Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Informatik
Computer Science

Distributed Systems Seminar – Spring 2012

Eigenbehaviors: identifying structure in routine

Nathan Eagle & Alex Sandy Pentland (MIT Media Lab, 2009)

Presented by:

César Fuentes (ETH D-INFK)

Abstract

- From longitudinal data → identify structure inherent in daily behavior
- Represent structure: principal components, set of characteristics vectors → **“eigenbehaviors”**
- Approximations with the first few eigenbehaviors
- Used for:
  - Compact representation
  - Prediction
  - Infer community affiliations

Past challenges & Motivation

- Repeating & identifiable routines in people's lives
  - More apparent when behavior is contextualized → time, space, social circle
- Before: lack of contextualized behavioral data → NOW: smart phones data
- Traditional methods (e.g. Markov models) cannot manage temporal patterns across different timescales.
- New method: Principal Component Analysis

Applications

- Compact representation
  - 90% accuracy with 6 primary eigenbehaviors
- Prediction
  - If first 12h of a day's activities are known, the last 12h can be predicted with ~79% accuracy
- Characterization of groups
  - Groups of friends have collective "behavior space"
- Identification of affiliations and similarities
  - Using the Euclidean distance between individual behavior and a community's behavior subspace

Related work

- CSCW: Techniques of rhythm modeling within the workspace (Begole et al.) → last week
- Electronic badges → 80's, early 90's
  - location-based applications, detection of face-to-face interactions
- GPS → location detection & classification (but not indoors)
- Correlating cell tower ID with a user's location
- Pattern recognition, computer vision
  - "**Eigenfaces**" → many analogies in characterization of individuals
  - Also: new technologies provide wealth of training data

Data Source: Reality Mining Dataset

100 subjects @ MIT during 2004-2005 academic year

- **75 lab students/faculty**
  - 20 incoming masters
  - 5 incoming freshmen
- **25 business school students**

100 **Nokia 6600** smartphones, with **“Context”** app.
(http://www.cs.helsinki.fi/group/context/)

Call logs

Bluetooth devices in proximity

Cell tower IDs (location)

Application usage

Phone status

$\underbrace{\hspace{10em}}_{\sim\ 400\ 000\ \text{h of data}}$

Limitations and concerns

- Justifiable **privacy** concerns
  - Legitimate, but NOT addressed in this work
  - Dataset from social experiment, with consent of subjects
- Techniques not only applicable to humans → animal behavior studies
  - Prediction can be actually more accurate (animals less “inventive”)
- Subjects in the RM study may not be a representative sample of society, but…
  - Regularity in routines is normal for everyone

Limitations and concerns

- Justifiable **privacy** concerns

Underlying assumptions

•Similarity of behaviors across time → predictability

•Similarity of different individuals' behaviors within the same social group → homophily

•Can be defeated with unexpected behavior (spontaneity)

•But good enough for most cases…

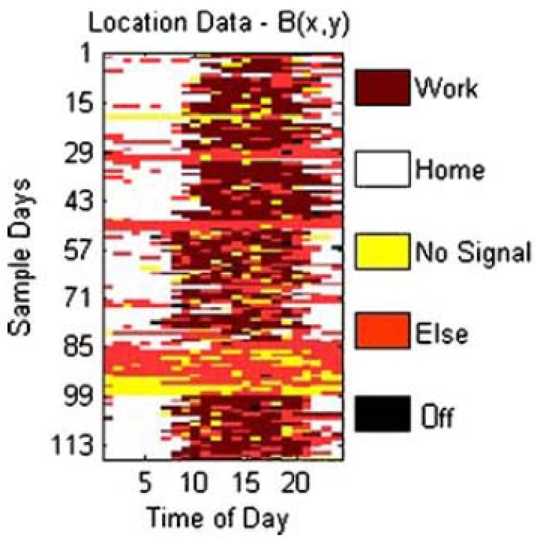
sample of society, but…

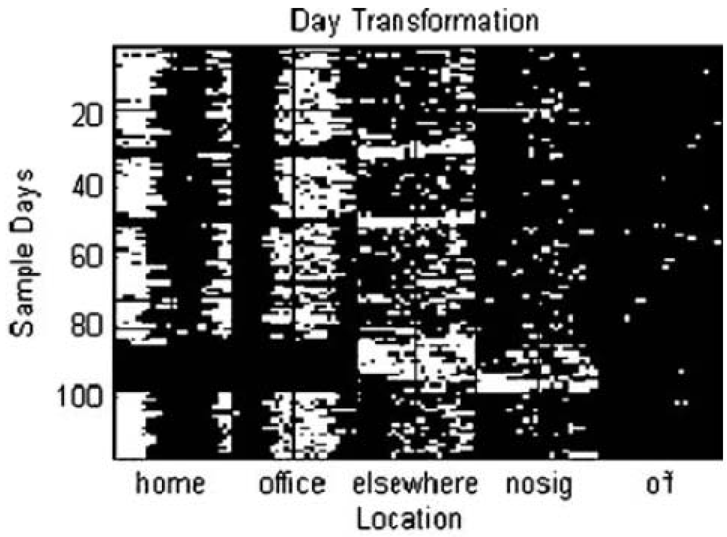
- Regularity in routines is normal for everyone

Data Modeling: Temporal Location Data

- Characterize person I as matrix B of size D×24
  - D → # of days in study; columns for 24h
- B contains *n* "location" labels = {*Home, Elsewhere, Work, No Signal, Off*}
  - Labels obtained in previous work, here assumed as ground truth
- B → B' : matrix of D×H (H=24×*n*) binary values
- Days are not scattered across the 120-dim. space → they live in a low dimensional "behavior space"
  - Space defined by a subset of vector of dimension H

Data Modeling: Temporal Location Data





$$\mathbf{B} = \underbrace{\begin{bmatrix} 1 & 2 & \dots & 1 \\ 2 & 2 & & 1 \\ \vdots & \vdots & & \vdots \\ 5 & 4 & \dots & 3 \end{bmatrix}_{D\times 24}}_{\text{24 hours}} \Big\}\ \text{D days} \quad \Rightarrow \quad \mathbf{B}' = \underbrace{\begin{bmatrix} \boxed{10000 \quad 01000 \quad \dots \quad 10000} \\ 01000 \quad 01000 \qquad\quad 10000 \\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \\ 00001 \quad 00010 \quad \dots \quad 00100 \end{bmatrix}_{D\times H}}_{\text{120-dim. space}} \rightarrow \Gamma_i \in \{0;1\}^H$$

Eigenbehaviors for individuals

For each subject: set of behaviors

$$\Gamma_1, \Gamma_2, \ldots, \Gamma_D \in \{0;1\}^H$$

Average behavior of the individual

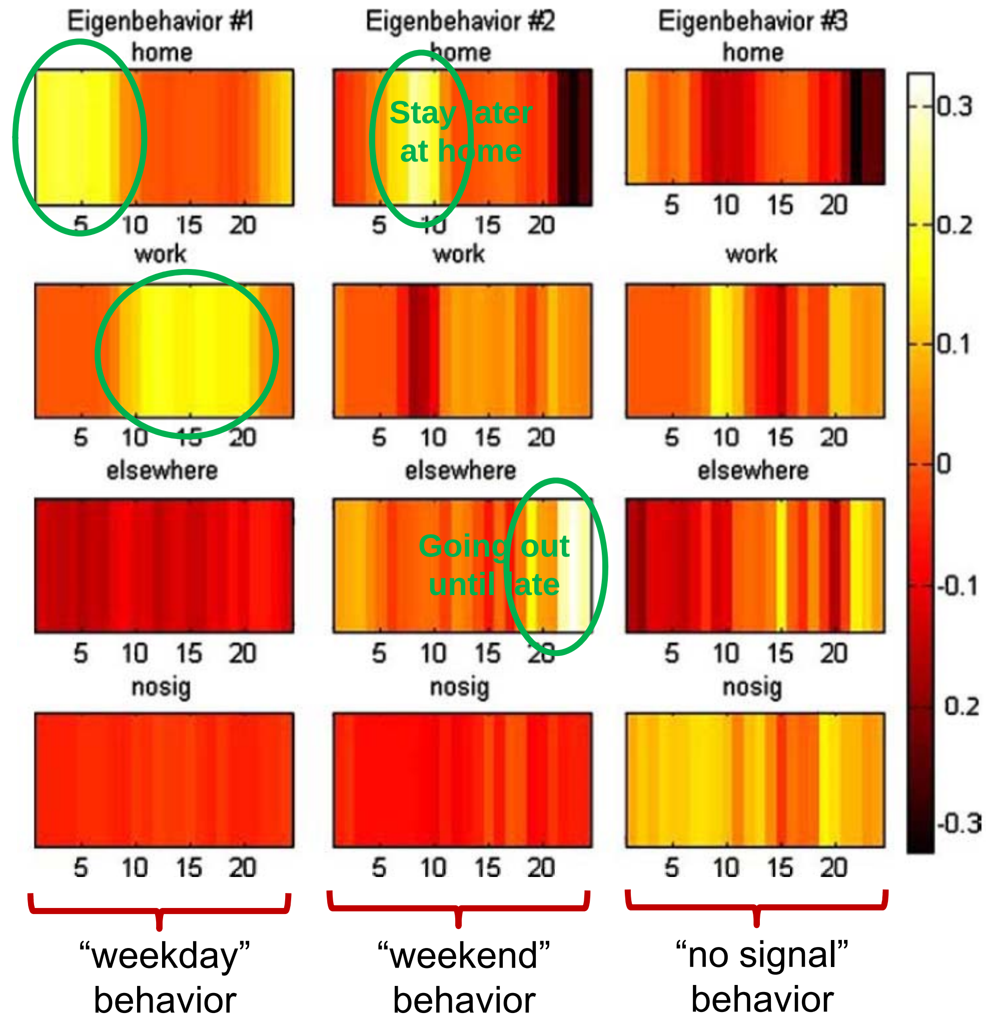
$$\Psi = \frac{1}{D}\sum_{n=1}^{D}\Gamma_n \qquad \Phi_i = \Gamma_i - \Psi$$

PCA on these vectors: eigenvectors of the covariance matrix

$$C = \frac{1}{H}\sum_{n=1}^{H}\Phi_n\Phi_n^T = AA^T$$

$$C = U\Lambda U^T$$

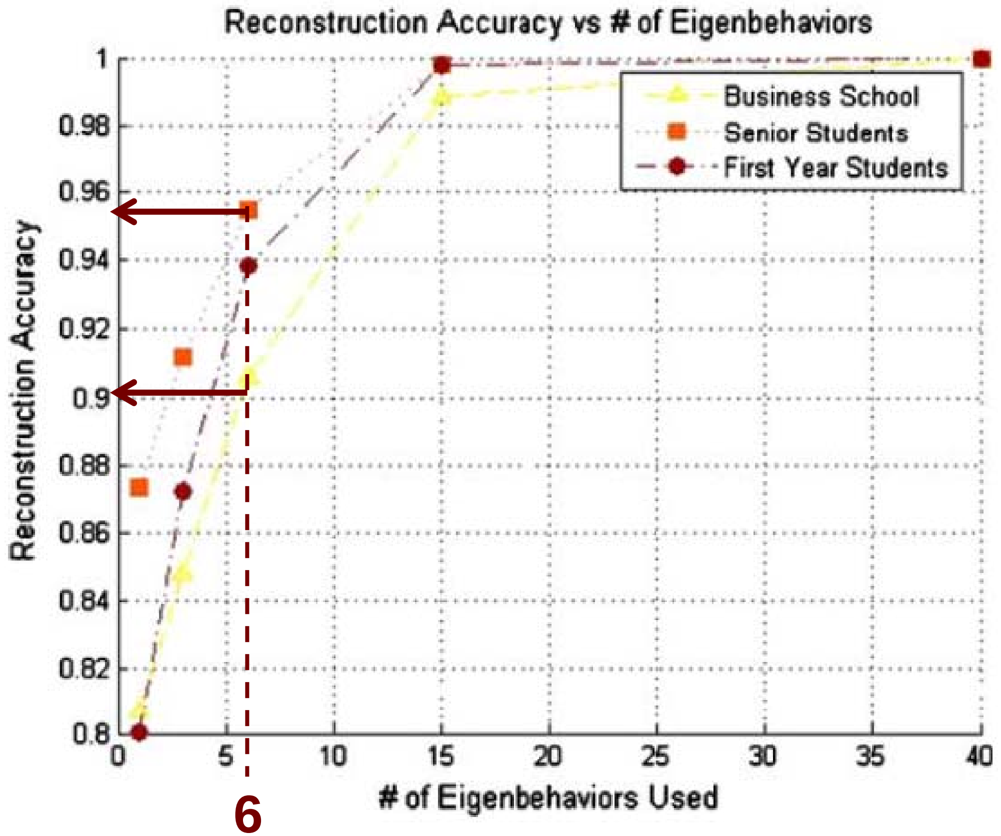
$$U = \underbrace{\begin{bmatrix} u_1 & u_2 & \ldots & u_H \end{bmatrix}}_{\text{Keep 6 largest eigenbehaviors}}$$



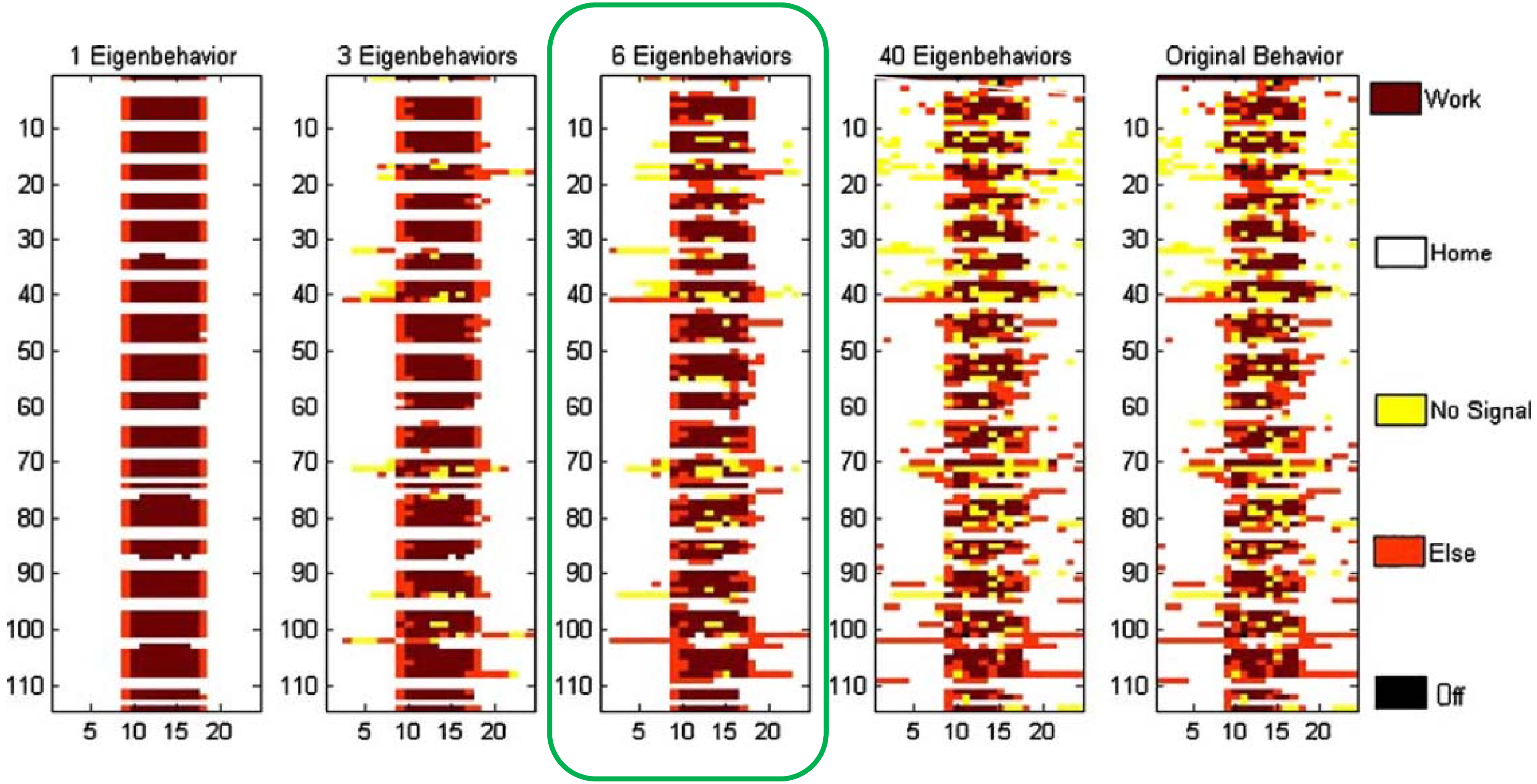
Eigenbehaviors for individuals

- How many eigenbehaviors to keep?

Senior lab students behave more regularly than business school students!

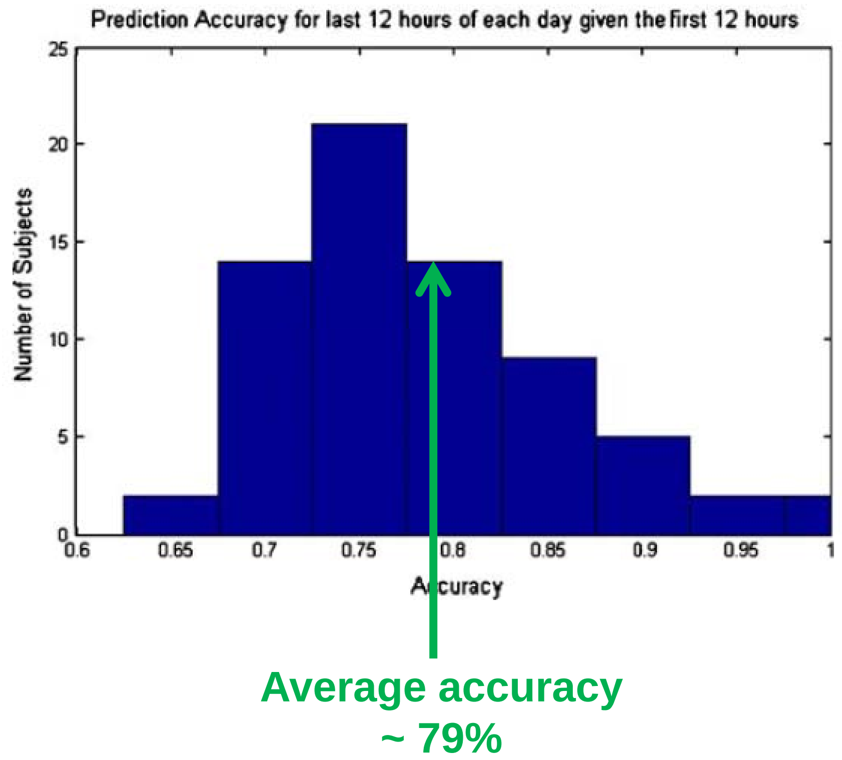


Eigenbehaviors for individuals



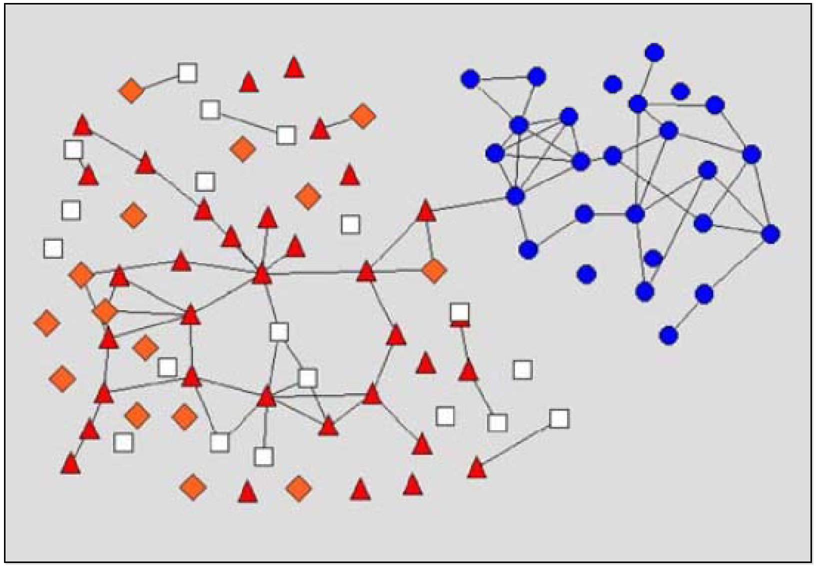
Prediction of an individual's behavior

- For each subject, calculate behavior space with:
  - Individual's 6 primary eigenbehaviors
  - Weights from first 12h of the day
- Linear combination of weights and primary eigenbehaviors → vector of predicted locations created
- (mechanism is similar to a recommender system)



Eigenbehaviors for social networks

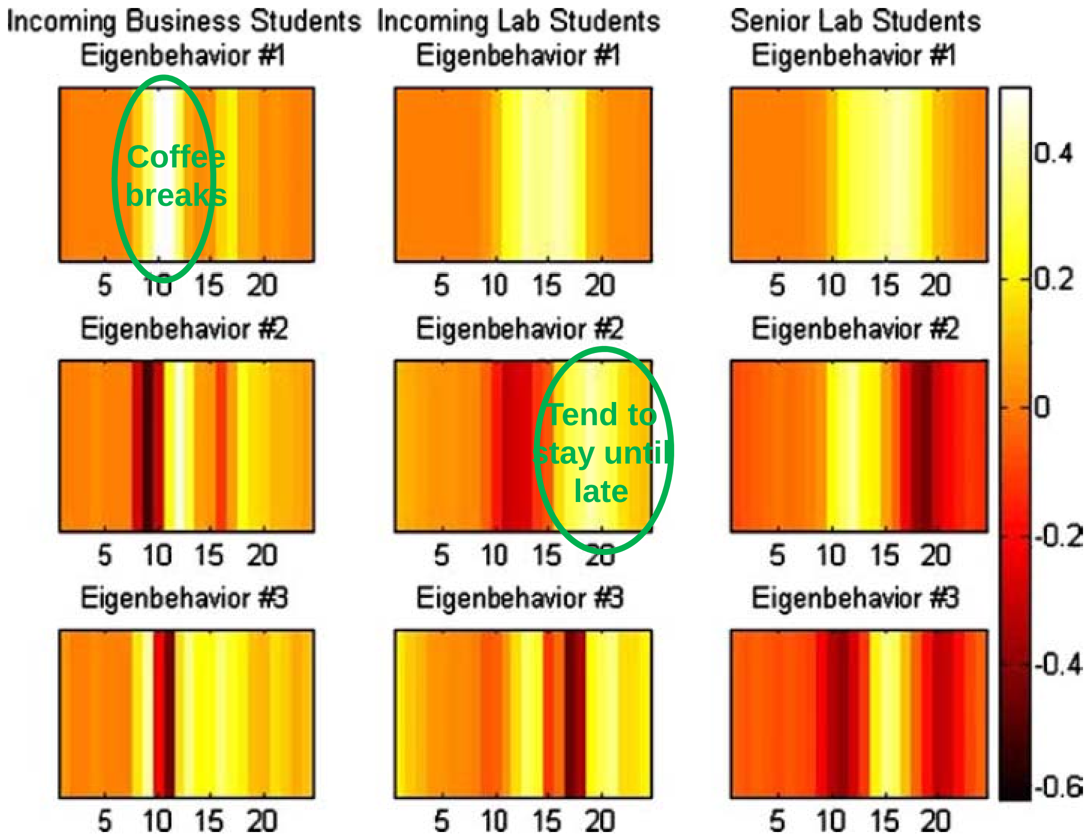
- Goal: infer relationships & affiliations from comparison of eigenbehaviors.
- RM social network: high amount of clustering
  - Reasonable to assume that each group has characteristic behaviors
  - Identify eigenbehaviors of communities; project individuals onto the behavior space
  - Affiliation inferred from Euclidean distance btw. individual behavior & principal comp.
  - Also: distance btw. pair of subjects within a community ~ probability of friendship



● Business school students
▲ Senior lab students
♦ Incoming lab students
□ Lab staff and faculty

Eigenbehaviors for social networks

- Math similar to the previous case, but now…
  - Matrix B: (M×H) → each row is the average behavior of an individual in the community
  - Same transformation B→B'
- For this example: only Bluetooth proximity data
  - # of devices discovered in each hour of scanning
- Principal eigenbehaviors exhibit main characteristics



Eigenbehaviors for social networks

- To determine similarity of members:
  - how accurately the behavior can be approx. by the community's primary eigenbehaviors
- A behavior can be projected onto the community *j* space

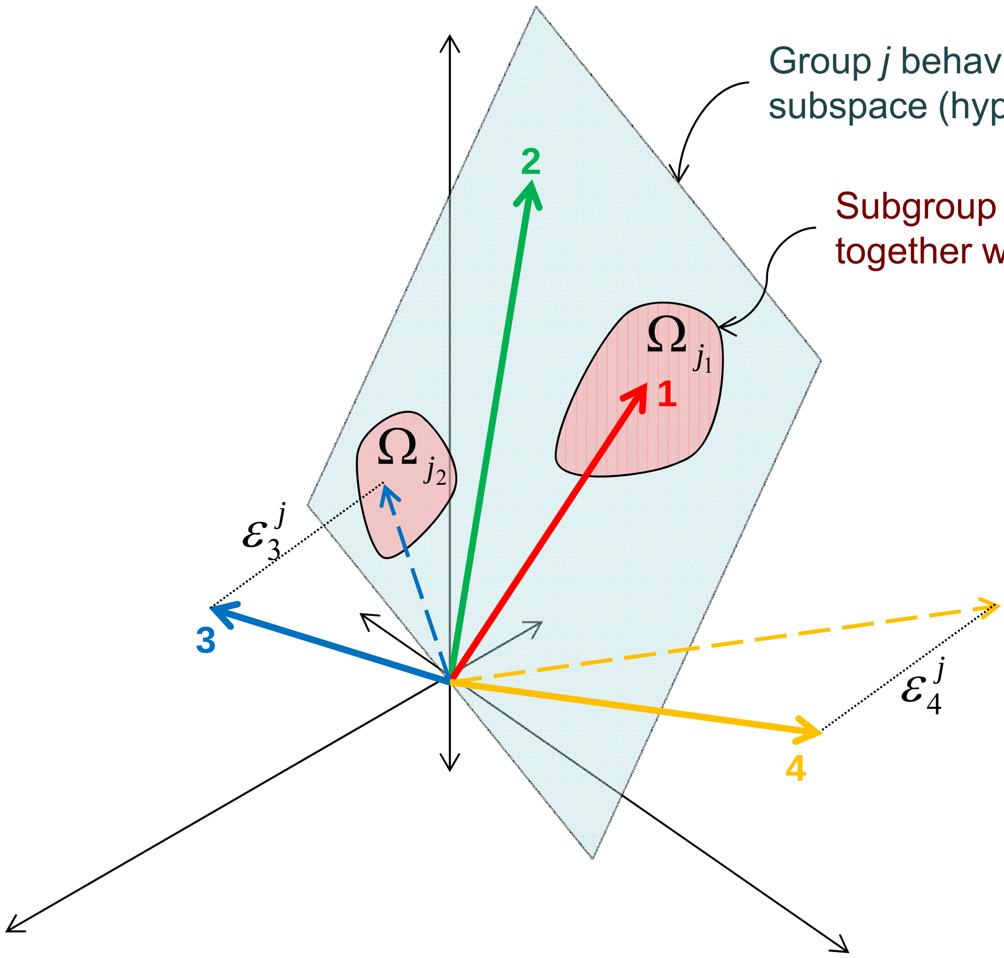
$$\omega_k^j = u_k^j\left(\Gamma - \Psi_j\right) \Rightarrow \Omega_j = U_j^T\left(\Gamma - \Psi_j\right)$$

- Vector Ω_j : optimal weights to get the behavior closest to the behavior space
  - Euclidean distance used to determine person *k* in *j* closest to the individual $\varepsilon_{jk}^2 = \left\|\Omega^j - \Omega_k^j\right\|^2$

Eigenbehaviors for social networks

- Method also used for determining most similar days
- Also: how much an individual "fits in" with a community → (classification)
  - Distance btw. original behavior (mean-adjusted) and its projection onto the community subspace
  - Projection: $$\Phi_b^j = \sum_{i=1}^{M'_j} \omega_i^j u_i^j = U_j \Omega_j$$
  - Distance: $$\varepsilon_j^2 = \left\| \Phi^j - \Phi_b^j \right\|^2$$
  - There are four possible outcomes of affiliation

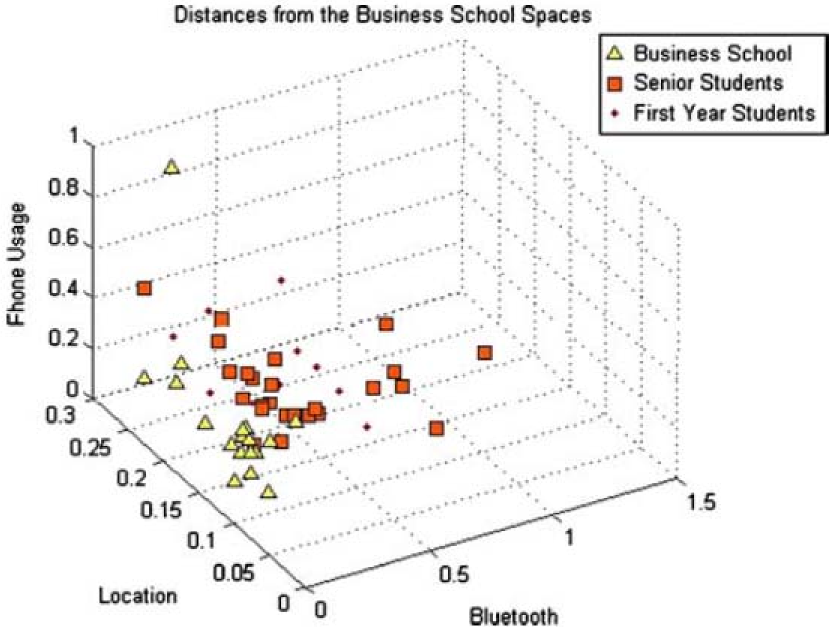
Affiliations in the behavior space



- Ind.1: lives in the subspace, can be affiliated to subgroup of individuals 1.
- Ind. 2: lives in the subspace, but is not close to other individuals
- Ind. 3: shares something with some individuals, but does not lie in the behavior space
- Ind. 4: disparate input neither near the behavior space nor any individual in the space.

Eigenbehaviors for social networks

- Until now: working with datasets independently → multimodal analysis also possible!
  - Generate set of eigenbehaviors for each type of data captured
  - Calculate an individual's Euclidean distance from each space
  - Points closest to the origin are more related to the community from where the spaces originate
  - Classification accuracy ~ 96%
- Distance btw. two points ~ probability of the pair being connected

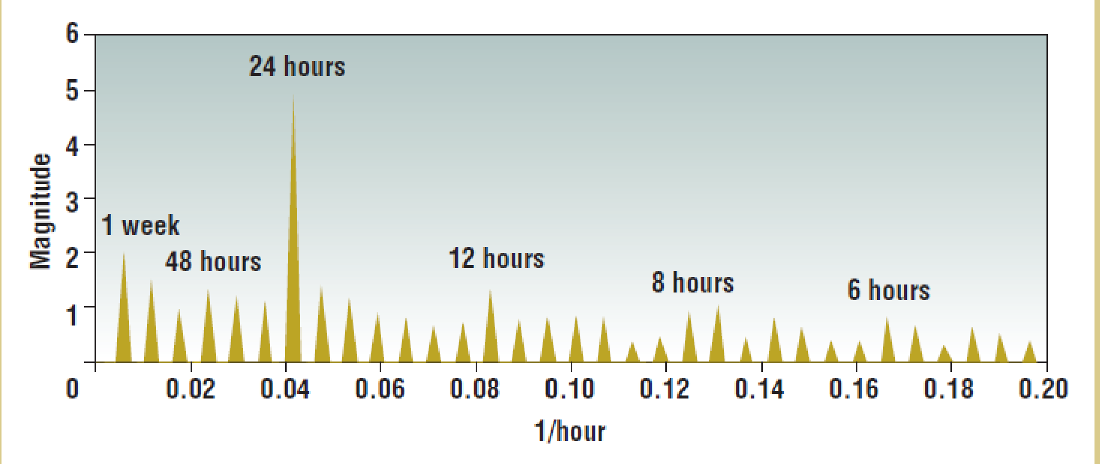


Another approach: Eigenplaces

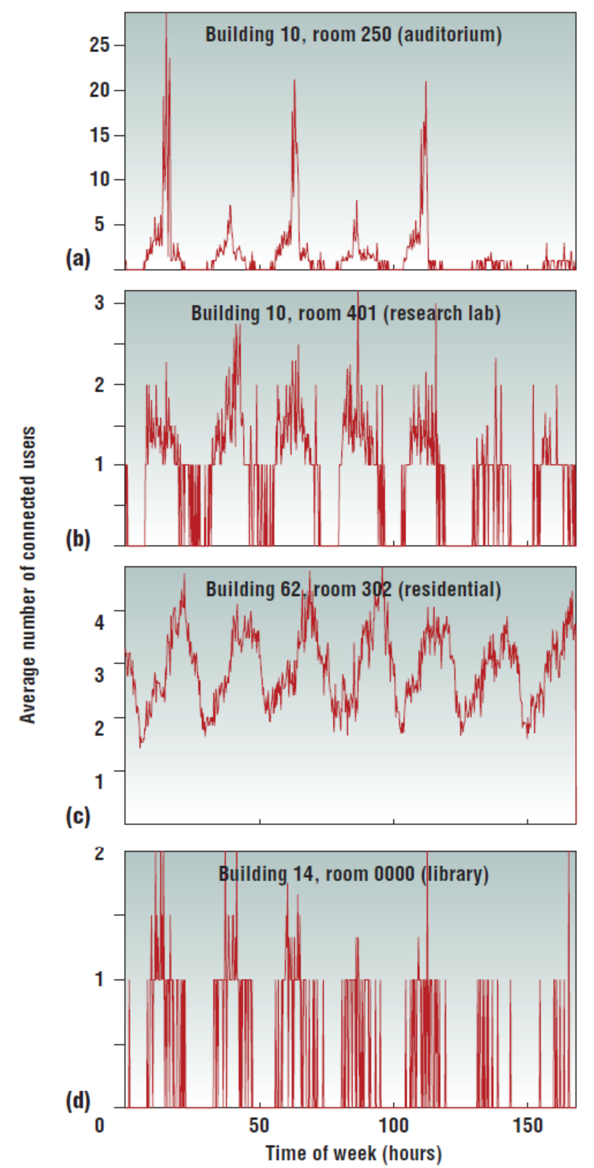
- Use of eigen-decomposition to leverage MIT's Wi-Fi network activity data and analyze its correlation to the physical environment.
- MIT campus covered with unified Wi-Fi network (APs)
  - 20 000 users, 250 000+ sessions/day
  - 73% students bring laptop to campus → network activity reasonable proxy of students activities
- **Experiment**: 2006 spring semester
  - Polled 3053 APs at 15-min intervals → determine # of connected users
  - No access to content → only spatiotemporal access profiles, preserving anonymity

Dataset preparation

- Holidays removed, average data → view of typical week
- Fourier transform shows daily & weekly access cycles
- Use of MIT's spaces database: 10 broad spatial types (e.g. classroom, administrative, residential, library, public space, etc.)
- Average # of connected user per week for each space type: graphs show distinctive characteristics

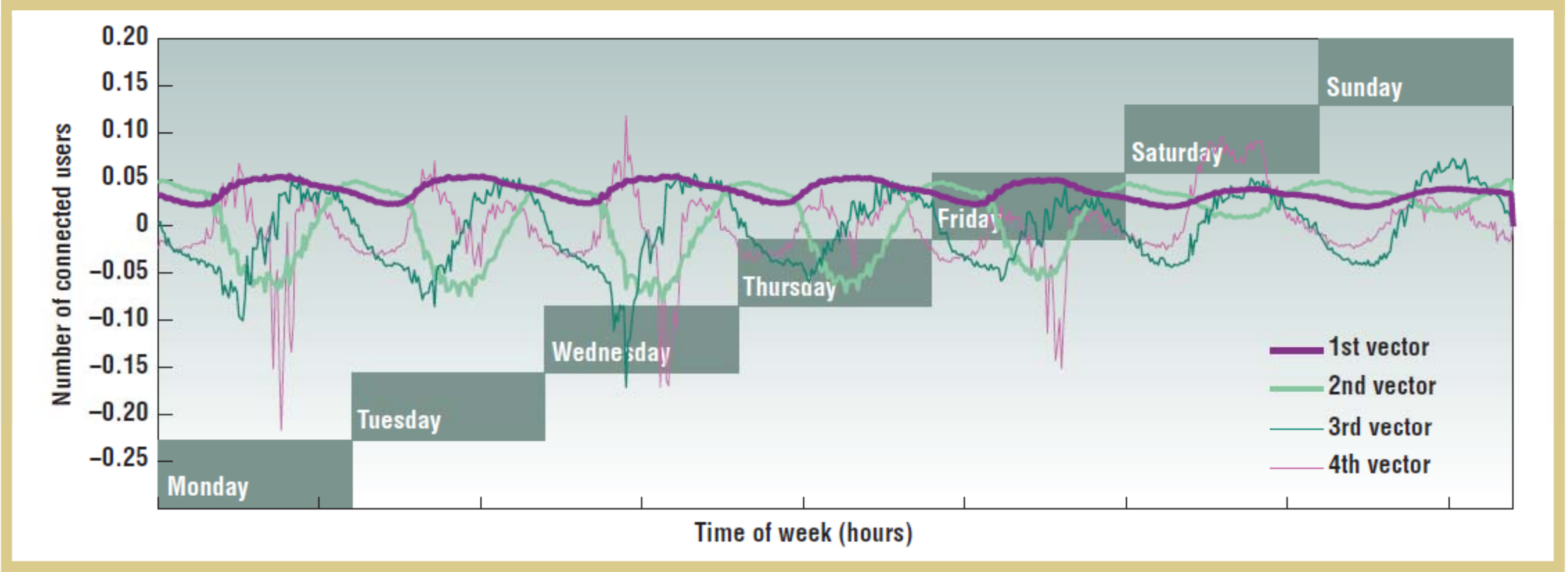


Fourier transform of the average week usage



Eigenplaces: Application of PCA

- # connections to an AP over a week → vector of 24×7=168 elem.
- All APs observations assembled into a single covariance matrix
- First **4** eigenvectors enough for keeping relative error < 0.1
  - V1: daily cycle, V2: evening activity, V3: not clear interpretation, V4: usage pattern of largest auditorium

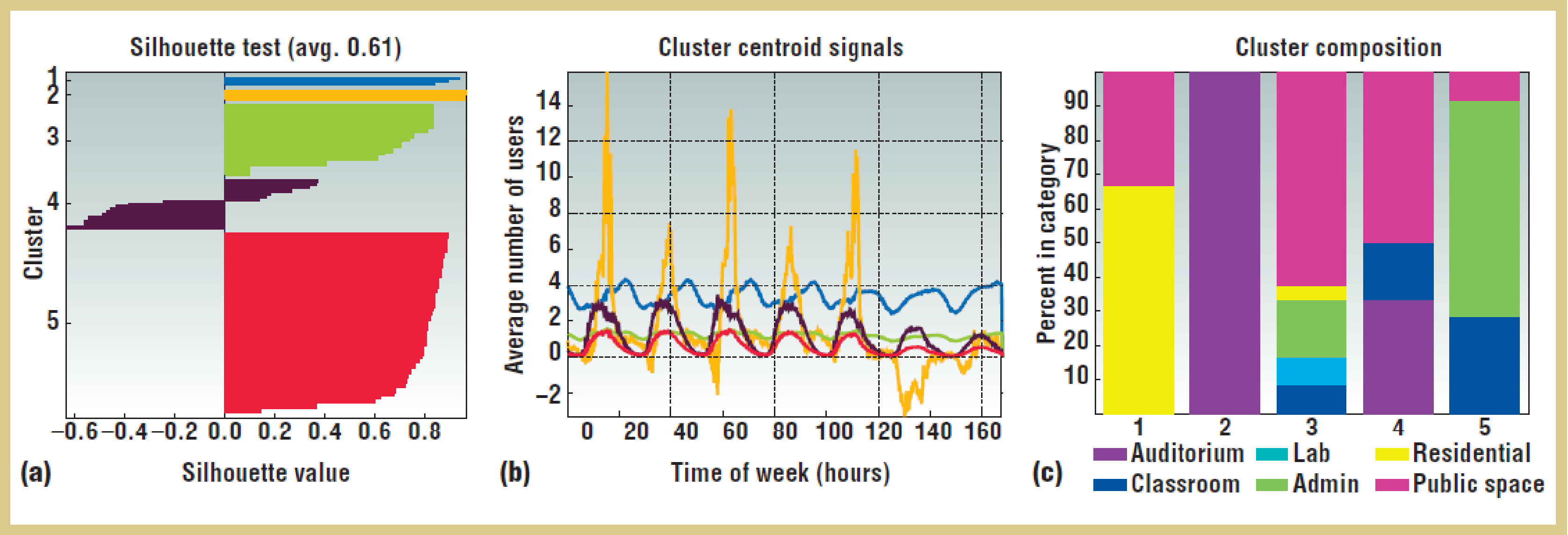


Eigenplaces: Application of PCA

- Key benefit: compression
  - Difference between APs captured entirely in coefficients
- Vector of coefficients describing each AP → **Eigenplace**
  - Comparable to any other place described with same vector set
  - Possible to cluster APs based on their distance in the space (similarity)
- **Clustering**: unsupervised k-means
  - Requires number of clusters → unknown!! Previous work used 3
  - BUT: use silhouette plot for finding optimal # of clusters!
  - Each AP silhouette value ~ how suited it is to its cluster and how far it is from other clusters. s-value in interval [-1, +1]
  - Tests showed that 3 clusters is NOT an optimal number

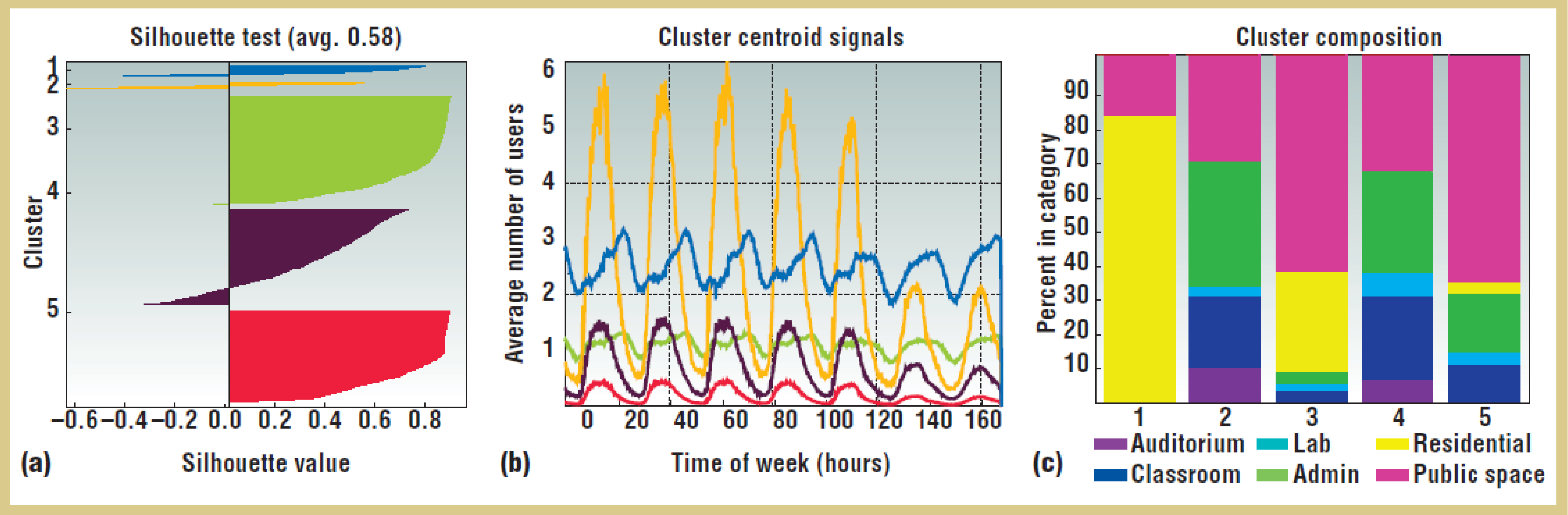
Cluster Training on partial data set

- Selected APs from 3 representative buildings
- 5 clusters maximized the average silhouette value (s-value = 0.61)
- Centroid signals → average of clusters in the eigenplace space, then taken back to the 168-dim. usage time space
- Comparison with “true” usage type classification shows consistency



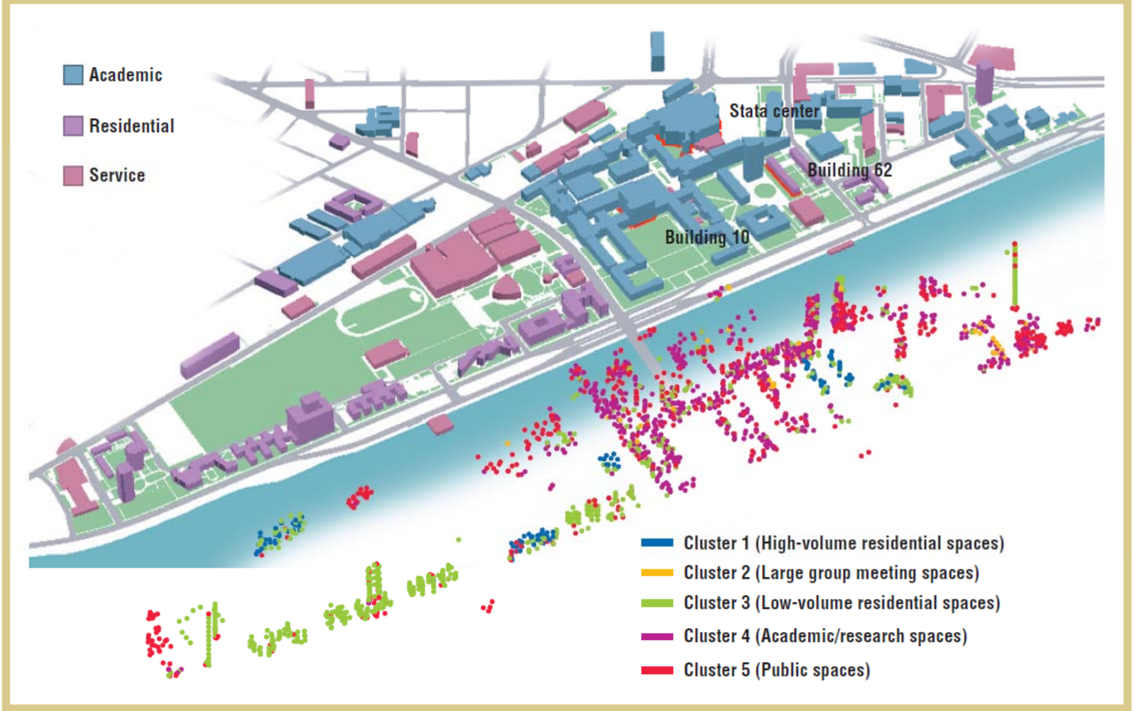
Cluster Analysis on full data set

- Previous step reduced risk of non-optimal solutions
- Full data fit is slightly weaker, but still quite coherent (s-value = 0.58)
- Clusters exhibit distinctive characteristics: 1 – public APs with very high traffic levels, 2 – small number of high-traffic public spaces, 3 – public APs from residential blocks, 4 – core buildings, 5 – most accessible ground



Successful approach

- Results of clustering all APs in campus show very distinctive features
- More than 3000 APs classified without personal inspections; possible to have continuous results at minimal cost.
- **Applications**: understand resource usage across a large-scale network; large advertising-supported systems



Critique

- **Overall rating:** average **4.0** (accept)
- **Technical strength:** average **3.8** (agree)
  - Greatly reduce the complexity of behaviors
  - Authors used large & solid data set
  - Efficient classification and prediction; good accuracy
  - BUT: revealed patterns are somewhat trivial, lacks proofs of correlation with ground truths, calculation of friendship probability not very clear
- **Originality:** average **4.0** (agree)
  - Known methods, but innovation is in the **application** to behavioral models
  - **Prediction** using eigenbehavior spaces is also very innovative
  - Reduction to a clustering problem for determining group affiliations

Critique

- **Presentation:** average **3.9** (good)
  - PROS: nicely written, easy to follow, good use of colored graphs, length
  - CONS: some typos, graphical representation of vectors needed
- **Contribution:** average **4.0** (strongly) → introduction of eigenbehaviors
  - Model to represent structure in routines
  - Insights for understanding behavioral data using dimensionality reduction
  - Understand what is important for characterization of ind./comm. behaviors
- **Future work:**
  - <u>Building concrete applications for the proposed methodology</u>
  - Make use of the prediction capabilities; use different/larger data sets
  - Compare/correlate affinity results with other social networks' data (e.g. FB)

Thanks for your attention.

Questions?