Principal Component Analysis of User Association Patterns in Wireless LAN Traces

Wei-jen Hsu and Ahmed Helmy

Department of Electrical Engineering
University of Southern California
{weijenhs, helmy}@usc.edu
http://nile.usc.edu/MobiLib/

Motivation

- **Are users similar to one another in their association pattern in long run?**
- **Does individual user show consistent daily association pattern across multiple days?**
- **If the answer to Q2 is yes, then how do we find some summarized presentation of the daily association pattern of a user?**
- **Can I group users using the summarized presentation obtained in Q3, leading to groups that show similar association pattern?**

Wireless LAN traces used

Trace source	Unique users	Unique APs	Unique buildings	Trace duration	User type	Environment	Analyzed part in this paper	Users in analyzed part	Labels used in paper
USC	5,580	79 ports	73	Dec 03-Now (trap) Apr 20 '05-Aug 7 '05 (detail)	Generic	Whole campus	Apr. 20, '05 to Aug. 7 '05	5,580	USC-05su
Dartmouth[6]	10,296	623	188	Apr. '01 to Jun. '04	Generic	Whole campus	Apr. 5 '04 to Jun. 4 '04	6,599	Dart-04sp
UCSD[5]	275	518	N/A	Sep. 22 '02 to Dec. 8 '02	PDA only	Whole campus	Whole trace	275	UCSD-02f

- **Traces from environments with various settings.**
- **In each trace we have AP association history of individual nodes, and we can further derive association time of each node to each AP.**

Matrices representation

- **Group association matrix- Do users have similar association pattern across a period of time?**
- **Individual association matrix- Does individual user have similar association pattern each day?**

Each column for a user

Each row for an AP

$$\begin{bmatrix} t_{u1,AP1} & \cdots & t_{un,AP1} \\ t_{u1,AP2} & \vdots & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ t_{u1,APm} & \cdots & t_{un,APm} \end{bmatrix}$$

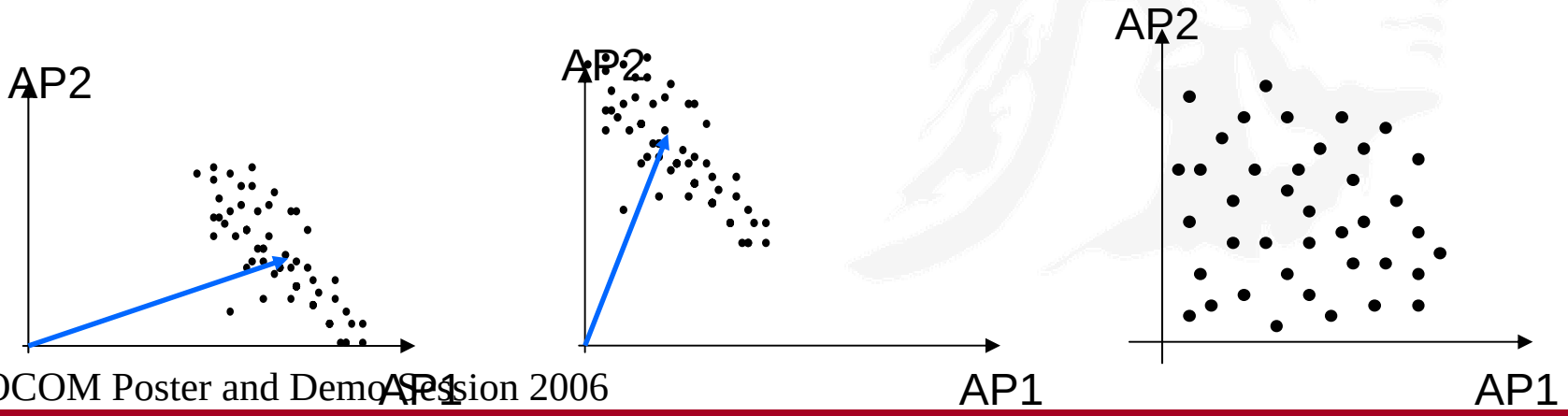
Each column for a day

Each row for an AP

$$\begin{bmatrix} t_{d1,AP1} & \cdots & t_{dn,AP1} \\ t_{d1,AP2} & \vdots & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ t_{d1,APm+1} & \cdots & t_{dn,APm+1} \end{bmatrix}$$

Principal Component Analysis

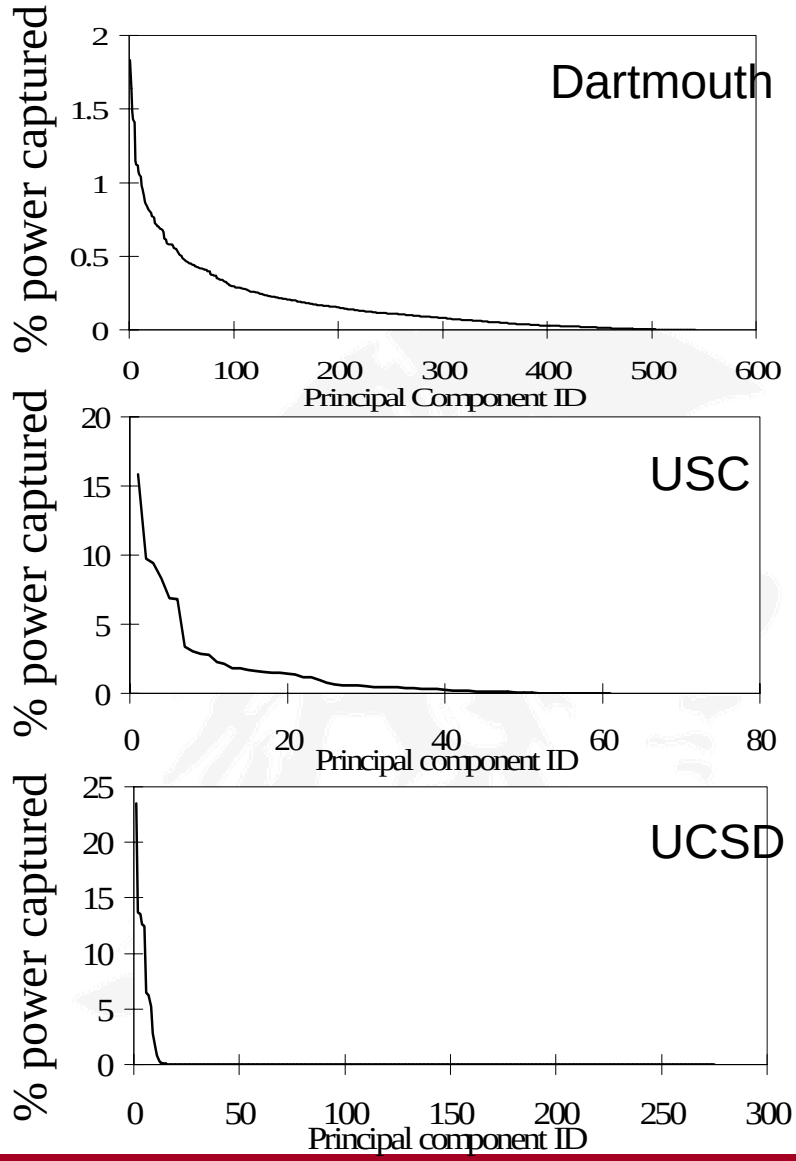
- **Applied as a tool to find the directions that carry most power, or the strongest association trend, in the matrices.**
- **If the data points show consistent trend, the first few PCs have high relative weights. Otherwise weights distributed across many PCs.**



Group Association Matrices

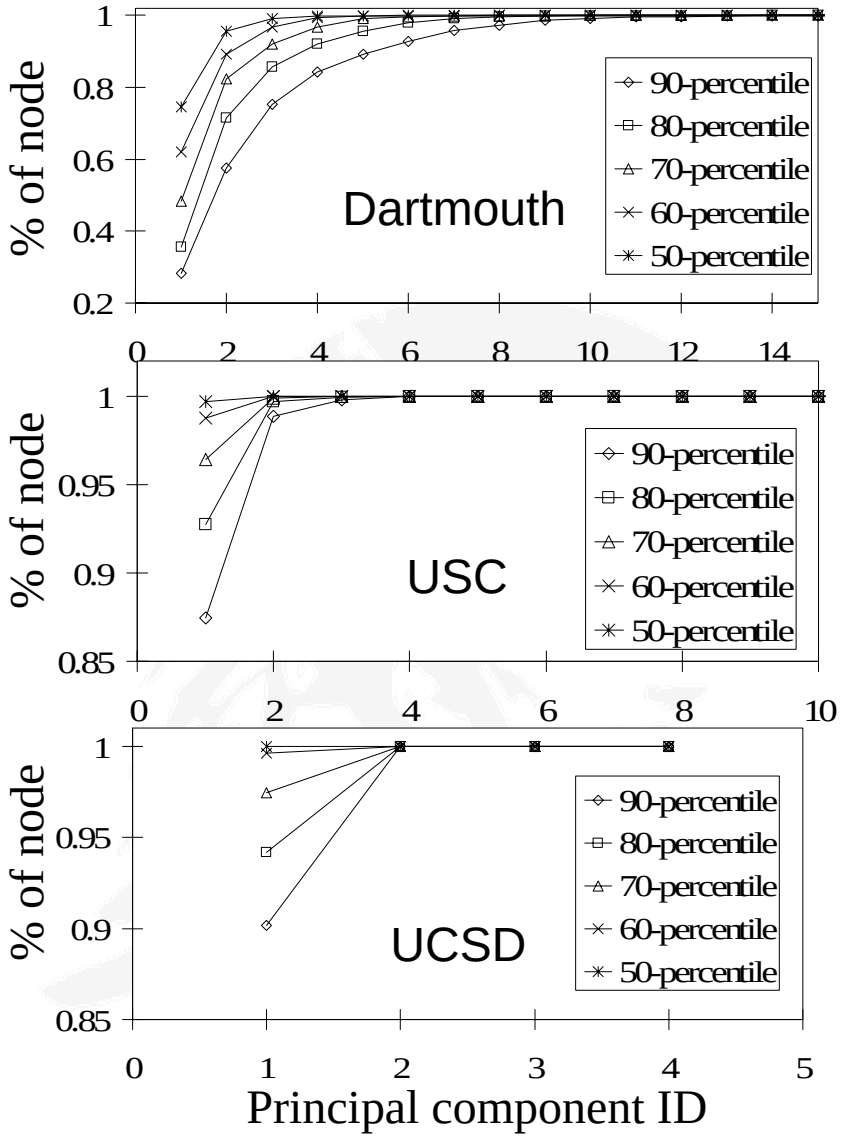
- **For large, heterogeneous user group, the long-run association patterns are diverse. Many PCs carry some weights, without dominant ones. (Dartmouth, USC)**
- **For smaller, homogeneous user group, dominant PCs exist. (UCSD)**
- **Weight of i-th PC =** $\dfrac{\lambda_i}{\sum_{k=1}^{n} \lambda_k}$

$$\lambda_i \equiv \text{i - th eigenvalue}$$



Individual Association Matrices

- **For most individuals, their daily association matrices have few PCs that carry most of the power, indicating that nodes show less diverse association pattern across days.**
- **The graphs show the percentage of nodes for which its p-percentile of power in individual association matrices is carried by x PCs with highest weights.**



Similarity between Individual Nodes

- **Principal components of individual association matrices with high weights are the "axes" along which the power of data points can be maximized.**
- **PCs of a user are unit-length vectors describing the major trends of its association pattern. Each PC's relative importance is given by the eigenvalues.**
- **We can compare the similarity between 2 sets of individual association pattern by weighted inner-product of the PC set.**

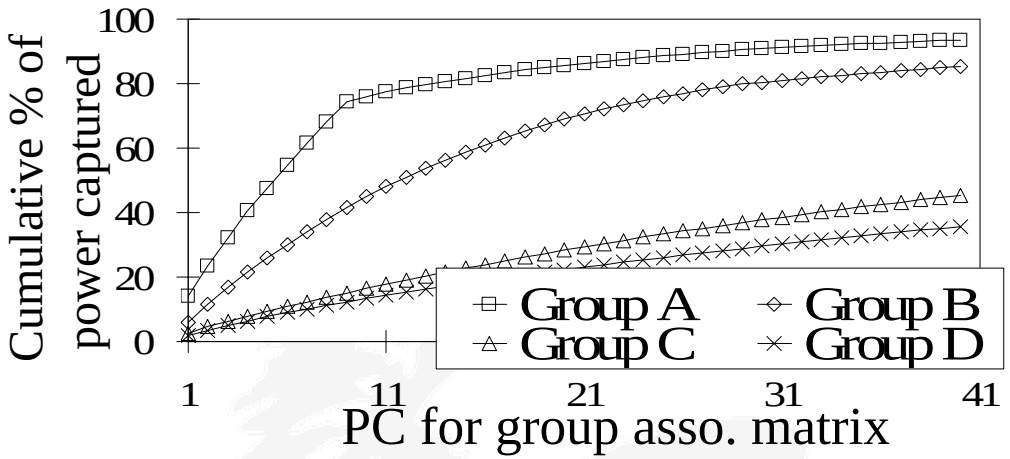
Similarity between Individual Nodes

- **Similarity index =**

$$Sim(U,V) = \sum_{i=1}^{rank(U)} \sum_{j=1}^{rank(V)} w_{u_i} w_{v_j} \left| u_i \cdot v_j \right|$$

- **Following the definition of similarity index, we can put nodes into sub-group, and more weights are carried by fewer PCs for group association matrices of these sub-groups.**

ui, vj: PCs from different users
Wui, Wvj: Weights for PCs in its set

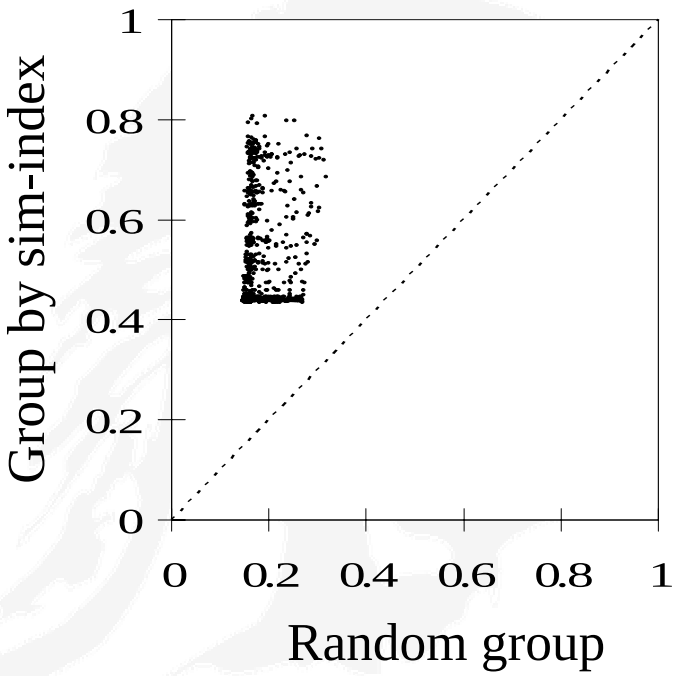


- **Group A/B: based on Sim-index, putting nodes having similar individual association pattern in a group.**
- **Group C: random group.**
- **Group D: the whole population group.**

Similarity between Individual Nodes

- **Comparing the weights carried by top-10 PCs between groups formed by sim-index (users are in same group if sim-index higher than threshold) and random groups of the same size.**
- **If we use higher threshold for similarity index, the groups chosen show higher common trend in association patterns.**

(Sim-index threshold = 0.8)



Conclusion

- **PCA is applied to find common association trends across multiple users, and across multiple days for individual user. Users do not share a common trend, but most individual users show consistent association trend across days.**
- **Principal components of a node provides a set of summary vectors to represent the strongest trends of its association pattern.**
- **Utilizing the similarity indexes, we can group similar users in sub-groups.**

Future Work

- **Utilize the grouping technique developed to design a routing protocol taking advantage of existing similar user sub-groups.**
- **Utilize the PCs of nodes to design a anomaly detection scheme to identify when a user deviates from its typical behavior.**

References

- Longer tech-report available at http://nile.usc.edu/~weijenhs/PCA-TR.pdf
- [1] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural Analysis of Network Traffic Flows," ACM SIGMETRICS, New York, June 2004.
- [2] M. McNett and G. Voelker, “"Access and mobility of wireless PDA users,”" ACM SIGMOBILE Mobile Computing and Communications Review, v.7 n.4, October 2003.
- [3] T. Henderson, D. Kotz and I. Abyzov, “"The Changing Usage of a Mature Campus-wide Wireless Network,”" in Proceedings of ACM MobiCom 2004, September 2004.