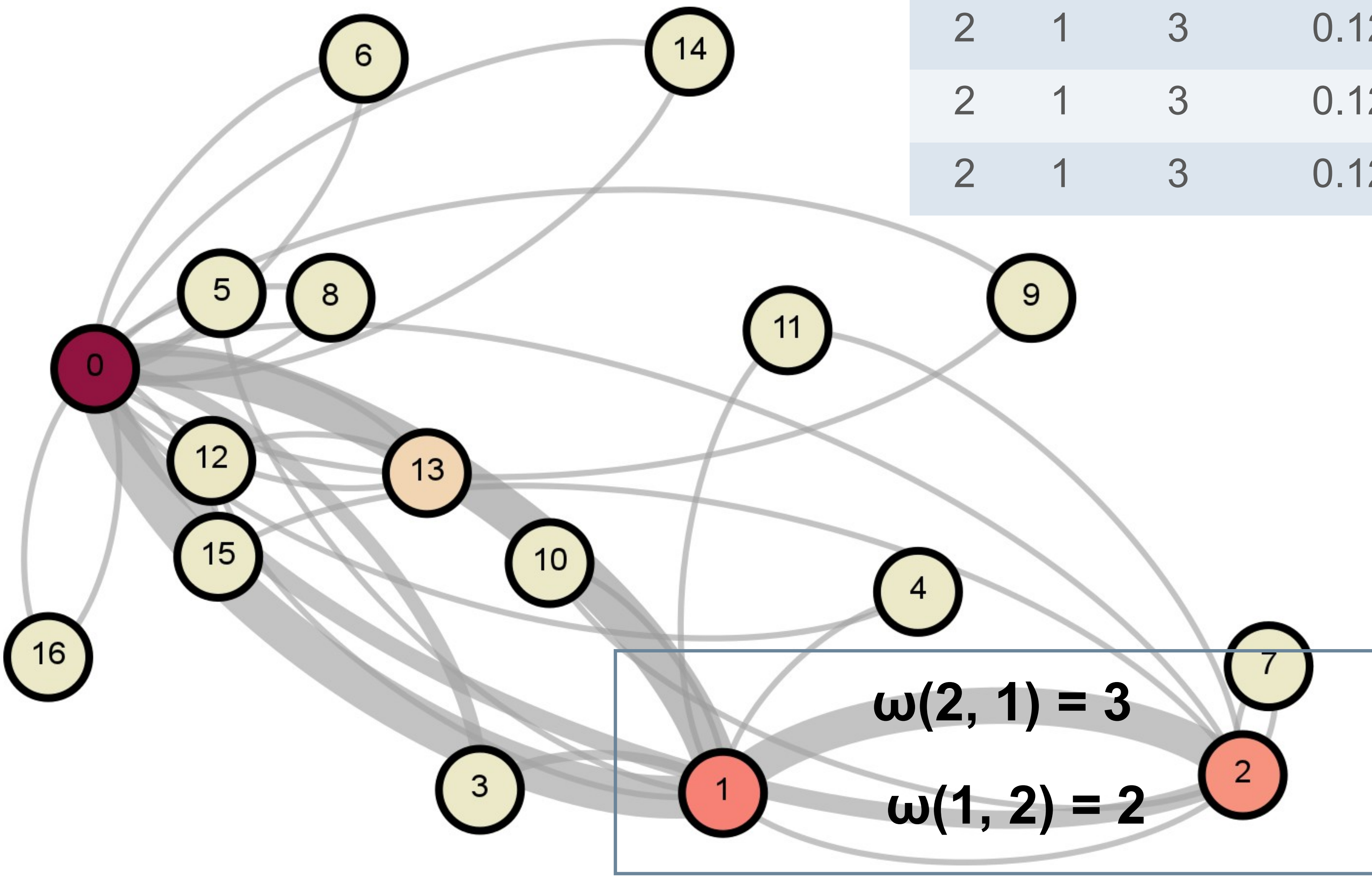


Trip Features
Length
Duration
Time Interval
Average Speed

Network Features	
centrality	clustering coefficient average path length
predictability	entropy
hubbiness	degree betweenness
volume	edge weight flow per location

Extracting the IMN

from	to	weight	ccFrom	ccTo	duration
1	2	2	0.22	0.12	10 min
1	2	2	0.22	0.12	5 min
2	1	3	0.12	0.22	4 min
2	1	3	0.12	0.22	6 min
2	1	3	0.12	0.22	4 min

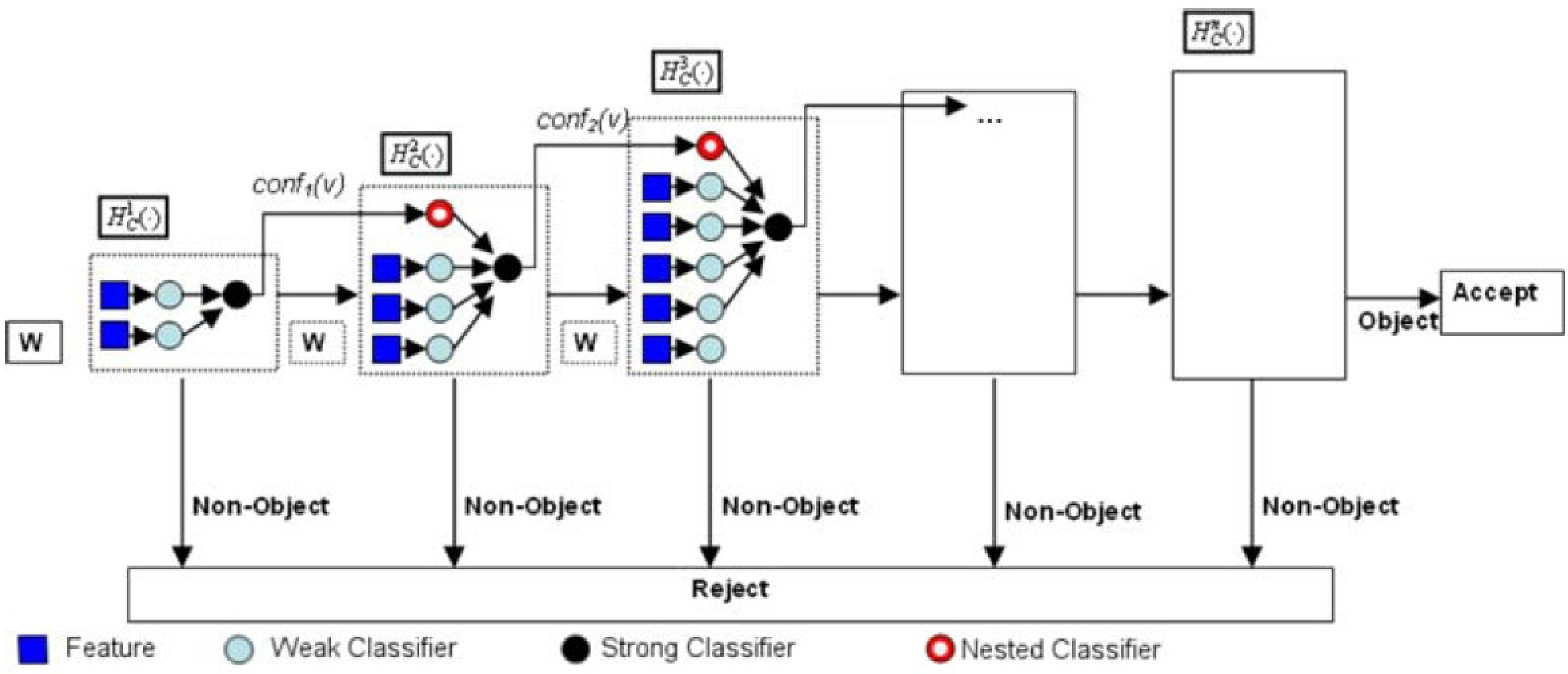


ABC Classifier

- Principles:
  - The activities of a user should be predicted as a whole, not separately
  - Some activities are easy to classify
  - Other activities might benefit from contextual information obtained from previous predictions
- E.g.: a place frequently visited after work might be more likely to be leisure / shopping

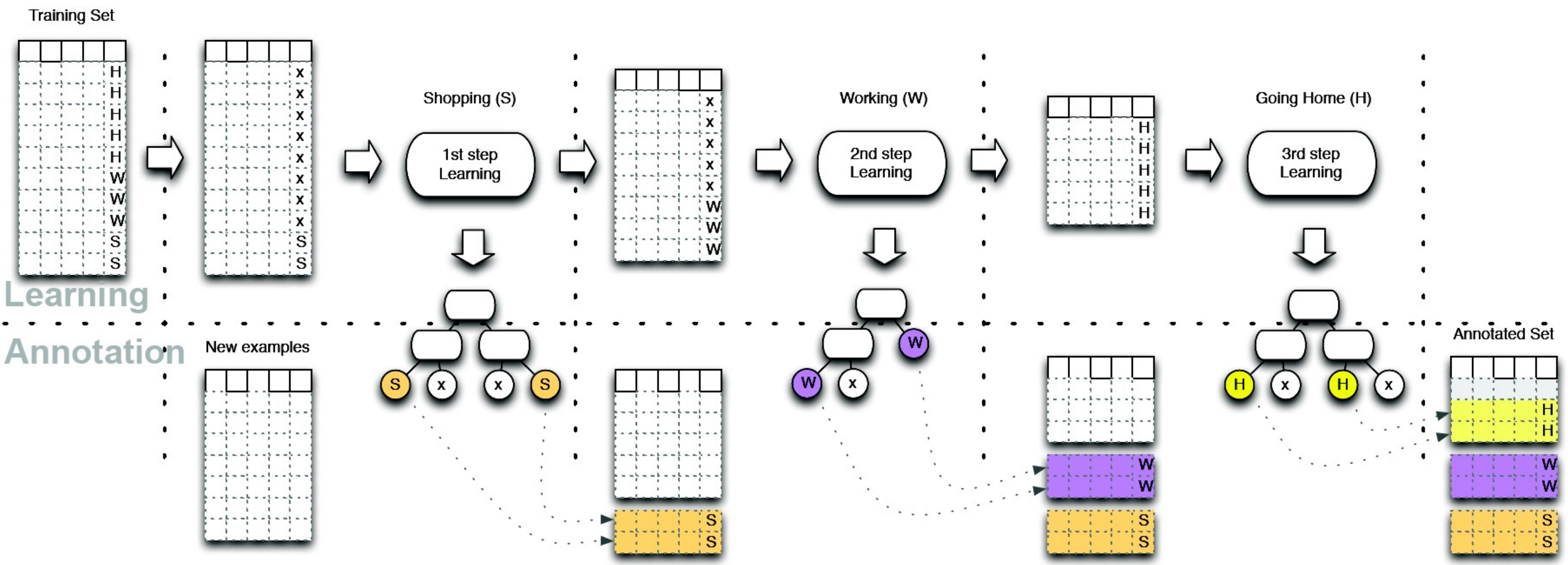
ABC Classifier

- Inspired by Nested Cascade Classification

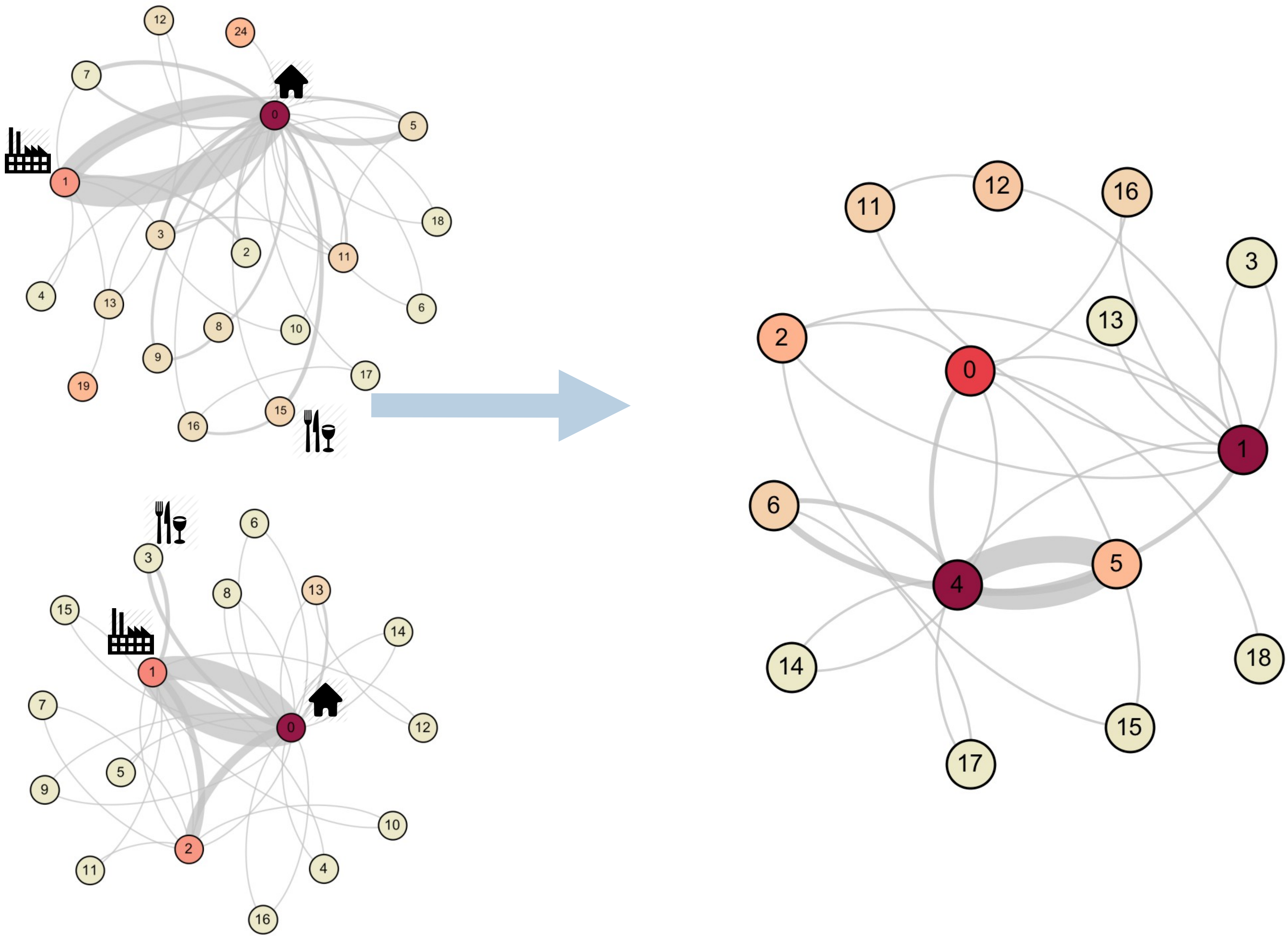


ABC Classifier

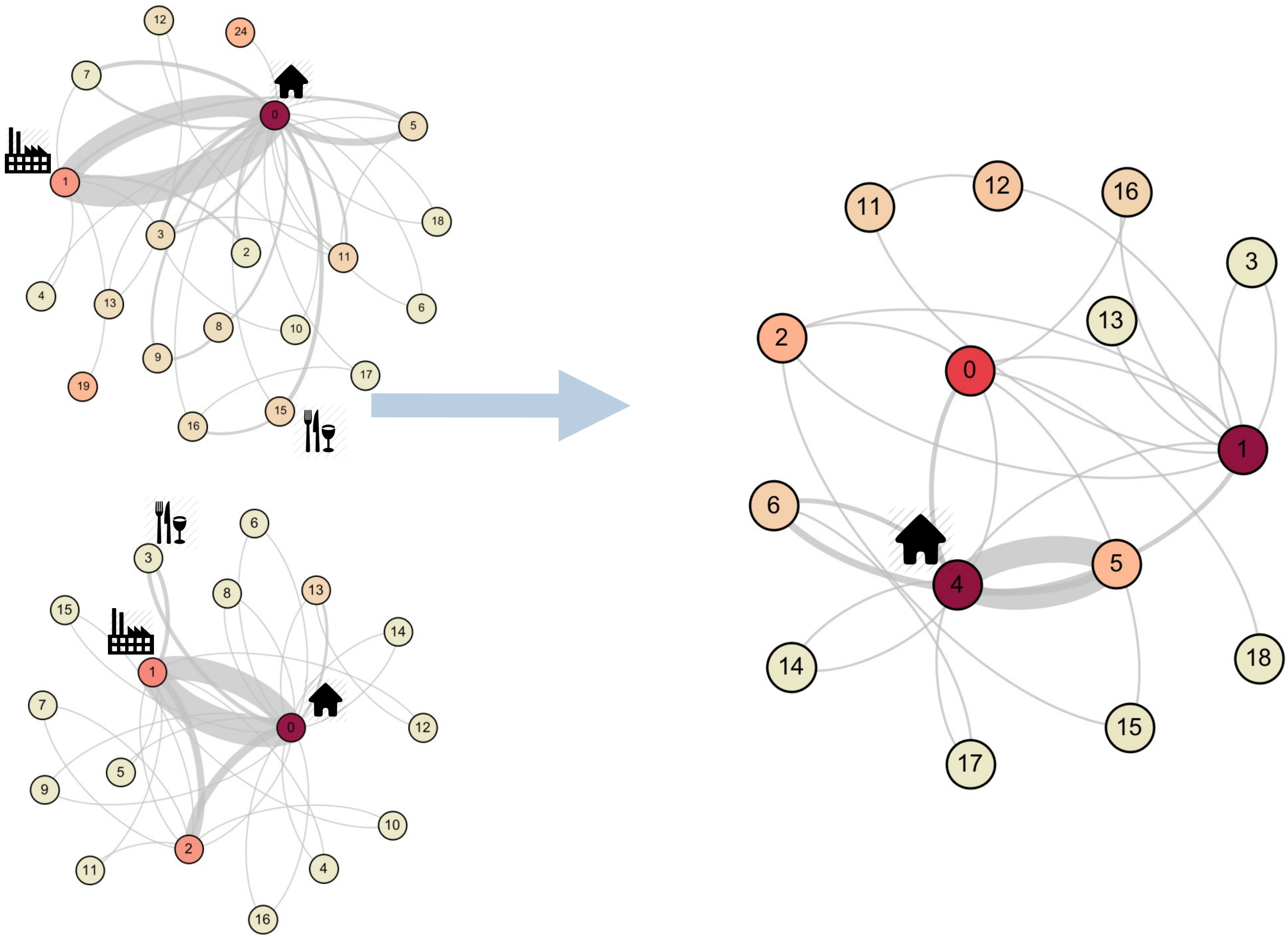
- Inspired by Nested Cascade Classification



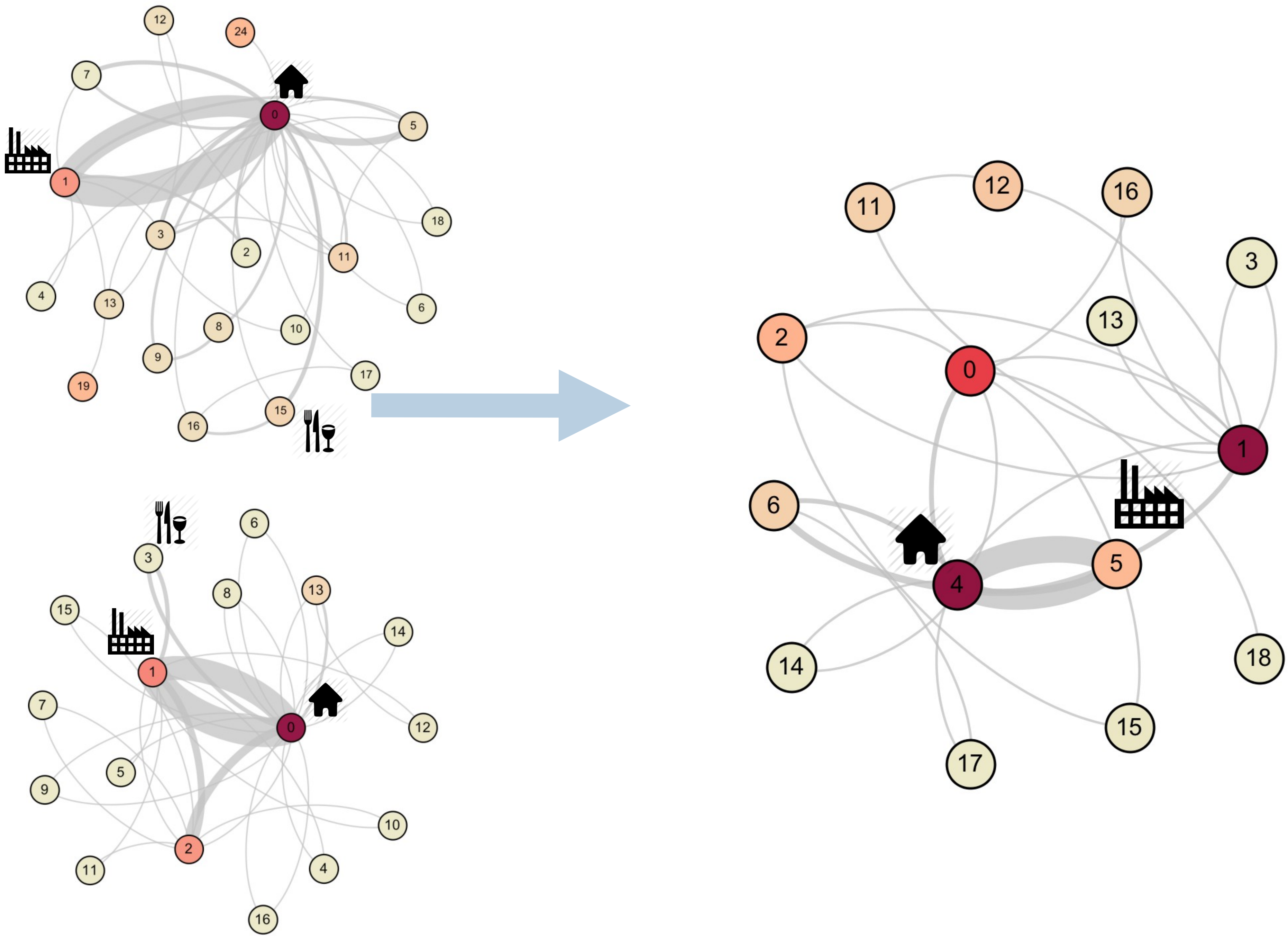
The ABC classifier



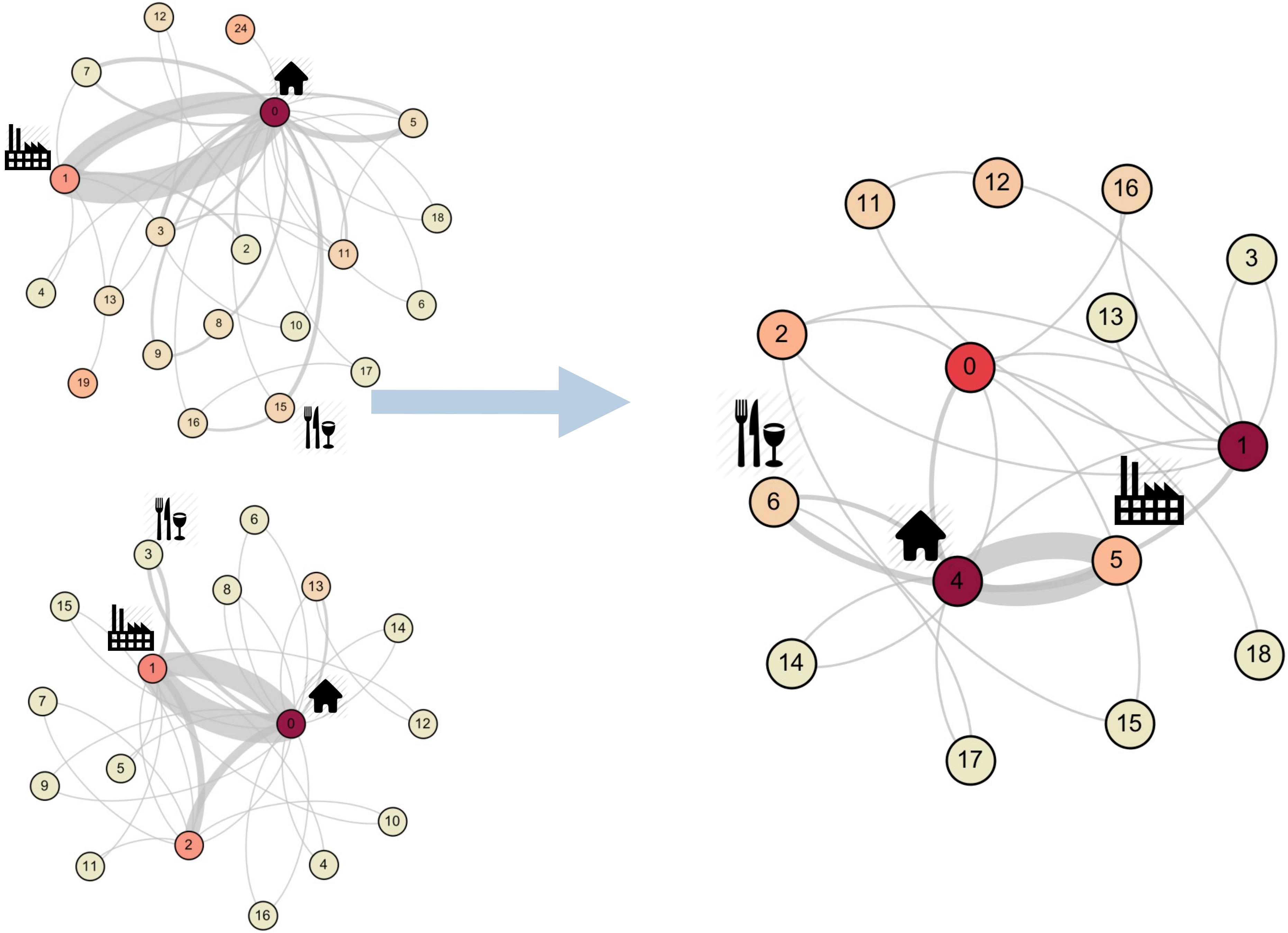
The ABC classifier



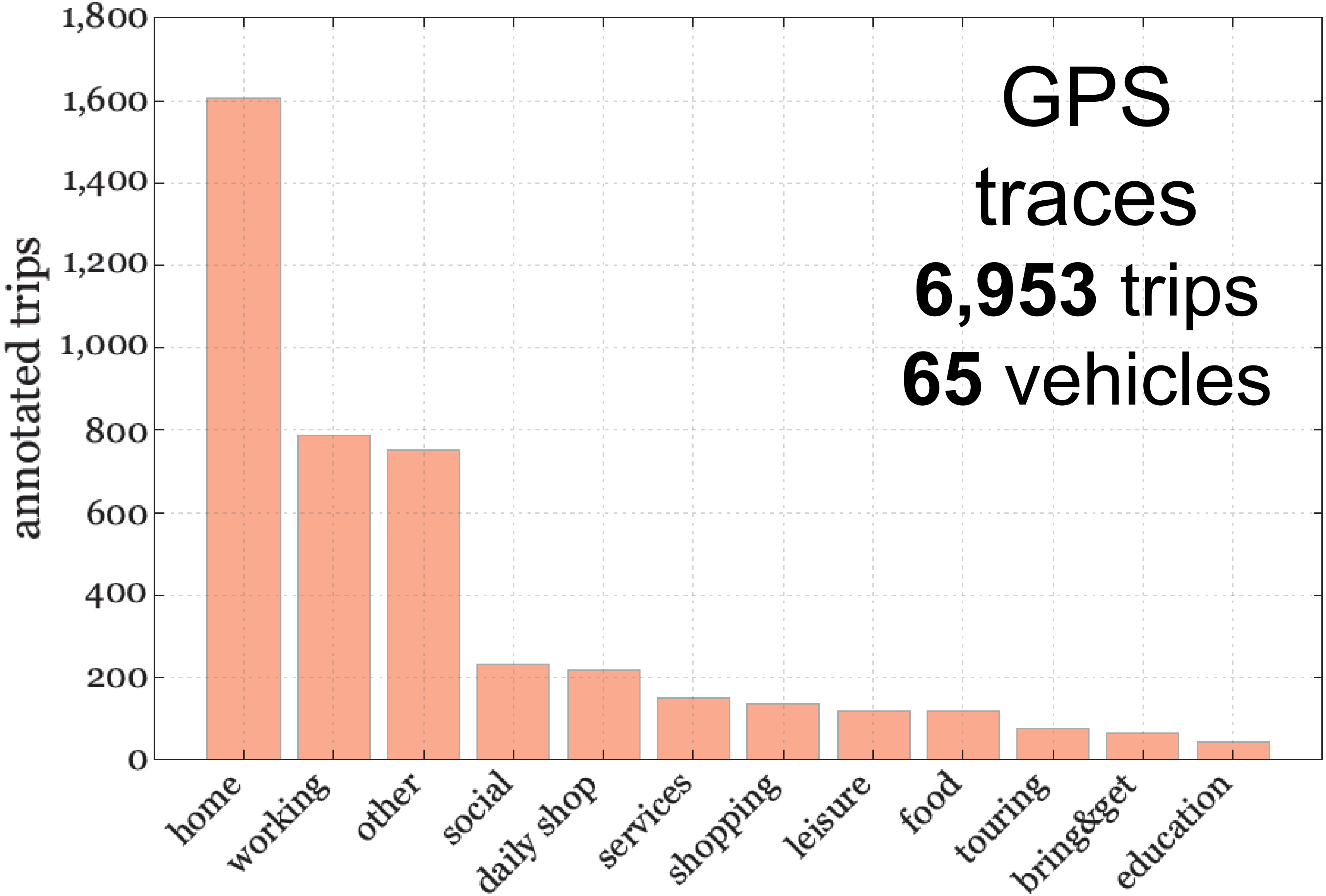
The ABC classifier



The ABC classifier



Experiments

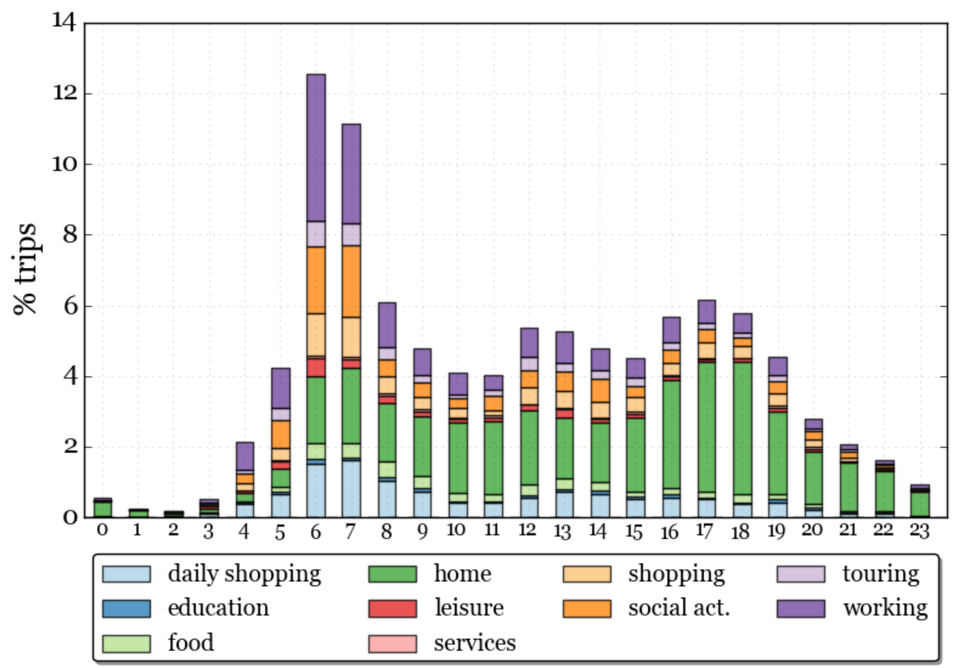


GPS
traces
6,953 trips
65 vehicles

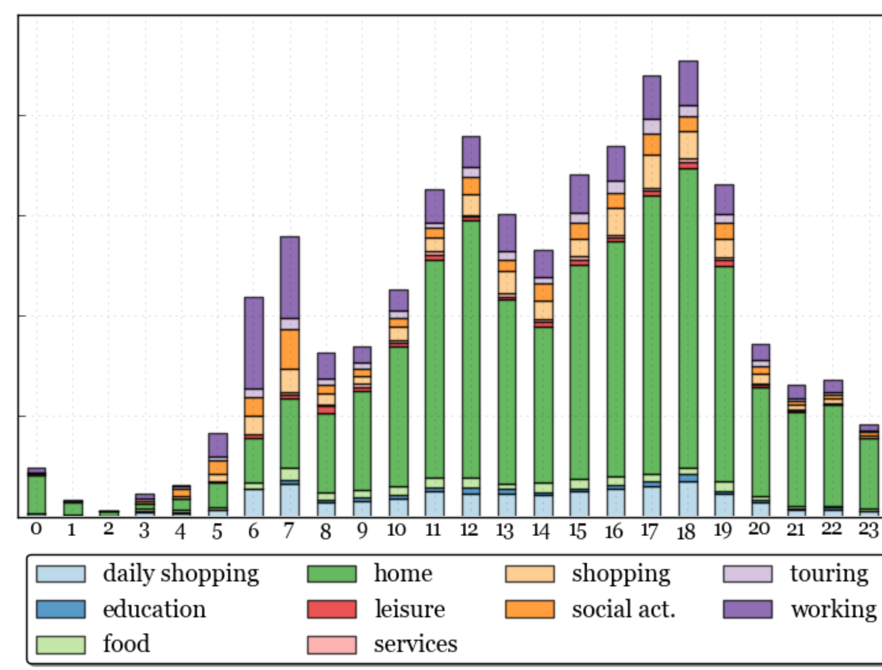
Semantic Mobility Analytics

Temporal Analysis

- Pisa traffic



In

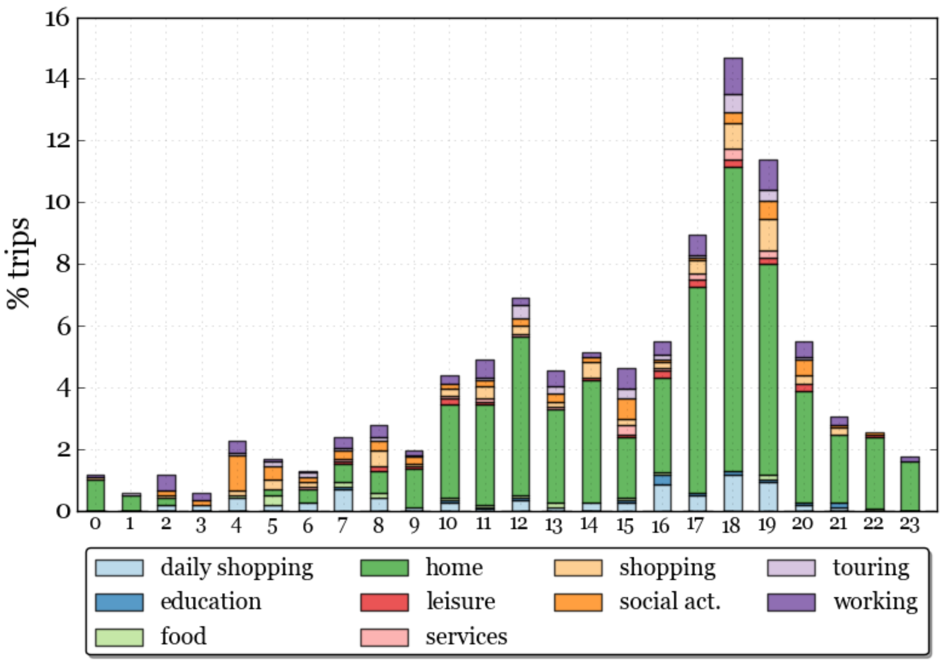


Out

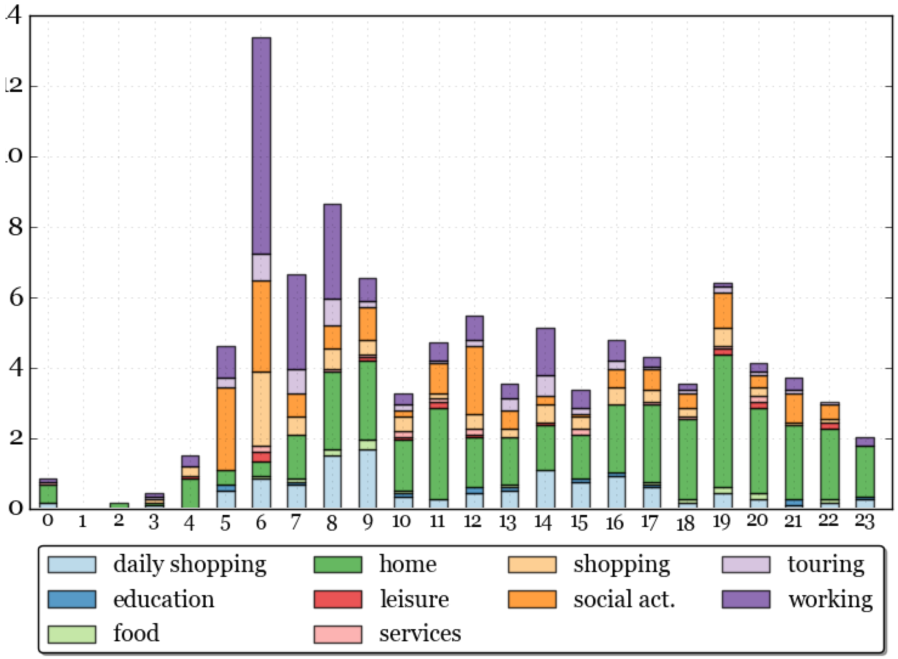
Semantic Mobility Analytics

Temporal Analysis

- Calci traffic



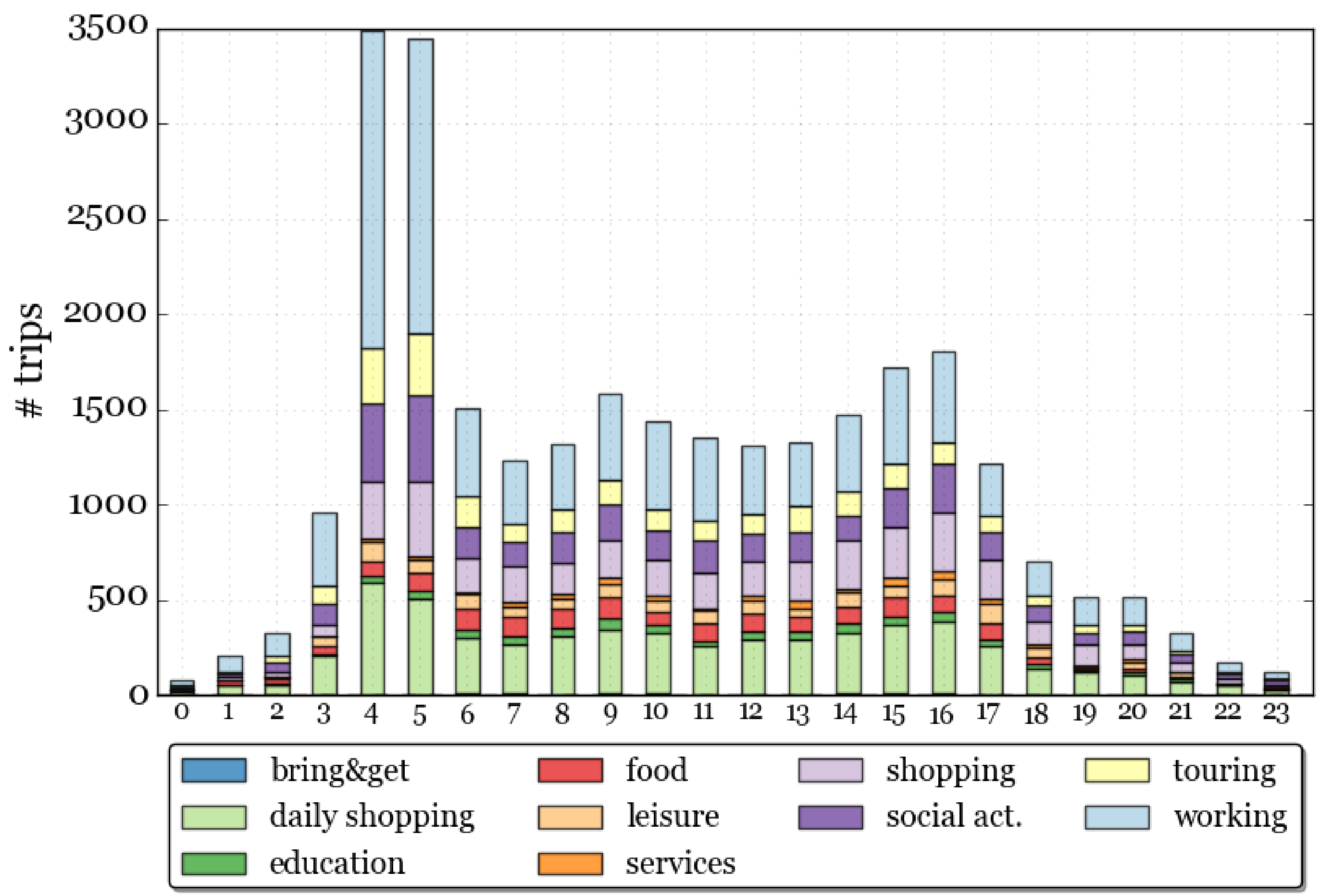
In



Out

Semantic Mobility Analytics

Temporal Analysis



User Profiling

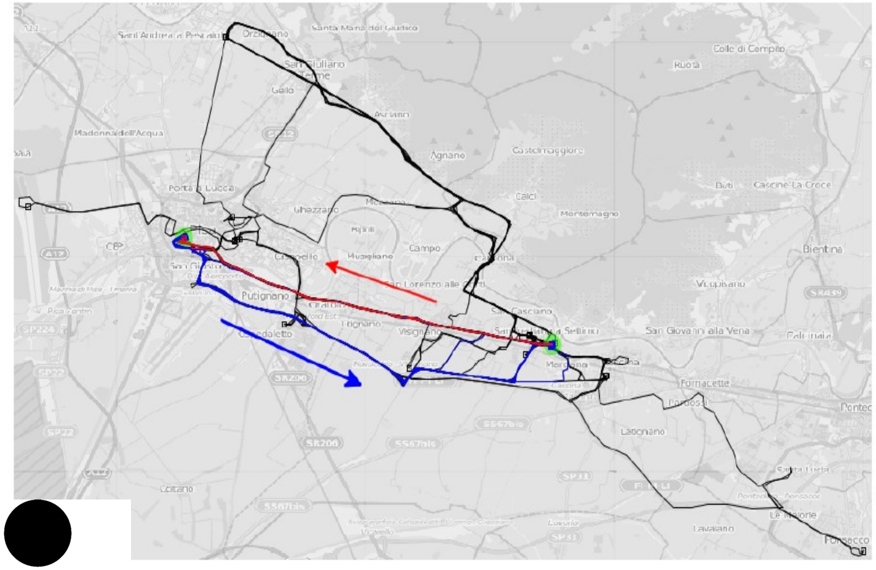
- In computer science, is the process of construction and extraction of models representing user behavior generated by computerized data analysis.
- Are employed to study, analyze and understand human behaviors and interactions.
- Are exploited by many applications to make predictions, to give suggestions etc.

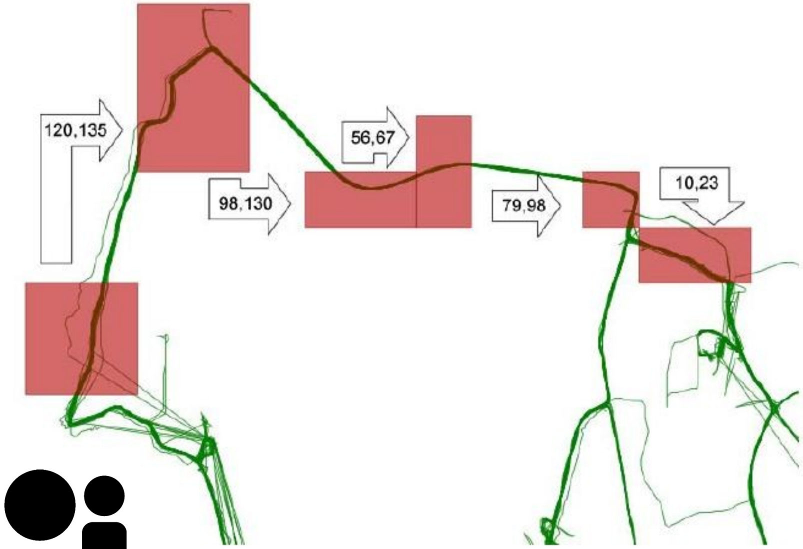
MYWay:
Trajectory Prediction

Individual and Collective Profile

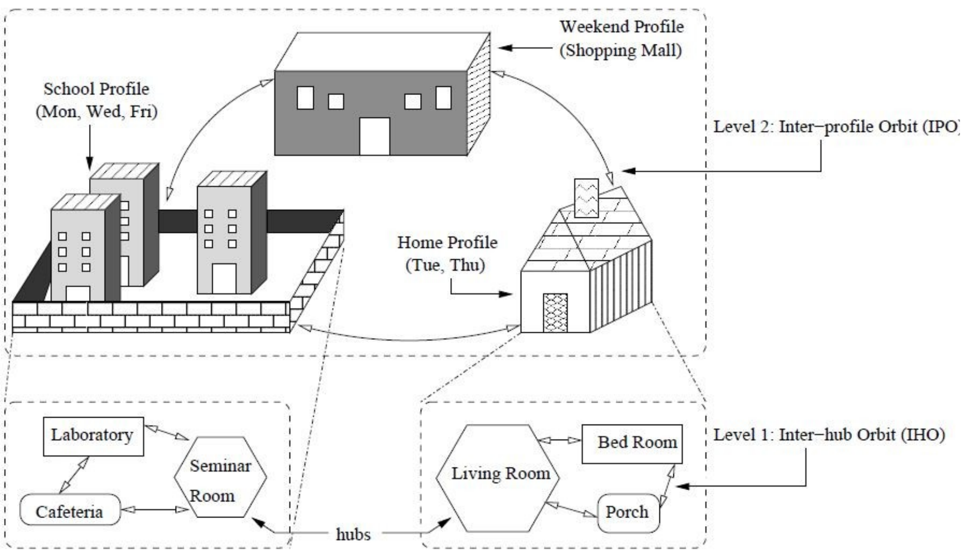
- **Individual Profile**
- Input: Individual Data
- Output: Individual Patterns

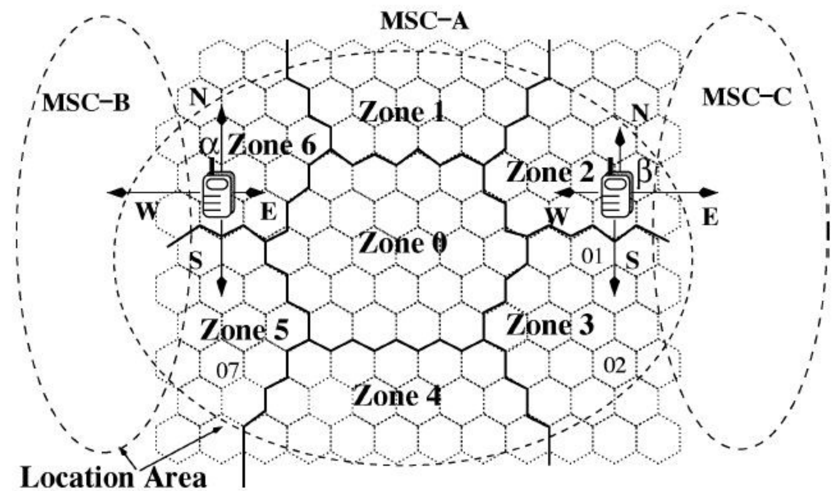
- **Collective Profile**
- Input: Collectivity Data
- Output: Collective Patterns





Prediction using probability mixture models





J. Ghosh, M. J. Beal, H. Q. Ngo, and C. Qiao. **On profiling mobility and predicting locations of wireless users.** 2006.

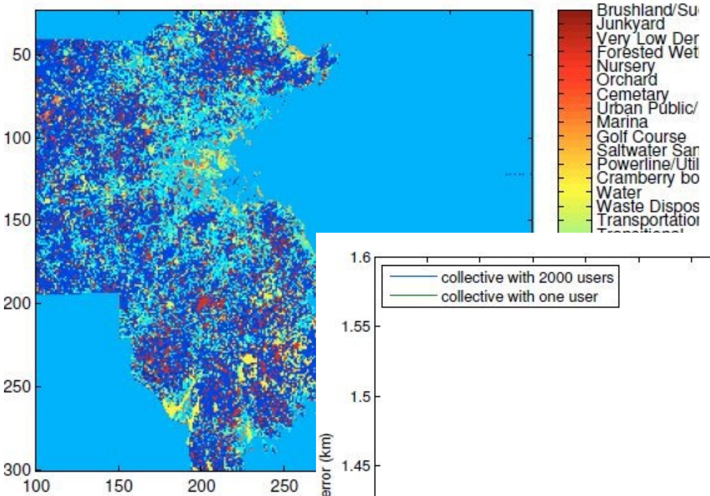
J. Ghosh, H. Q. Ngo, and C. Qiao. **Mobility profile based routing within intermittently connected mobile ad hoc networks (icman).** 2006.

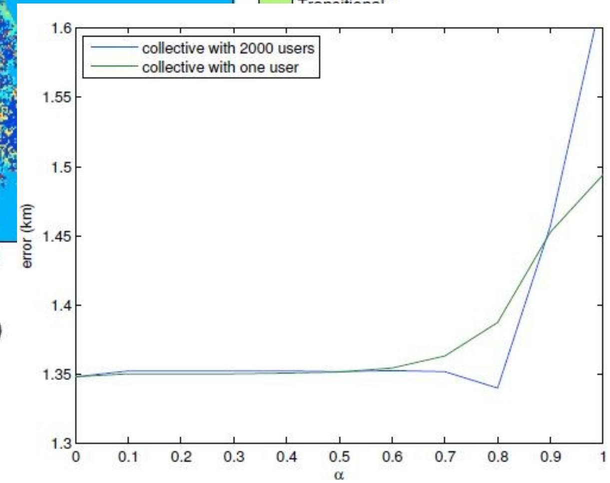
I. F. Akyildiz and W. Wang. **The predictive user mobility prole framework for wireless multimedia networks.** 2004.

Prediction based on individual and collective preferences

$$P_I(x_{k+1} = j | x_k = i) = \frac{\sum_{k_t=1}^{\lfloor k/T \rfloor} f_I(x_{k-Tk_t+1} = j | x_{k-Tk_t} = i)}{\lfloor k/T \rfloor}$$
$$\forall j \in N_n, \qquad (1)$$

$$P_C(x_{k+1} = j | x_k = i) = P_{LU}(x_{k+1} = j) P_{POI}(x_{k+1} = j) \cdot$$
$$\cdot P_D(x_{k+1} = j | x_k = i) \cdot$$
$$\left(\sum_{j'=1}^{n} P_{LU}(x_{k+1} = j') P_{POI}(x_{k+1} = j') P_D(x_{k+1} = j' | x_k = i) \right)^{-1}$$
$$\forall j \in N_n. \qquad (2)$$

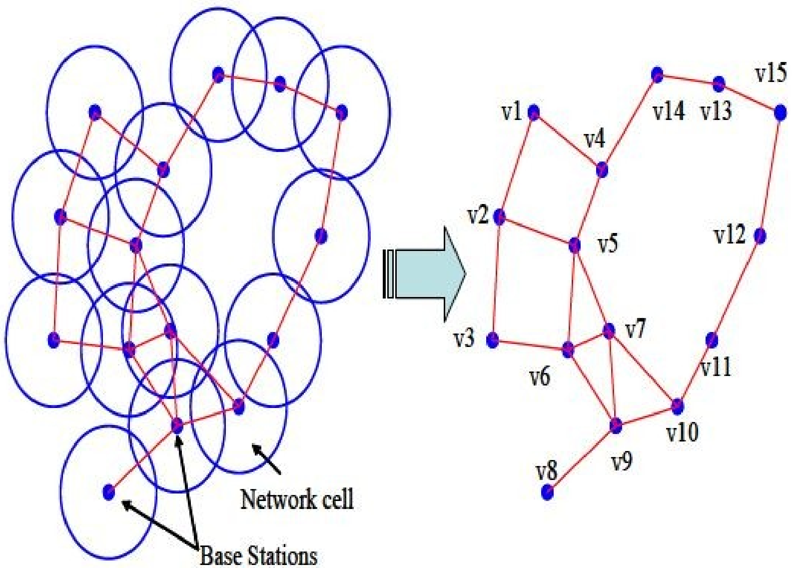


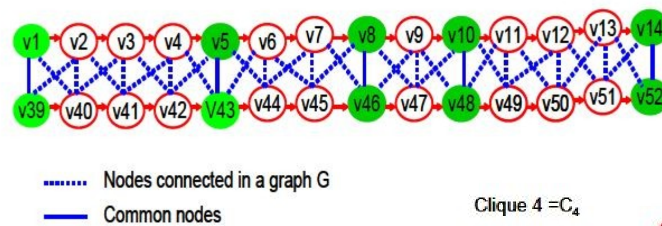


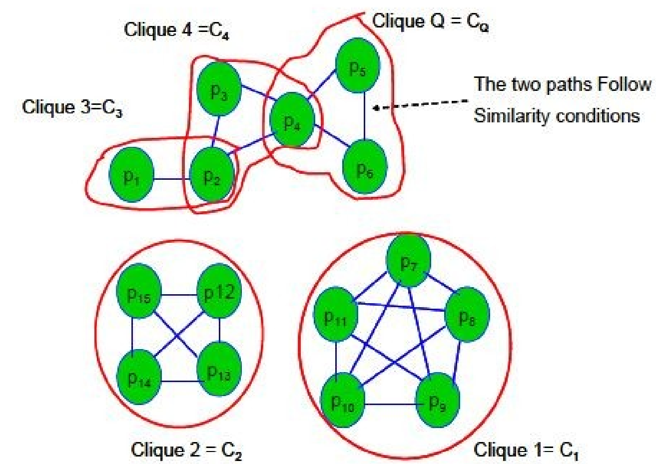
$$P(x_{k+1}(u) = j | x_k(u) = i) = \qquad (3)$$
$$(1 - \alpha(k)) P_I(x_{k+1}(u) = j | x_k(u) = i) +$$
$$+ \alpha(k) P_C(x_{k+1} = j | x_k = i),$$
$$\forall j \in N_n,$$

F. Calabrese, G. Di Lorenzo, and C. Ratti. **Human mobility prediction based on individual and collective geographical preferences.** 2010.

Prediction using complex networks and probability



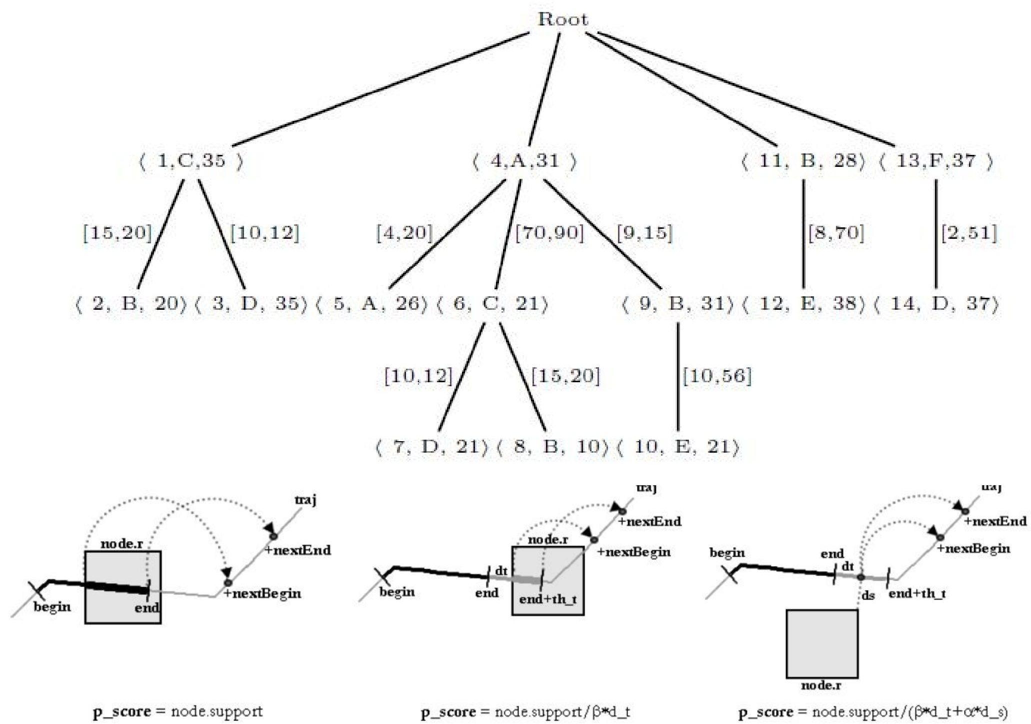




D. Barth, S. Bellahsene, and L. Kloul. **Mobility prediction using mobile user profiles.** 2011.

D. Barth, S. Bellahsene, and L. Kloul. **Combining local and global proles for mobility prediction in lte femtocells.** 2012.

Collective prediction using t-patterns

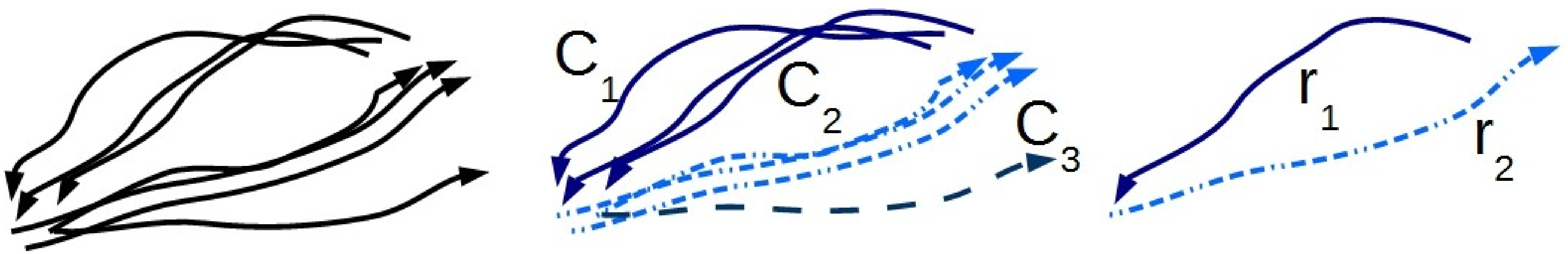


F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. **Trajectory pattern mining.** 2007.

A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. **Wherenext: a location predictor on trajectory pattern mining.** 2009.

Mobility Profiling

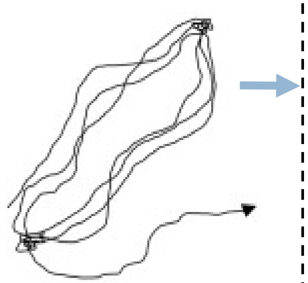
A concise model ables to describe the user's mobility in terms of representative movements, i.e. routines.



This model is called Mobility Profile.

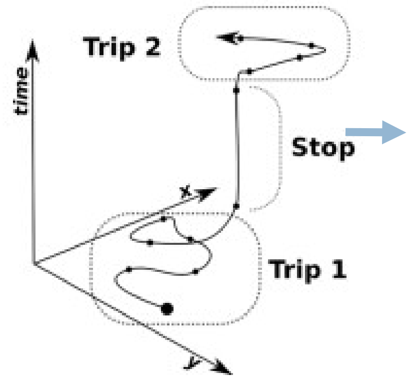
Mining mobility user profiles for car pooling. Trasarti, Pinelli, Nanni, Giannotti. KDD 2011

Derived patterns and models: mobility profiles



User history

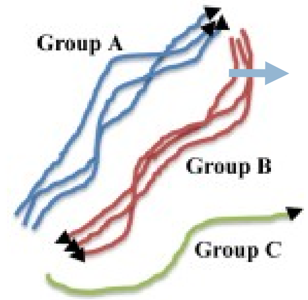
An ordered sequence of spatio-temporal points.



Trips construction

Cutting the user history when a **stop** is detected

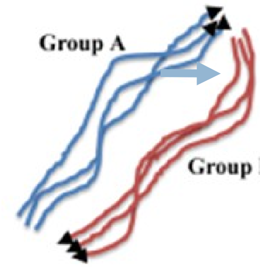
Stops Spatial Threshold
Stops Temporal Threshold



Grouping

Performing a density based **clustering** equipped with a **spatio temporal** distance function

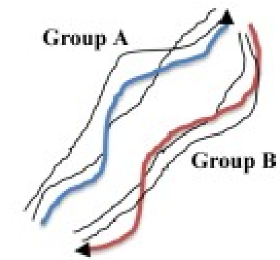
Spatial Tollerance
Temporal Tollerance
Spatio temporal distance



Pruning

Groups with a small Number of trips are Pruned

Support Threshold



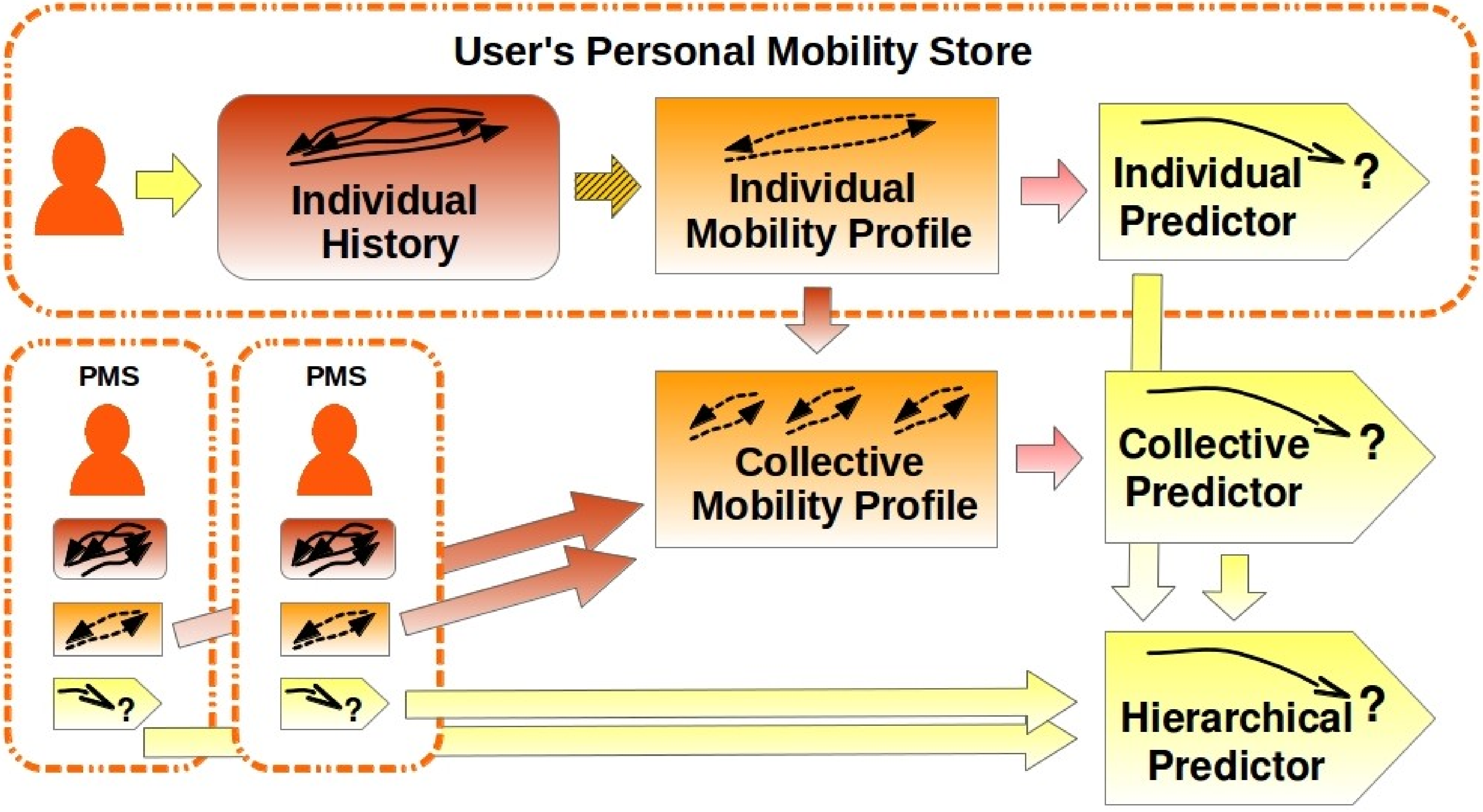
Profile extraction

The **medoid** of each group becomes user's **routines** and the all set become the user's **mobility profile**

Trasarti, Pinelli, Nanni, Giannotti.
Mining mobility user profiles for car pooling. ACM SIGKDD 2011

Idea in a nutshell

Use the mobility profile to predict the user's movements. If it is not able to produce a prediction, a collective predictor is used.
The collective predictor is built using the mobility profiles of the crowd.



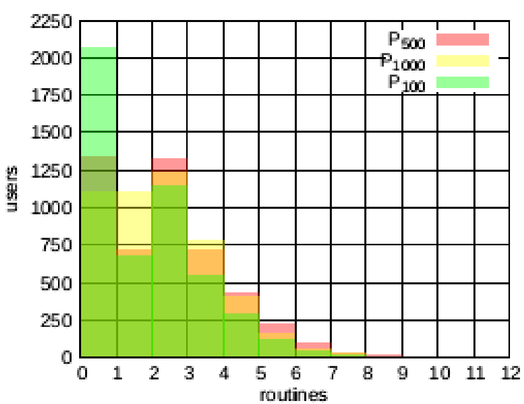
Experimental setting

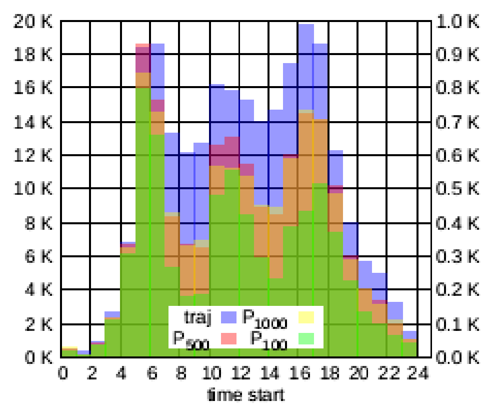
Starting from a dataset of 1 month of movements, 5.000 users and 326.000 trajectories. We divided the training set, i.e. 3 weeks and as test set the remaining last week.

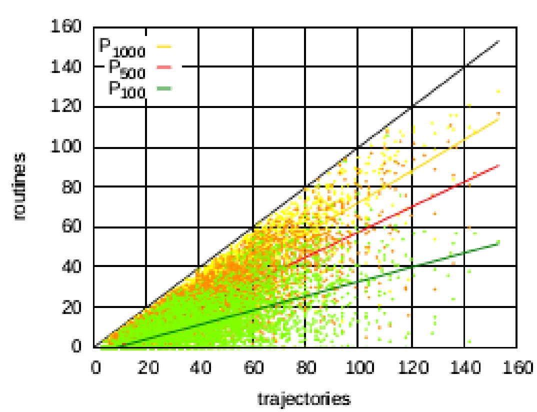
The trajectories in the test set are cut to become the queries for the predictor. The cuts tested are taking the first 33% or 66% of the trajectories.

Extracting the Mobility Profiles

The first step is to extract the mobility profiles from the training set. In order to assess the quality of them an empirical analysis is performed.





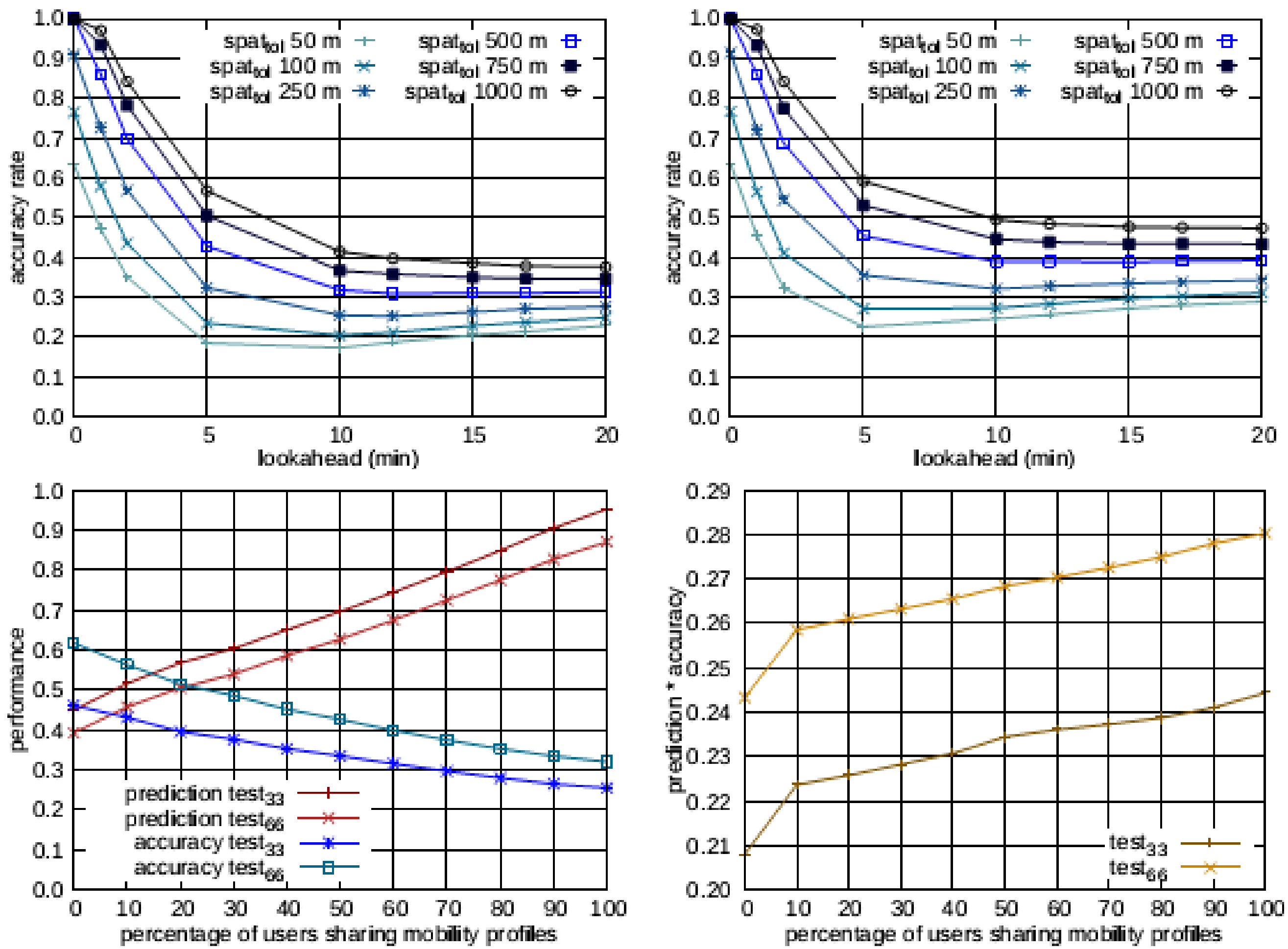


Routines per user distribution (left), trajectories and routines time start distribution (right) and the dataset coverage (bottom)

Results

MyWay obtains good results which are comparable to a global predictor built on top of the whole set of trajectories.

Test 33%



Test 66%

proactive car pooling

Project ICON

Carpooling cycle
Context

- Several initiatives, especially on the web

Carpooling cycle

Distinctive features

Traditional approach	vs.	ICON cycle
Users manually insert and update their rides	→	System autonomously detect systematic trips
Users search and contact candidate pals	→	System automatically suggest pairings
Users make individual, “local” choice	→	System seeks globally optimal allocation

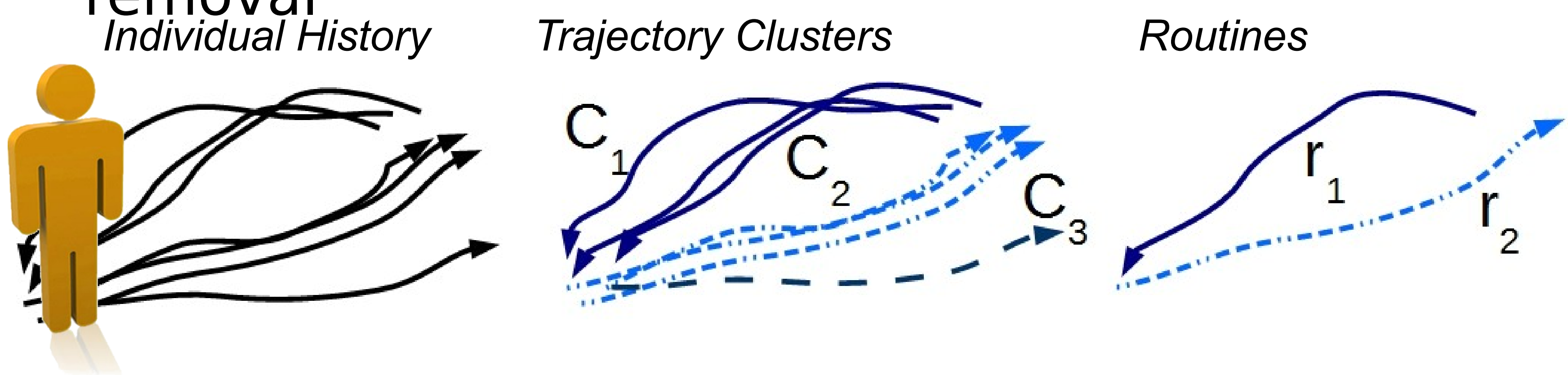
Carpooling cycle
Assumptions

- Users provide access to their mobility traces

Carpooling cycle

Step 1: Inferring Individual Systematic Mobility

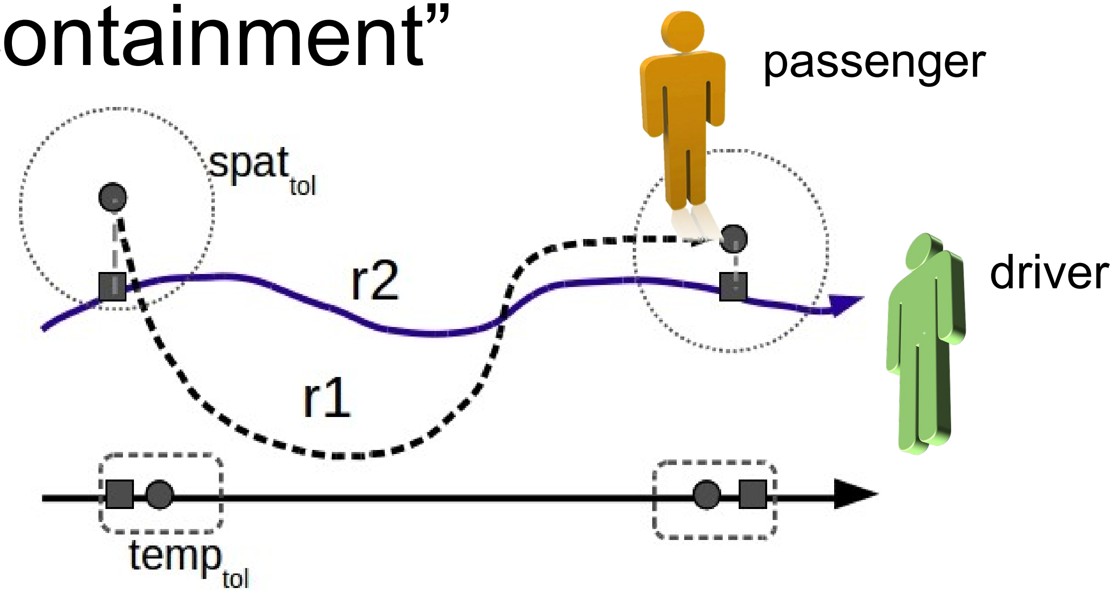
- Extraction of Mobility Profiles
  - Describes an abstraction in space and time of the systematic movements of a user.
  - Exceptional movements are completely ignored.
  - Based on trajectory clustering with noise removal



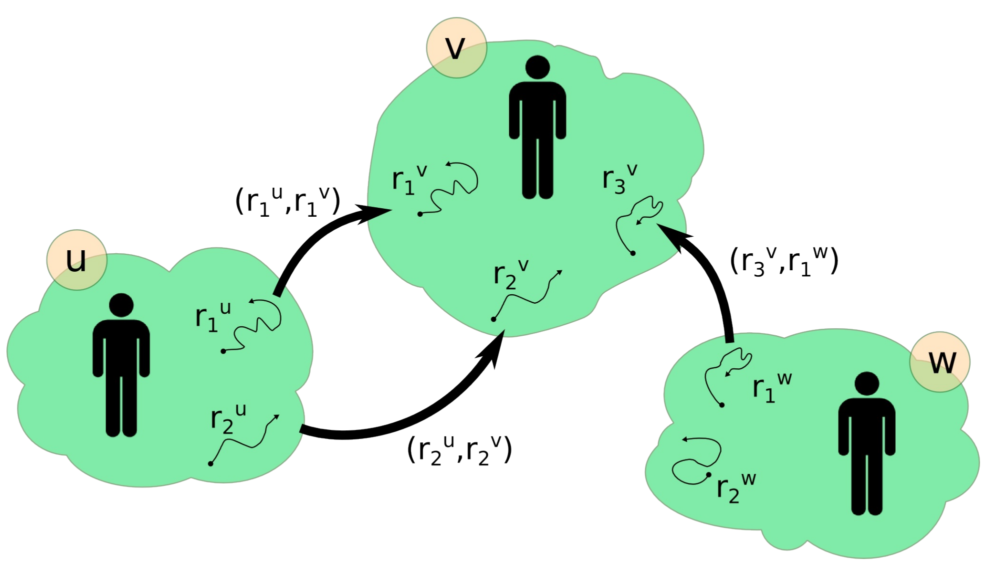
Carpooling cycle

Step 2: Build Network of possible carpool matches

- Based on "routine containment"
  - One user can pick up the other along his trip



- Carpooling network
  - Nodes = users
  - Edges = pairs of users with matching routines



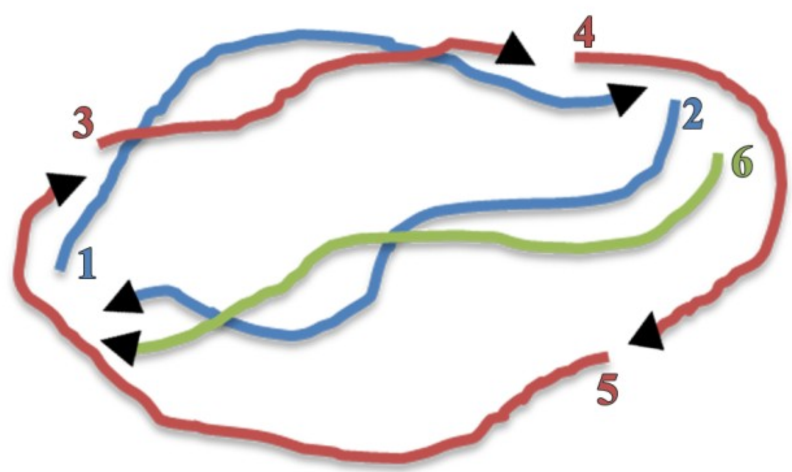
Application: Car pooling

Pro-active suggestions of sharing rides opportunities without the need for the user to explicitly specify the trips of interest.

Matching two routines:

$$contained(T_1, T_2, th^{walking}_{distance}, th^{wasting}_{time}) \equiv \exists i, j \in \mathcal{N} \mid 0 < i \leq j \leq m \wedge Dist(p^1_1, p^2_i) + Dist(p^1_n, p^2_j) \leq th^{walking}_{distance} \wedge Dur(p^1_1, p^2_i) + Dur(p^1_n, p^2_j) \leq th^{wasting}_{time}$$

Mobility profile share-ability:



	1	2	3	4	5	6
1	-	-	F	F	F	F
2	-	-	F	F	F	**T**
3	**T**	F	-	-	-	F
4	F	F	-	-	-	F
5	F	F	-	-	-	F
6	F	**T**	F	F	F	-

	-	0	1/2
	1/3	-	0
	1	0	-

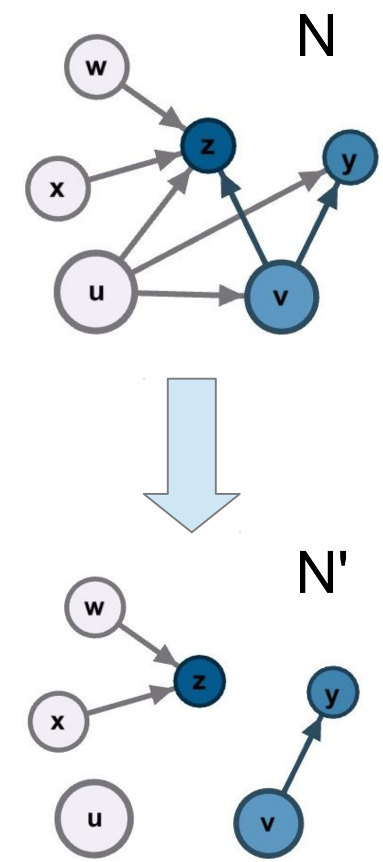
mobility profiles $\tilde{T}_1$ *and* $\tilde{T}_2$

$$profileShare(\tilde{T}_1, \tilde{T}_2, th^{walking}_{distance}, th^{wasting}_{time}) = \frac{\left| \left\{ p \in \tilde{T}_1 \mid \exists q \in \tilde{T}_2 . Share(p, q, th^{walking}_{distance}, th^{wasting}_{time}) \right\} \right|}{\mid \tilde{T}_1 \mid}$$

Carpooling cycle

Step 3: Optimal allocation of drivers-passengers

- Given a Carpooling Network N, select a subset of edges that minimizes |S|
    - S = set of circulating vehicles

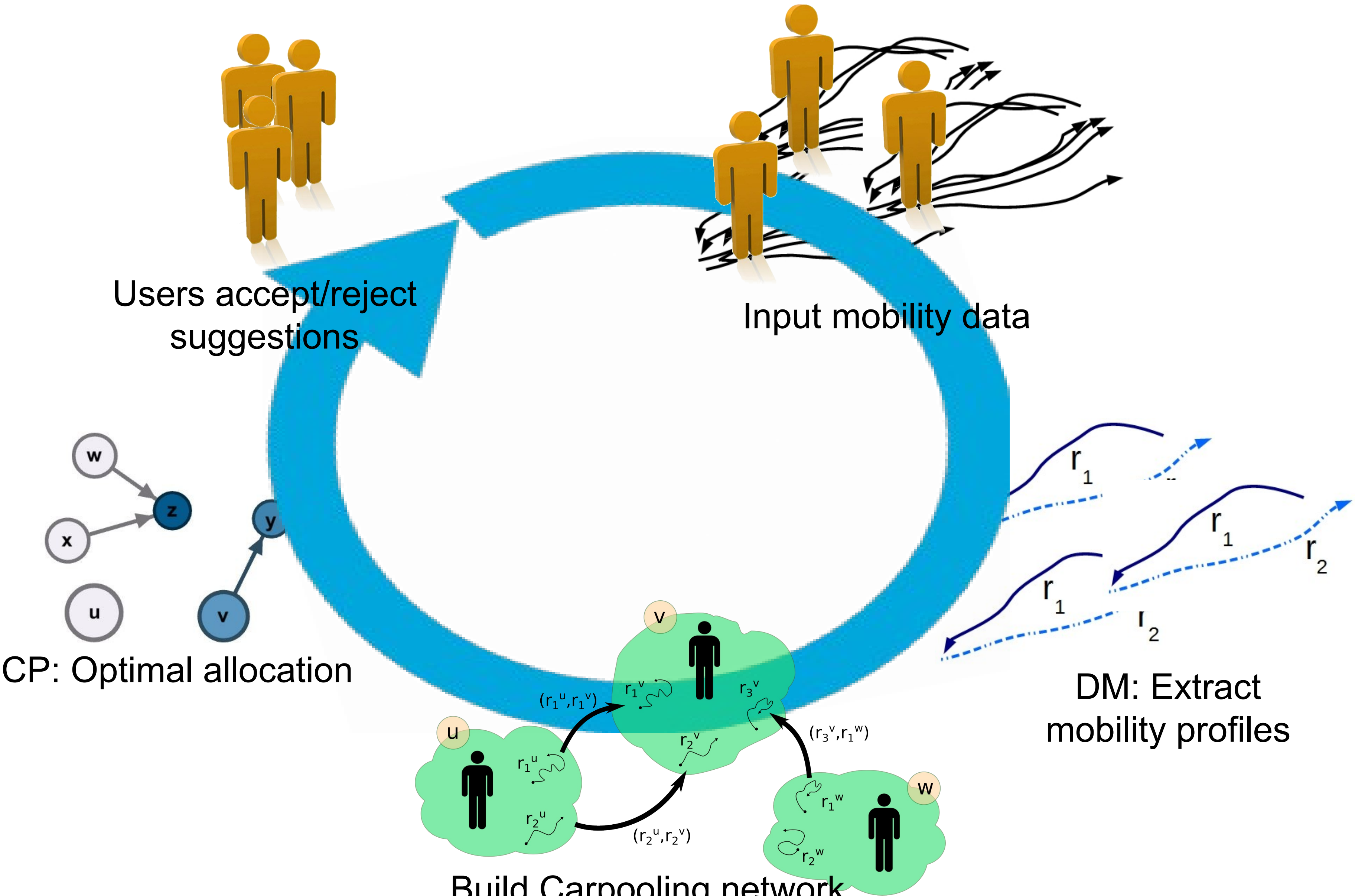


provided that the edges are coherent, i.e.:

- indegree(n)=0 OR outdegree(n)=0 (a driver cannot be a passenger)
- indegree(n) ≤ capacity(n)

Carpooling cycle

The "simple" ICON Loop



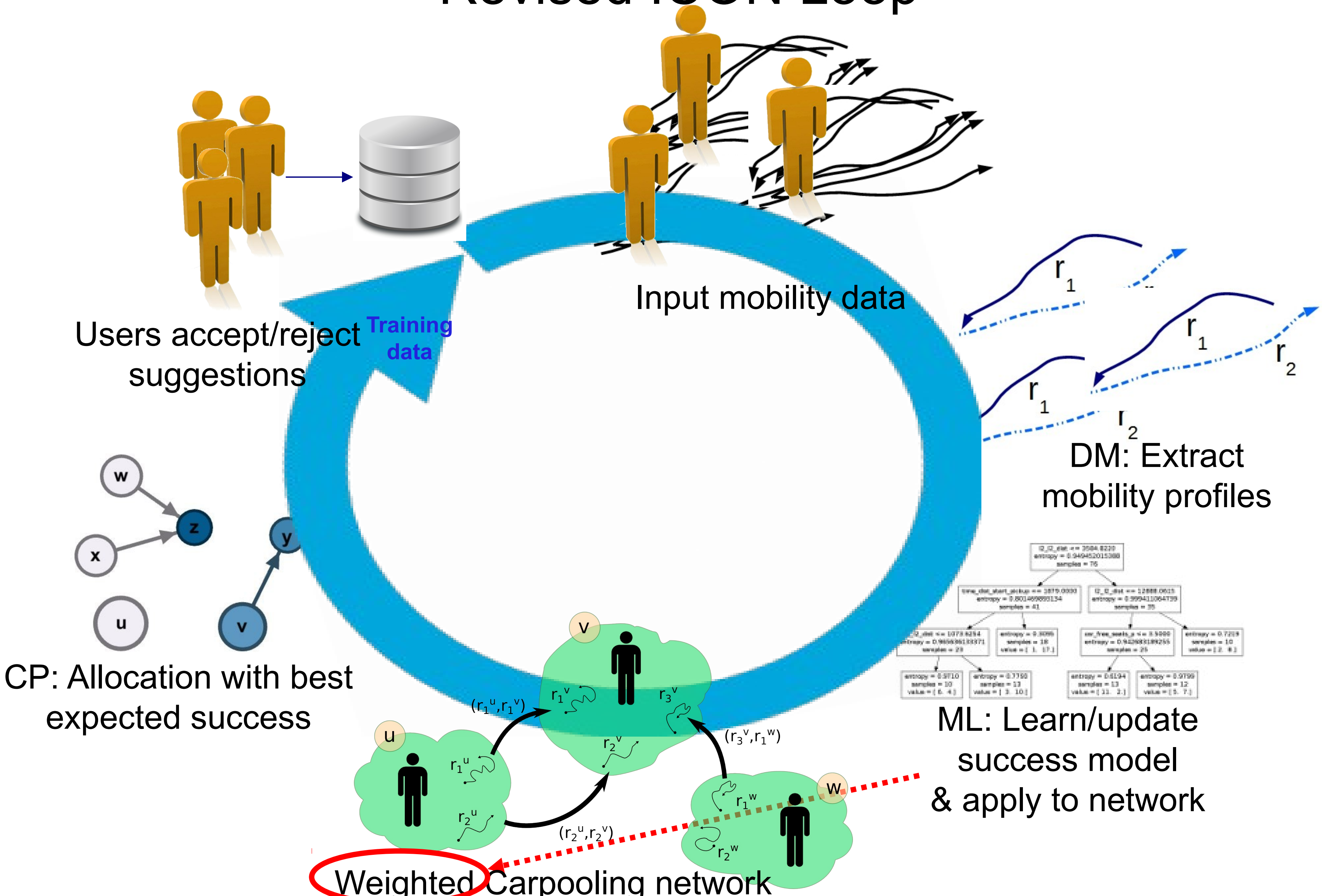
Carpooling cycle

Improvement

- In carpooling (especially if proactive) users might not like the suggested matches
  - Impossible to know who will accept a given match
  - Modeling acceptance might improve results
- Two new components
  - **Learning** mechanism to guess success probability of a carpooling match
  - **Optimization** task exploits it to offer solution with best <u>expected</u> overall success

Carpooling cycle

Revised ICON Loop



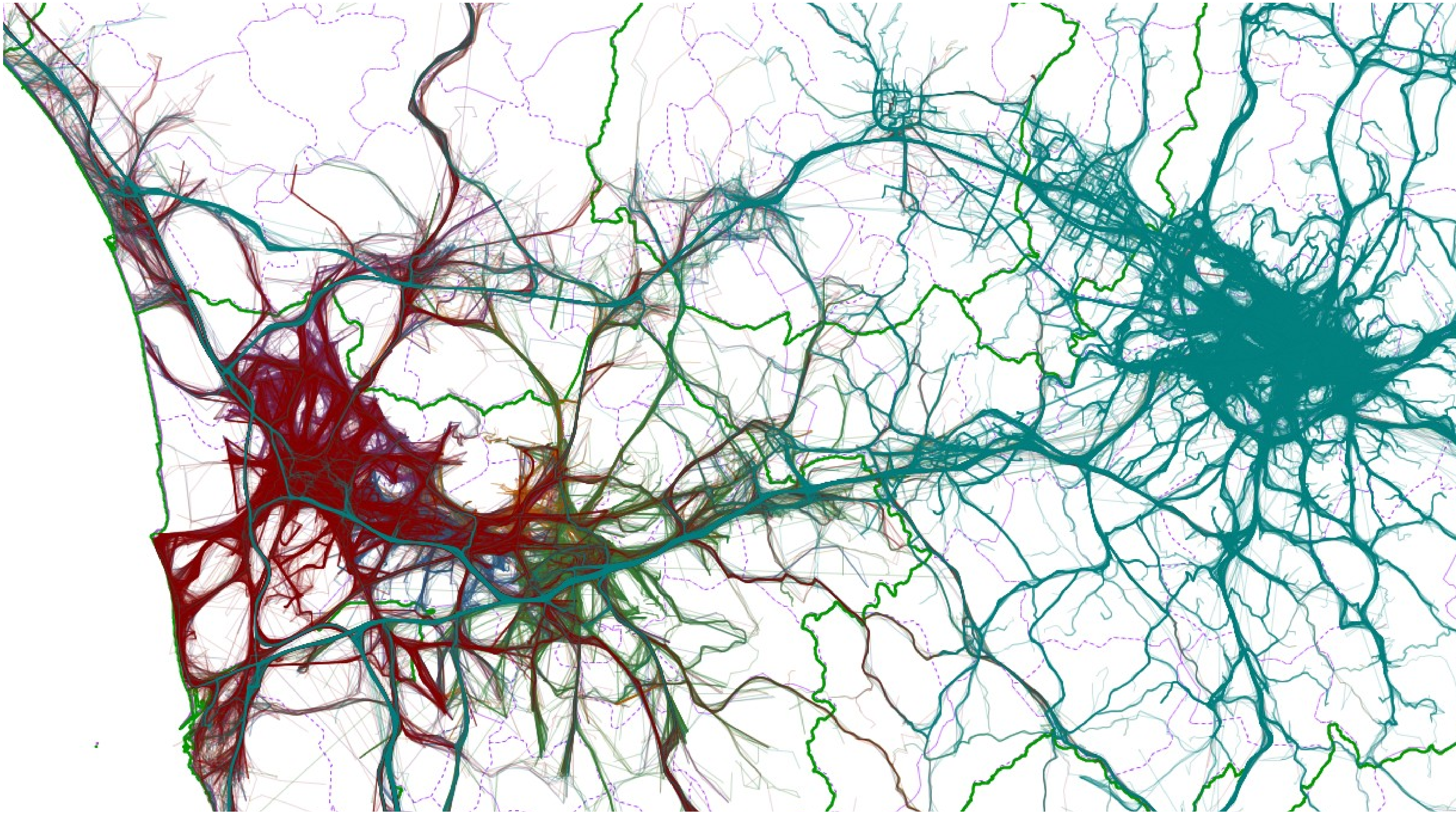
Networks as a mining tool

S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, F. Giannotti

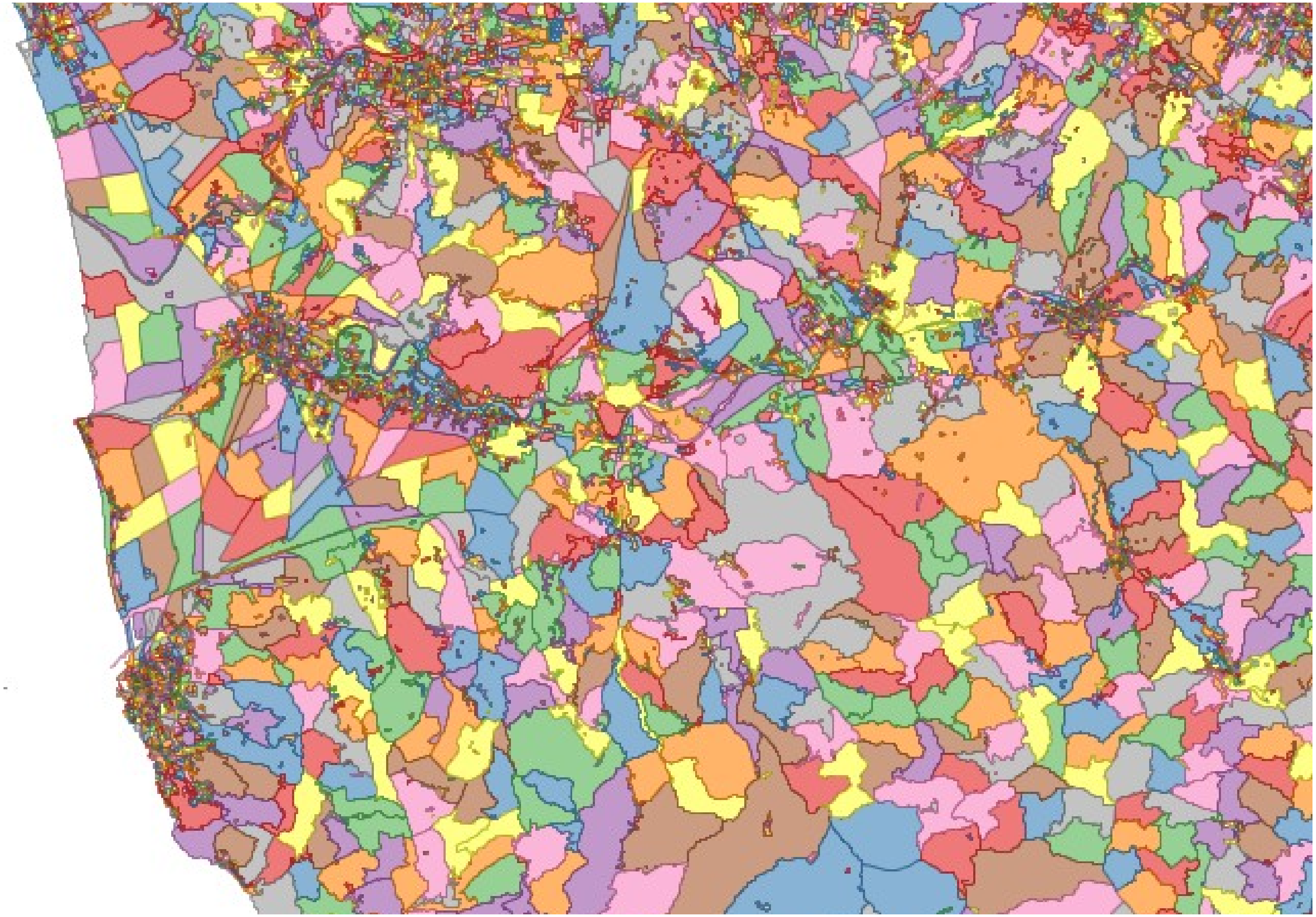
Discovering the Geographical Borders of Human Mobility

KI - Künstliche Intelligenz, 2012.

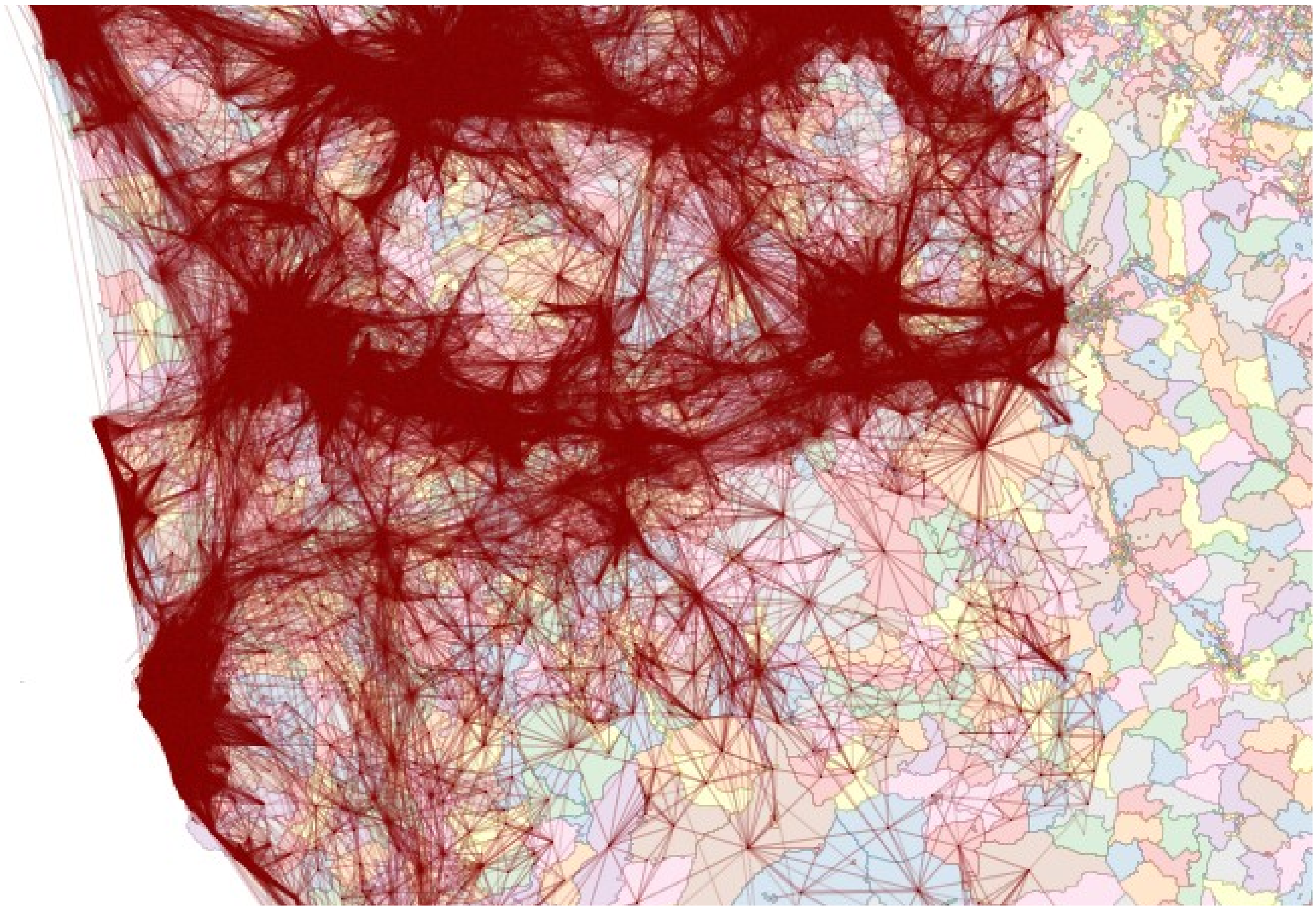
Mobility coverages



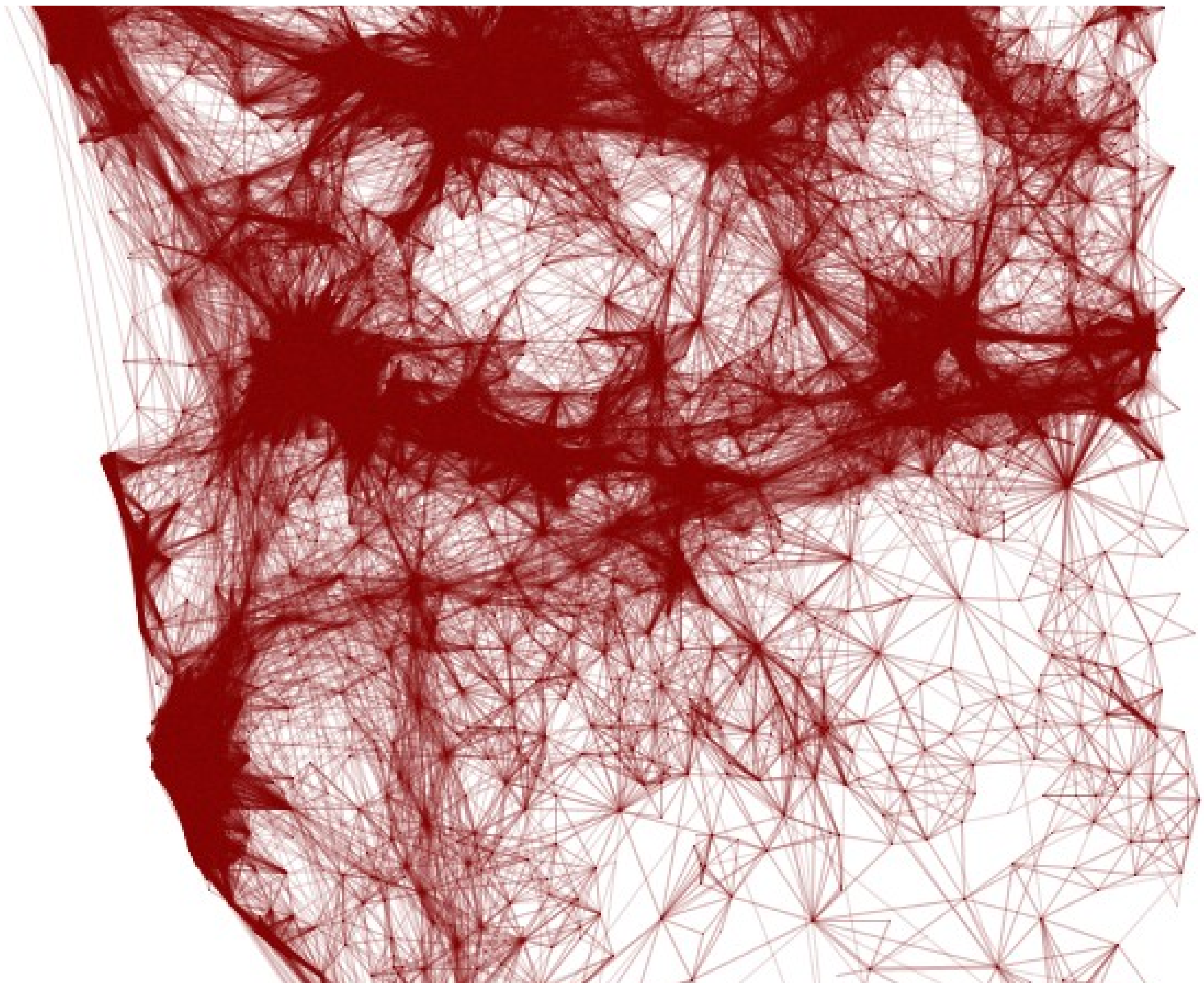
Step 1: spatial regions



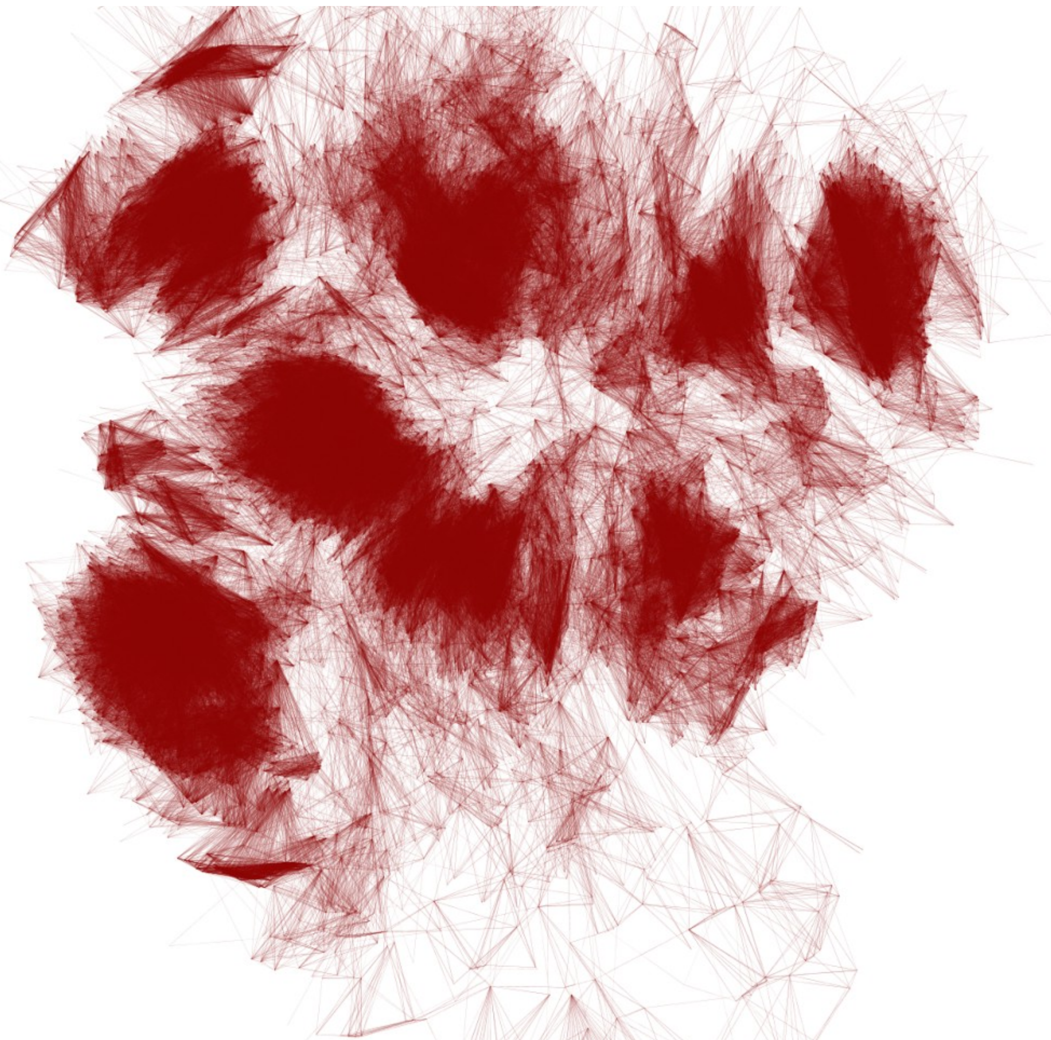
Step 2: evaluate flows among regions



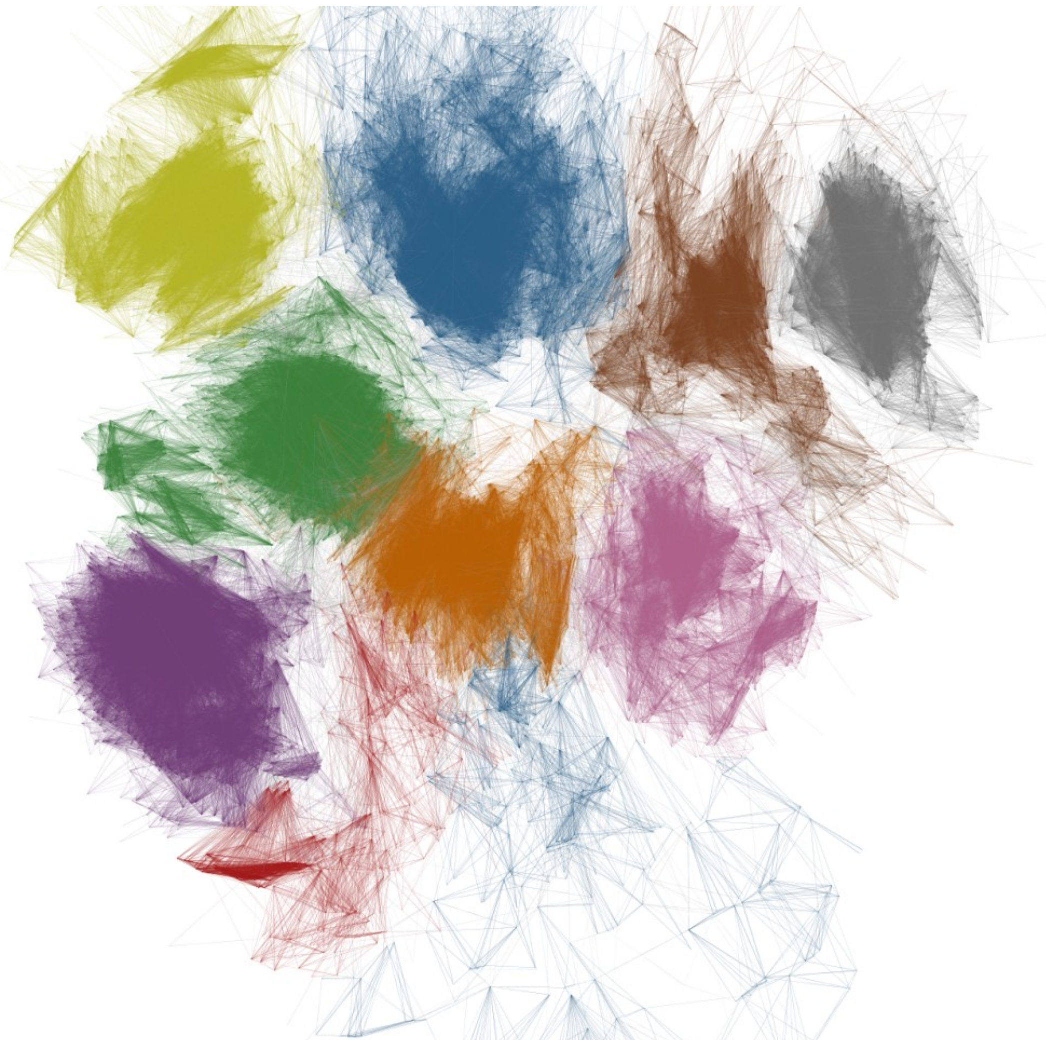
Step 3: forget geography



Step 4: perform community detection

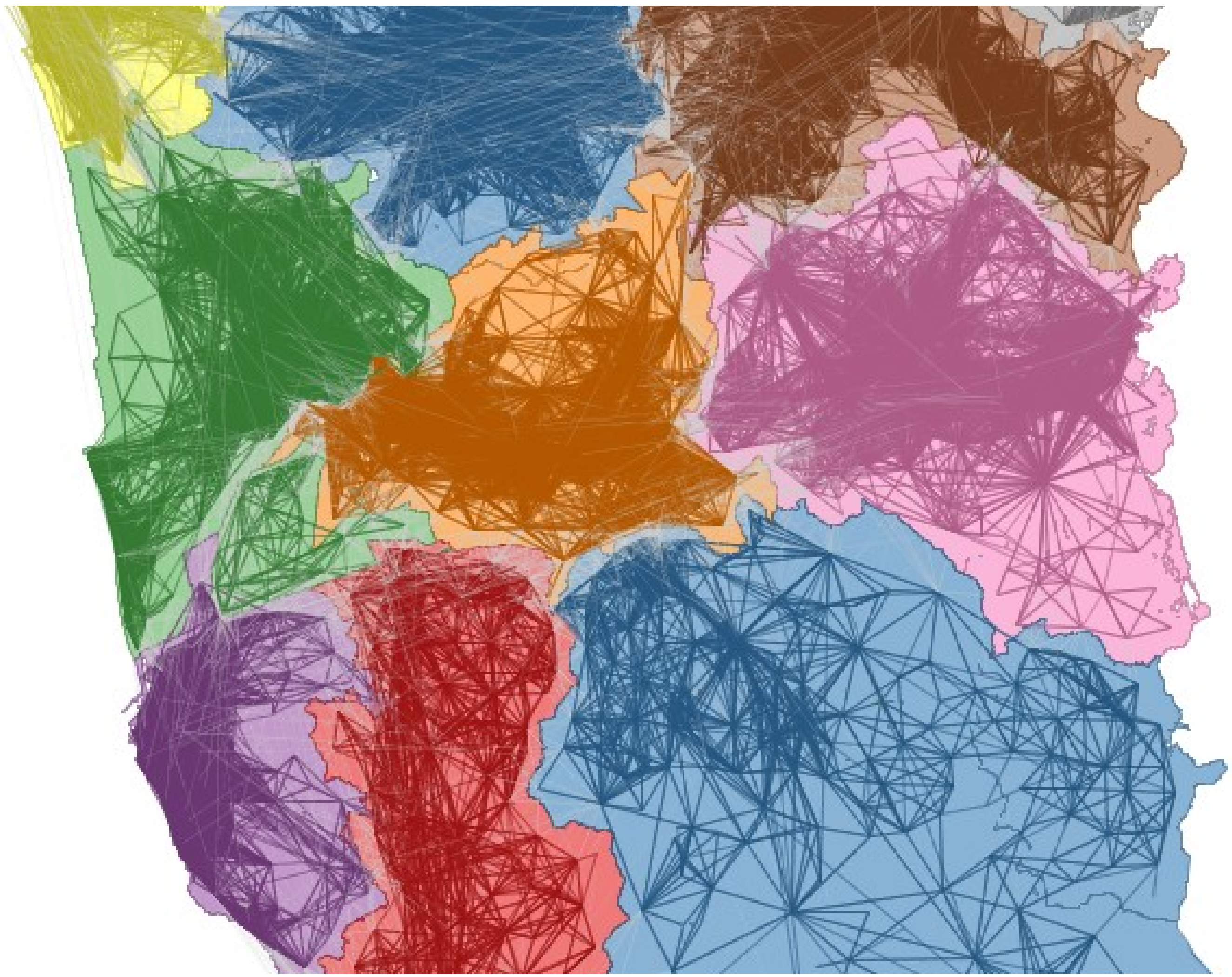


Step 4: perform community detection

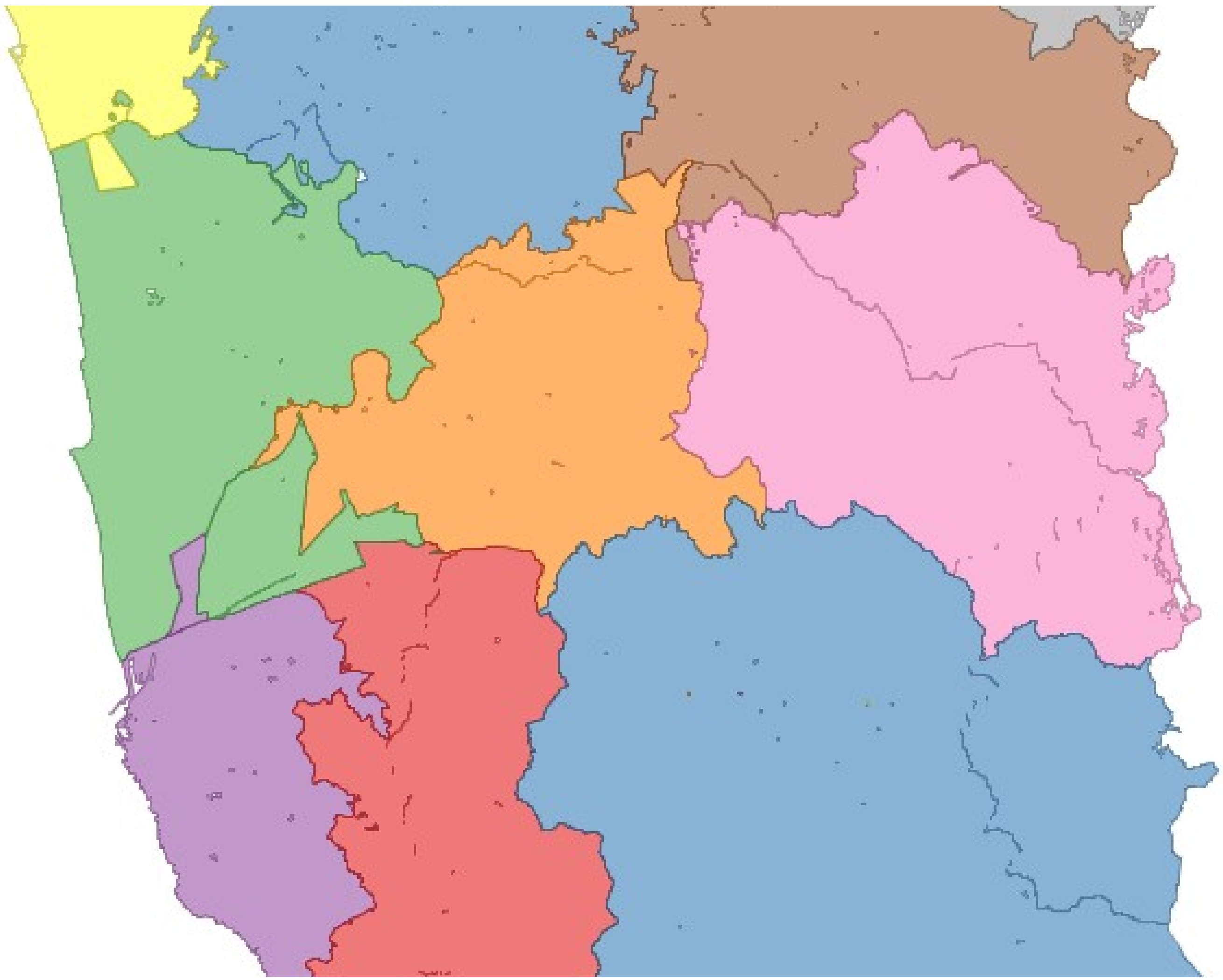


Step 5: map back to geography

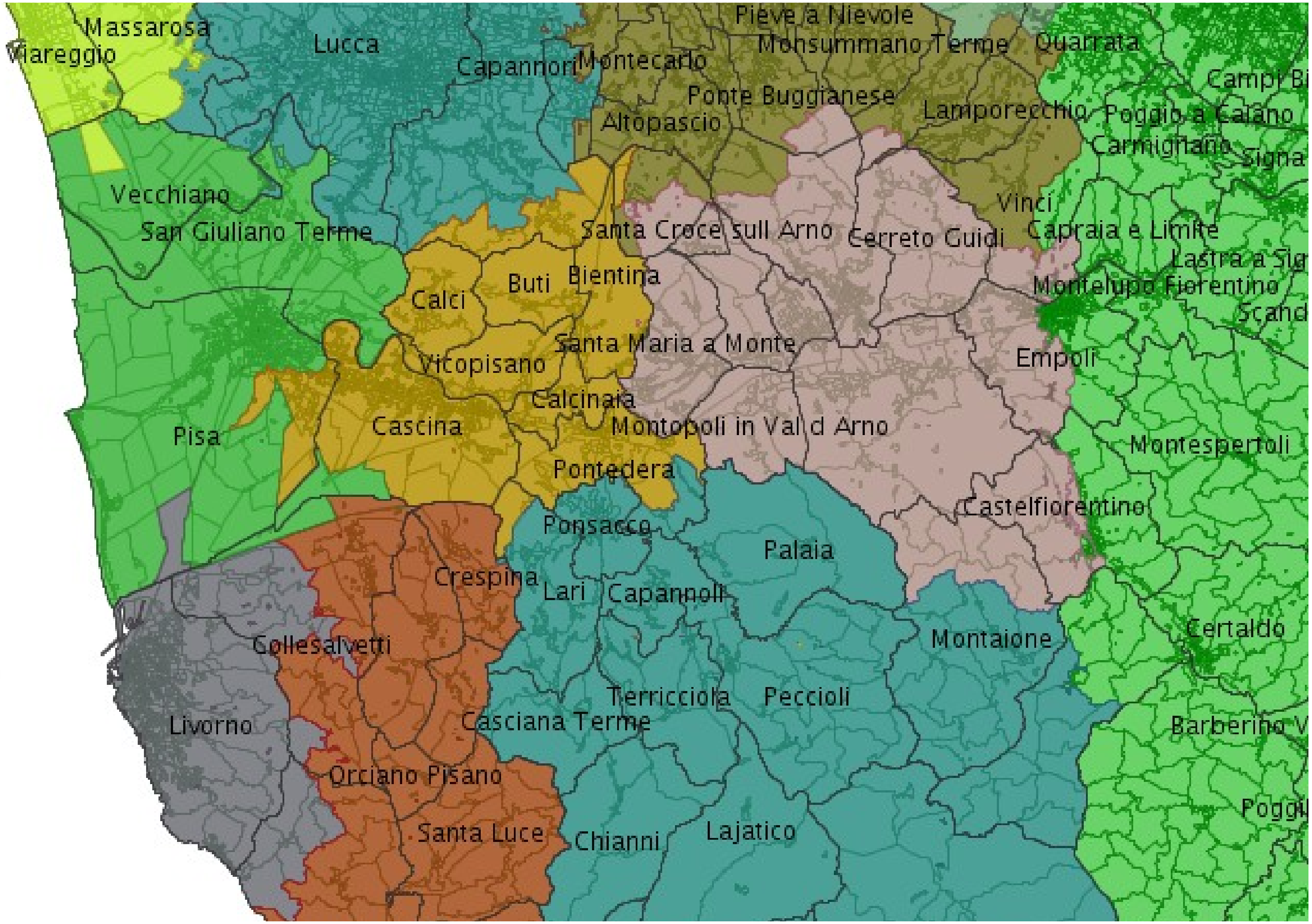
Step 6: draw borders



Final result

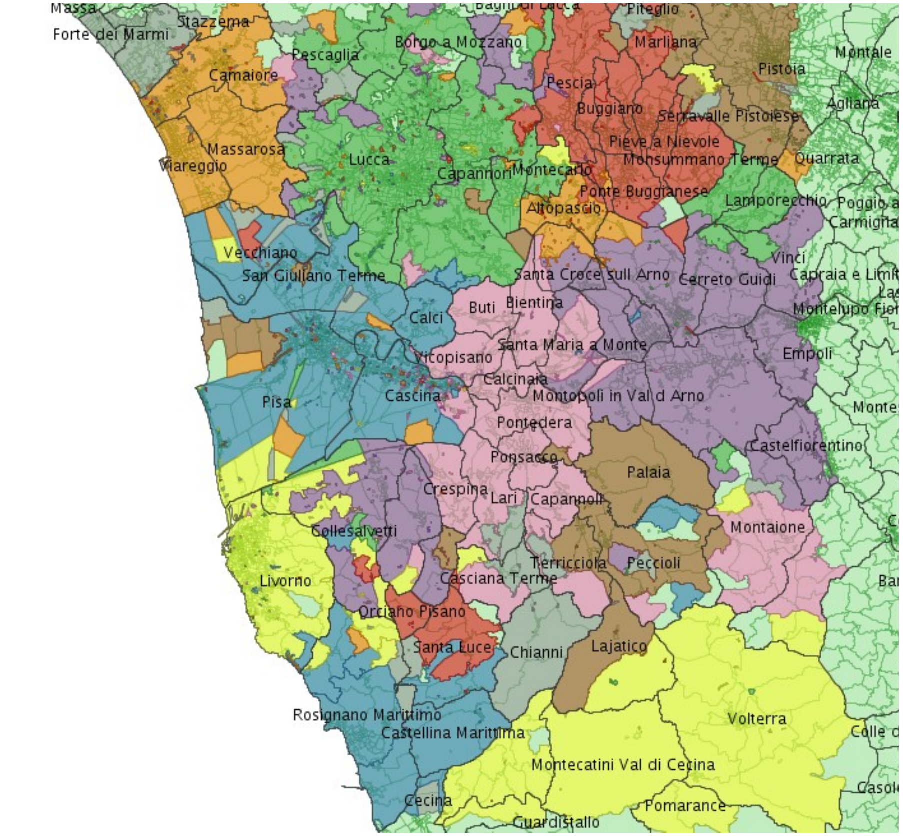


Final result: compare with municipality borders



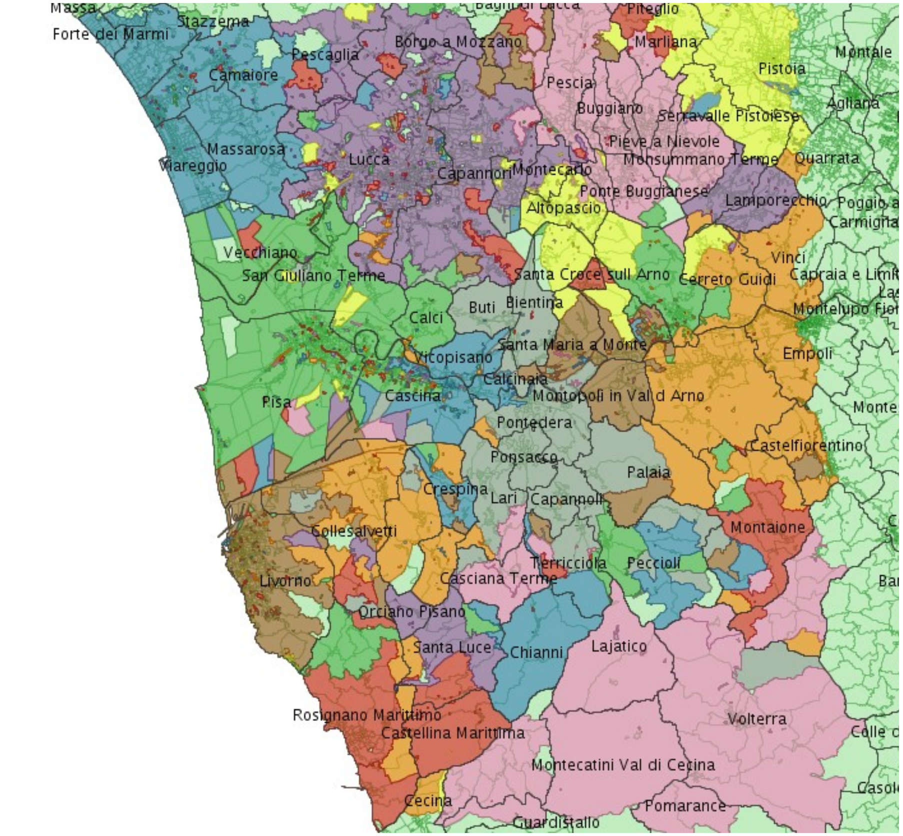
Borders in different time periods

Only weekdays movements



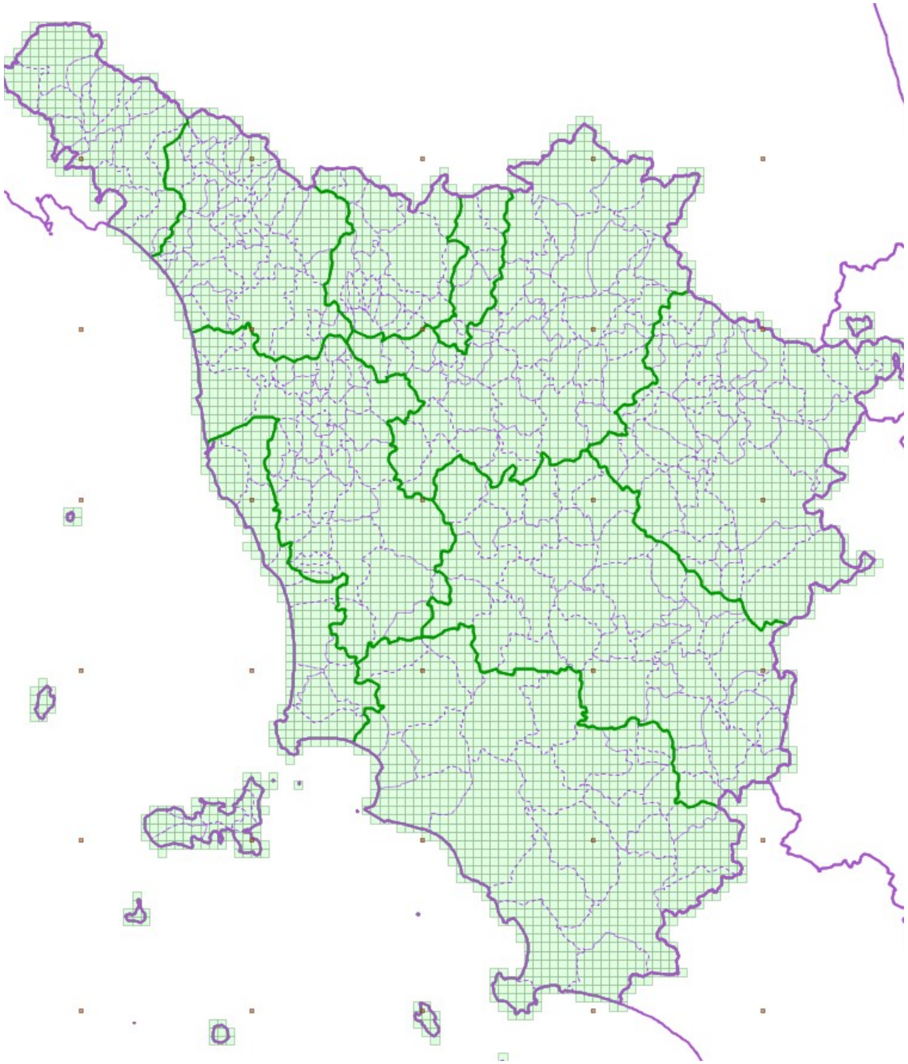
Similar to global clustering: strong influence of systematic movements

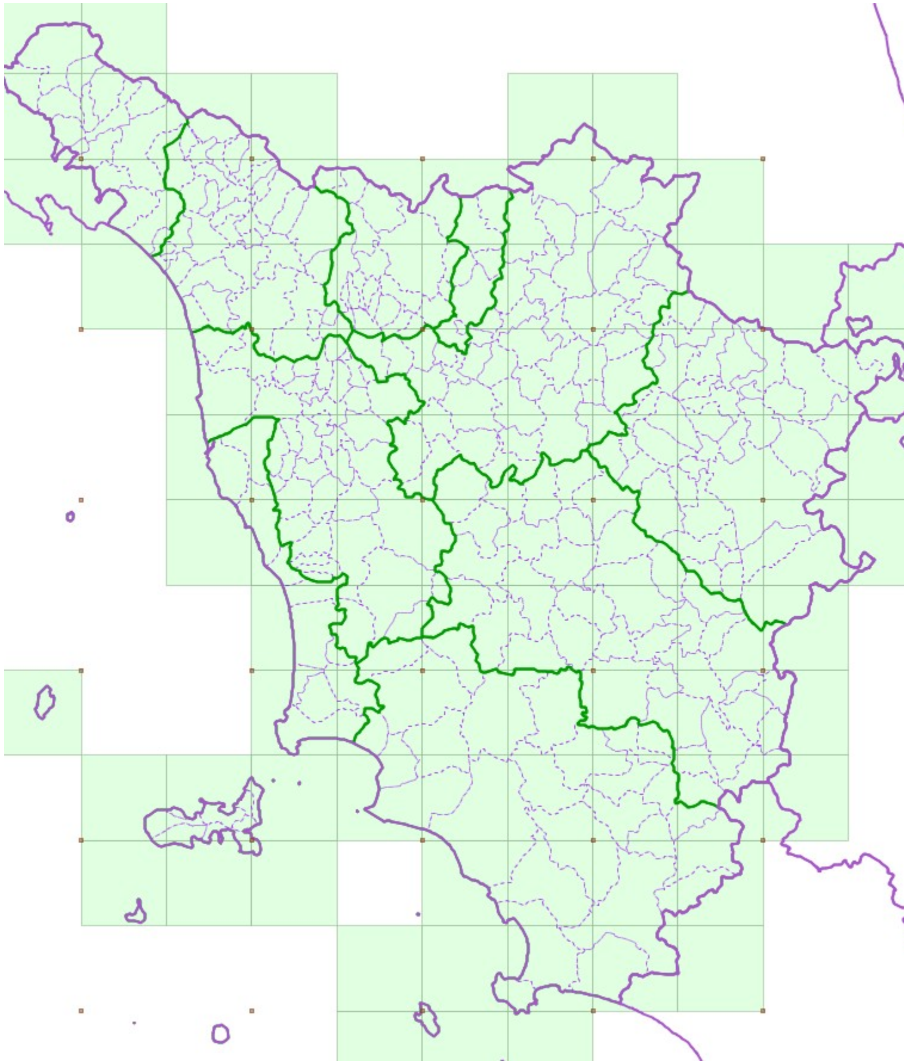
Only weekend movements



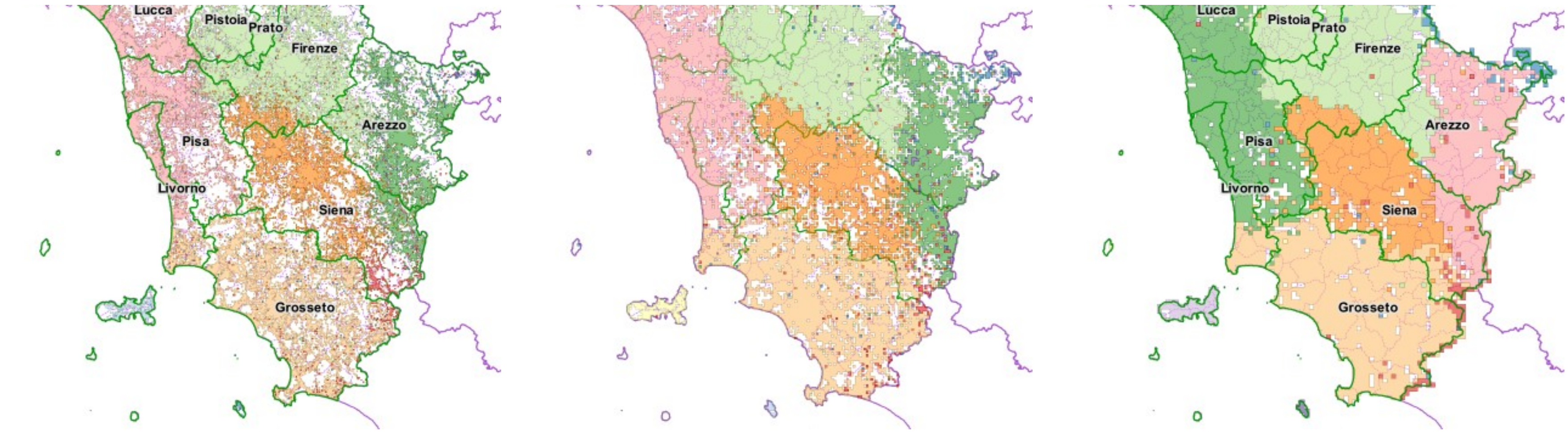
Strong fragmentation: the influence of systematic movements (home-work) is missing

Borders at regional scale

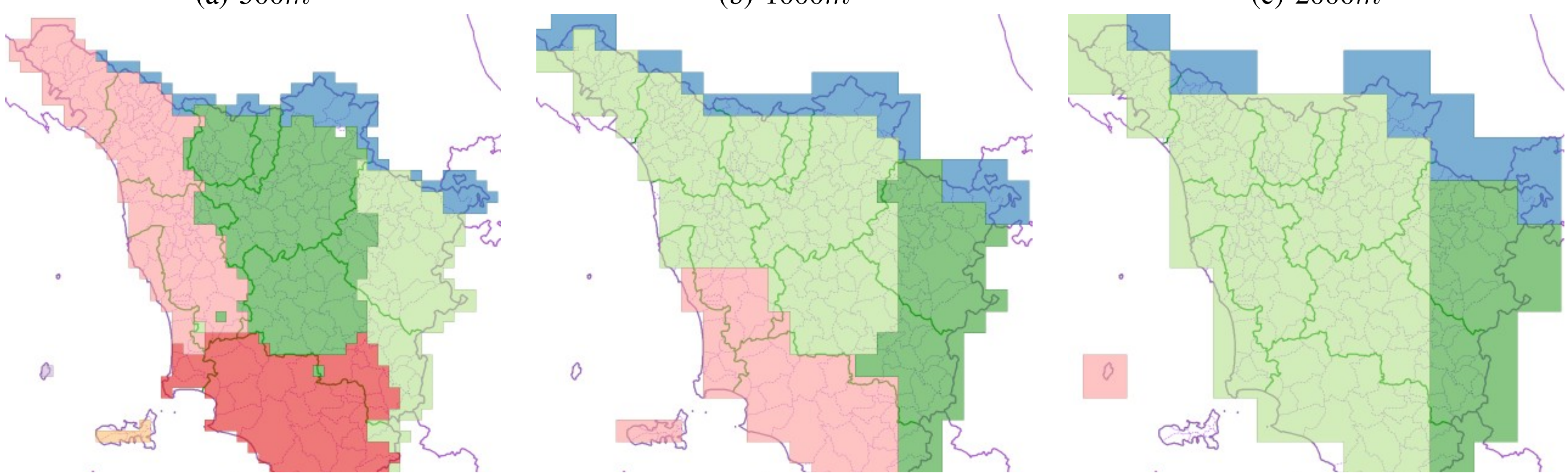




Final results



(a) $500m$ (b) $1000m$ (c) $2000m$



Comparison with “new provinces”