



MIT Open Access Articles

Eigenplaces: Segmenting Space Through Digital Signatures

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation	Calabrese, F., J. Reades, and C. Ratti. "Eigenplaces: Segmenting Space through Digital Signatures." <i>Pervasive Computing</i> , IEEE 9.1 (2010): 78-84. Print. © 2010 Institute of Electrical and Electronics Engineers
As Published	http://doi.ieeecomputersociety.org/10.1109/MPRV.2009.62
Publisher	Institute of Electrical and Electronics Engineers
Version	Final published version
Citable link	http://hdl.handle.net/1721.1/52542
Terms of Use	Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



senseable city lab:::

Francesco Calabrese
Jonathan Reades
Carlo Ratti

Eigenplaces: Segmenting Space through Digital Signatures

Eigenplaces: Segmenting Space through Digital Signatures

Francesco Calabrese, Jonathan Reades, and Carlo Ratti

Vol. 9, No. 1
January–March 2010

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

IEEE  **computer society**

© 2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

For more information, please see www.ieee.org/web/publications/rights/index.html.

Eigenplaces: Segmenting Space through Digital Signatures

Researchers use eigendecomposition to leverage MIT's Wi-Fi network activity data and analyze its correlation to the physical environment.

In the past two decades, wireless networks have permeated our public and private spaces, their usage shaped by the built environment's impact on user activity in the vicinity of transceivers. So, unlike unidirectional radio and television infrastructure, bidirectional wireless data networks can act as probes, propagating data about their users' environment back to a network observer. This fundamental difference lets us use the volume, timing, and distribution of packets across networks to study the "bricks and mortar" of physical space.

Francesco Calabrese
Massachusetts Institute
of Technology

Jonathan Reades
University College London

Carlo Ratti
Massachusetts Institute
of Technology

In contrast to the mobile network, Wi-Fi (IEEE 802.11) systems are particularly accessible to researchers because they're more modest in scale and are often operated by institutions with a vested interest in primary research. To date, most campus Wi-Fi deployment studies have focused on network performance and management¹ or inferred user mobility.^{2,3} However, as Jong

Hee Kang and his colleagues note, incorporating the concept of place allows a more sophisticated analysis and understanding of wireless environments.⁴ We propose a method to analyze and categorize wireless access points (APs) based on common usage characteristics that reflect real-world, place-based behaviors.

We use *eigendecomposition* to study the Wi-

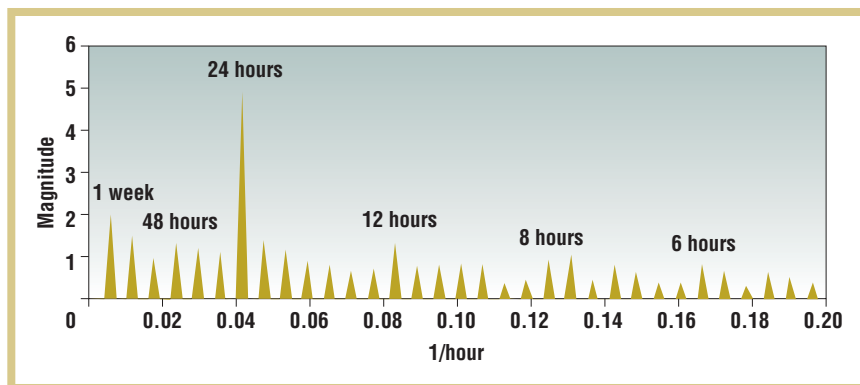
Fi network at the Massachusetts Institute of Technology (MIT), correlating data generated as a byproduct of network activity with the physical environment. Our approach provides an instant survey of building use across the entire campus at a surprisingly fine-grained level. The resulting *eigenplaces* have implications for research across a range of wireless technologies as well as potential applications in network planning, traffic and tourism management, and even marketing.

The MIT Wireless Environment

Like many universities, MIT has covered its campus with a unified Wi-Fi network; all APs share the MIT network name, enabling 20,000 users to establish more than 250,000 sessions a day. Filippo Dal Fiore and his colleagues found that 73 percent of MIT students bring their laptops to the campus either daily or on most days of the week.⁵ So, network activity is a reasonable proxy for many student activities, making it suitable for an aggregate spatial analysis.

During the 14-week 2006 spring semester, we polled each of the 3,053 APs in our data collection infrastructure at 15-minute intervals to determine the number of connected users. (For more information on the collection architecture, see "Mapping the MIT Campus in Real Time Using WiFi"⁶ and "Urban Activity Dynamics."⁷) Although we couldn't tell what types of content the students, staff, and faculty were accessing, we hoped that the APs' spatiotemporal access profiles would provide an interesting

Figure 1. A Fourier transform of signals from all access points (APs). The transformation highlights the underlying data's most important cycles—24 hours and one week—by representing the signal as the sum of a set of sinusoidal frequencies multiplied by coefficients.



window into campus activity while maintaining users' anonymity.

To control for the signal noise of day-to-day random variation, we took the entire data set, removed the holidays, and averaged the remaining data to create a composite view of a typical week. Therefore, the 10:00 a.m. Monday data point for any given AP is an average of all 10:00 a.m. Monday observations from that location during the term. Figure 1 shows a Fourier transform of connection data from all APs—representing the signal as the sum of a set of sinusoidal frequencies multiplied by coefficients—and highlights the daily and weekly access cycles.

We also obtained access to extensive spatial data that the university's Department of Facilities compiled. The database contained details on 33,000 campus "spaces," including their area, elevation (floor number), and use class. There were more than 90 usage types—ranging from classrooms to coat rooms, and elevator shafts to exhibition facilities—which we grouped into 10 broad spatial types: administrative, auditorium, classroom, food/café, library, public space, research lab, residential, support services, and "do not count." This last group contains uses such as animal quarters and vehicle storage, which seemed unlikely to have a distinct usage profile or, indeed, any usage profile. Armed with an appreciation of the campus's complexity, we set out to understand how network usage varied within these basic categories.

Figure 2 shows the average number of connected users by time of week for auditorium, research lab, residential, and library APs. We can readily identify

some spaces by their aggregate usage profile alone. The profile for building 10, room 250, one of the largest auditoriums on campus, even reveals the days it held lectures. This result is consistent with earlier research suggesting that areas containing large, dynamic populations are readily visible even in comparatively coarse wireless-traffic analyses.⁸ Residential spaces such as building 62, room 302 are also easy to identify by rising weekday evening use and heavy weekend use. Figure 2 also reveals the cycle of opening hours in the research lab in building 10, room 401—our own workspace—and building 14, room 0000—the Hayden Library lobby.

From Eigenvectors to Eigenplaces

Adapting a technique drawn from signal analysis and remote sensing, we applied eigendecomposition to extract the discriminant features from our time-series data (the AP signatures in Figure 2). We represented the number of connections to an AP over time as a vector and assembled the observations from all APs into a single covariance matrix.⁹ Following eigendecomposition, we expressed each AP's original signal as a sum of the matrix's eigenvectors \underline{V}_i , $i = 1, \dots, n$, each modified by a coefficient C_i , $i = 1, \dots, n$ particular to that AP. So, we describe a signature S_i observed at a randomly selected access point i by the equation $S_i = C_{i1} \cdot \underline{V}_1 + C_{i2} \cdot \underline{V}_2 + \dots + C_{in} \cdot \underline{V}_n$. We would describe a second AP signature S_j using the same vector set \underline{V}_1 through \underline{V}_n , but with dif-

fering coefficients C_{j1} through C_{jn} .

Applying eigendecomposition to MIT's network data yields many eigenvector and coefficient pairs; the latter's magnitude establishes the vectors' ranking according to their value in reconstituting the original data. Using the mean-square-error test, we determined that only the first four pairs were required to lower this error below a reasonable threshold of 0.1, letting us disregard the remaining eigenvectors and coefficients. Figure 3 shows the four eigenvectors that capture the decomposed signals' most significant aspects. Negative values on the y-axis are an unavoidable effect of eigendecomposition but aren't significant for this analysis.

The daily cycle in Figure 1—rapidly rising usage in the early morning followed by a steady decline in the afternoon and evening—is also evident in Figure 3's first eigenvector. The second vector shows an evening activity pattern that's sustained on weekends, suggesting residential usage. As we might expect, by the third vector, the plot becomes more difficult to interpret holistically because these vectors express the observed signals' lesser aspects. So, we were surprised to find that the fourth vector mapped quite clearly onto the usage pattern in building 10, room 250—the large auditorium.

Generating a single set of eigenvectors common to all APs has an important analytical benefit: compression. Because all spaces on campus share the same eigenvectors, we can capture the differences between APs entirely in the coefficients.

So, we can accurately represent thousands of different, noisy, complex, time-varying signals with just four scalar values per hotspot. Figure 4 illustrates how this approach captures the usage variation: the distinctive plot of the auditorium's four coefficients reflects its equally distinctive usage pattern. The second coefficient is positive for the AP in building 62, room 302 only, reinforcing the residential inference we drew from the raw-signal study.

We term the combination of coefficients describing each AP an *eigenplace* because it encapsulates the principal components of a space's telecommunications profile. The eigenplace's key analytical benefit is that it's quantitatively comparable to any other place described with the same characteristic vector set. Because the coefficients are simple scalars, we can cluster APs solely on the basis of the similarities and differences between the coefficients, then examine the groups' distributions across campus.

Although many clustering methods exist, we wanted a bottom-up mapping to avoid imposing our own expectations about campus life on the usage data, so we chose an unsupervised *k*-means clustering. This approach partitions data such that each observation is as much like its own group's members, and unlike other groups' members, as possible. However, the *k*-means method requires researchers to specify the desired number of clusters, which can allow other preconceptions to intrude. In "Urban Activity Dynamics," the authors imposed a constraint of three clusters, reflecting MIT's own

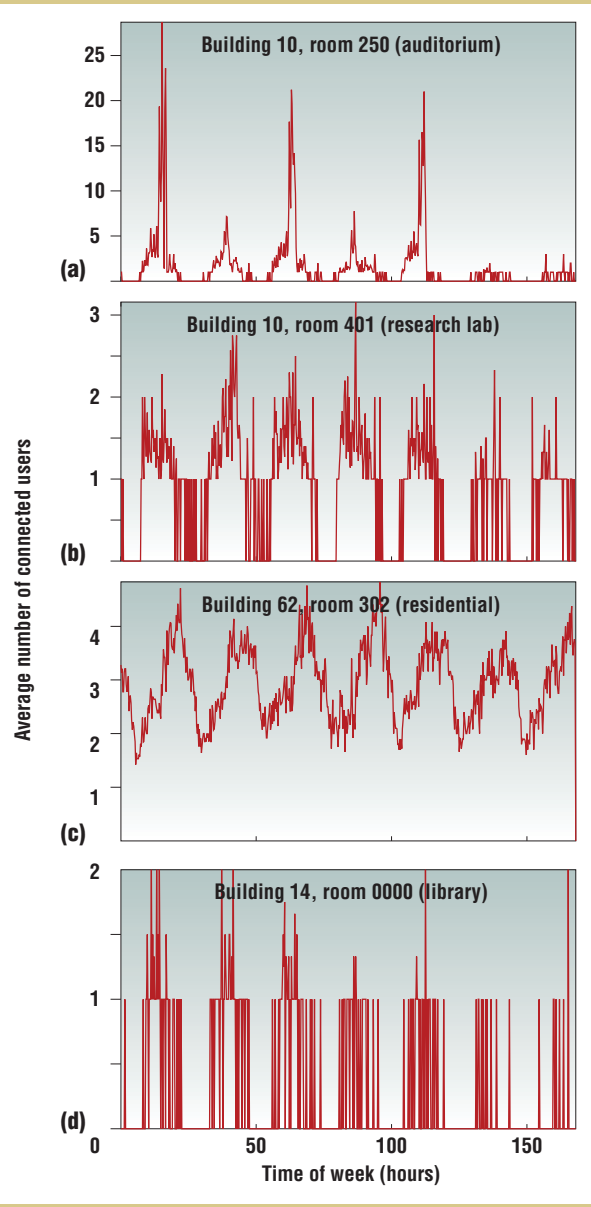


Figure 2. Aggregate network usage by location: (a) auditorium, (b) research lab, (c) residential, and (d) library. The academic spaces—auditorium, research lab, and library—all show evidence of typical working hours and days, whereas the residential access point in building 62, room 302 shows little variation across the week.

tripartite categorization of buildings into academic, residential, and service categories, and found that they could perfectly recreate this classification.⁷

Fortunately, we can gauge clusters' appropriateness both mathematically and visually using the silhouette plot.¹⁰

Each AP's silhouette value (*s*-value) measures how suited it is to its assigned cluster and how far—by whatever measure is appropriate—it is from any other cluster. We calculated the *s*-value using the squared Euclidean distance across the four dimensions abstracted from the eigen-decomposition process using the following formulation in Matlab: $S(i) = (\min(b(i,:), 2)) / \max(a(i), \min(b(i,:), 2))$ where $a(i)$ is the average distance from the *i*th point to all other points in the cluster, and each $b(i, k)$ is the average distance from the *i*th point to all points in another cluster *k*.

The silhouette plot simply shows the *s*-value for each cluster element, and the average silhouette measures how appropriately we clustered the data. An *s*-value close to +1 means that the element is appropriately clustered, whereas an *s*-value close to -1 suggests the element is quite different from the other elements in the cluster as measured by its distance from the centroid.

When we subjected our results to fitness tests, we were surprised to find that three clusters wasn't the optimal solution suggested by the data. We now wanted to investigate why evidence existed of more than three distinct Wi-Fi usage types, and determine whether these additional usage types had real-world behavioral correlates.

Cluster Training on a Partial Data Set

The APs' complex physical environment makes our clustering algorithm quite sensitive to initial conditions. To manage this risk, we employed a training process to create and calibrate pro-

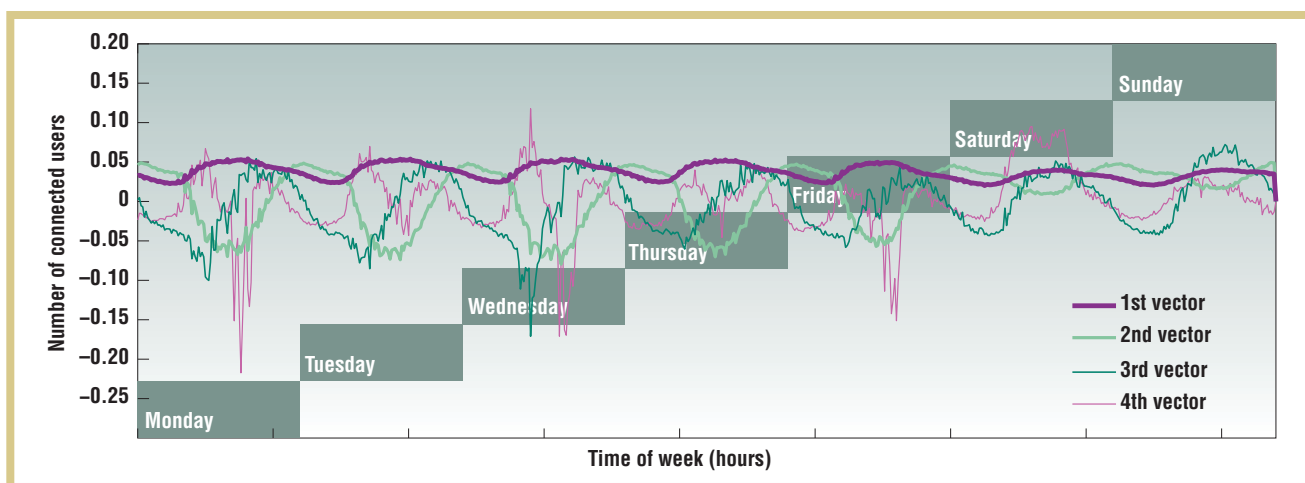


Figure 3. The four primary eigenvectors from MIT's Wi-Fi network. Much like the Fourier transform in Figure 2, the first eigenvector encapsulates the most important dimension of Wi-Fi usage at these access points. Depending on the coefficient, the second eigenvector either increases or dampens the first eigenvector.

prototype clusters using a subset of the data. Selecting APs from three representative buildings—10 (auditorium, classroom, and administrative), 62 (residential), and the Stata Center (auditorium, food/café, and administrative)—we found that five clusters maximized the average silhouette value; only the fourth cluster showed significant within-cluster distances (see Figure 5a).

The centroid signals in Figure 5b show the clusters' average signal, which we calculated for each cluster centroid eigenplace and then recombined with the original vectors to generate a composite signal. The cluster 1 centroid suggests residential origins because it maintains relatively heavy weekend usage; cluster 2 demonstrates the large auditoriums' impact. Clusters 3, 4, and 5 are more difficult to interpret solely on the basis of the centroid. Cluster 3 suggests public spaces because of a low-intensity pattern during both weekends and weekdays. The silhouette plot suggests that cluster 4 will be problematic regardless of our approach. Finally, cluster 5 appears to serve classroom and administrative functions because of the much lower average number of weekend users.

The Department of Facilities-supplied usage type classifications in Figure 5c reinforce our understanding of Figures

5a and 5b. The public spaces in cluster 1 are from the second floor and higher in building 62. The public spaces in cluster 3 are from the ground floors and basements of buildings 10 (academic) and 62 (residential). Cluster 5 is exclusively academic, incorporating classroom and administrative functions. Interestingly, all APs in cluster 4 come from just one building—the mixed-use Stata Center. We aren't sure why this building shows up in our training data this way, but we speculate that the complex floor plan and mix of uses create difficulties in our clustering approach.

Cluster Analysis of the Full Data Set

Using the centroids we obtained from the testing data to populate a second *k*-means clustering of the entire campus reduces the risk of nonoptimal solutions by ensuring that the test results respect the intercluster differences we identified in Figure 5. The issue arises because of the probability that usage at some outlying APs deviates so far from the norm that it skews the clustering process toward solutions in which most clusters contain just a few extreme APs. Figure 6 suggests that although the data fit is slightly weaker, the overall grouping remains remarkably coherent and

the centroids are still quite distinct.

Because we added the rest of the campus, the average *s*-value of 0.58 in Figure 6a is lower than the training data's *s*-value. The centroids for clusters 1 and 3 in Figure 6b demonstrate sustained weekend loads, suggesting important residential components. Cluster 2 has significant peaks every day of the week, indicating that it contains a variety of large-group spaces, which likely caused the large in-group variation in Figure 6a. We expected different departments to use their classroom spaces differently, leading to a lower in-cluster consistency for those APs, and this is the case for cluster 4.

Because public spaces are an important component of each cluster, we analyzed this use category in more detail and found that the spaces varied by cluster in specific ways. Cluster 1 contains public APs with very high traffic levels from buildings 62, 64, and 79 (Simmons Hall), all of which are undergraduate dormitories. Cluster 2 incorporates a small number of high-traffic public spaces, including some from the Sloan School of Management. Cluster 3's public APs come primarily from residential blocks, but almost exclusively from the second floor and higher, indicating that these aren't areas open to the general

public or nonresident student body. Cluster 4 public APs are principally from the first through fifth floors of the core research, administrative, and classroom buildings. Finally, public spaces from cluster 5 incorporate activity from the most accessible ground and first floors of academic buildings.

Our analysis of Figure 6c implies that researchers need a priori knowledge of each cluster's constituent APs to extract meaningful information from the data set. However, Figure 7 makes it clear that our approach can impart important information about activity distribution across campus without recourse to any reference data. In effect, Figure 7 is a user-generated campus map, created entirely from anonymous, aggregate wireless data. The features that emerge—the graduate towers and highly connected undergraduate dorms, the academic core, and the lecture halls—are entirely the product of the clustering.

We've classified more than 3,000 APs for an entire campus without having to inspect each one in person, and we've done so using a method that can provide continuously updated results over time at minimal cost. In combination with observations at a small, stratified sample of hotspots, a large network operator could use an eigenplace analysis to understand the drivers of resource usage across an urban- or national-scale network.

Limitations

One challenge when working with wireless network analysis is signal propaga-

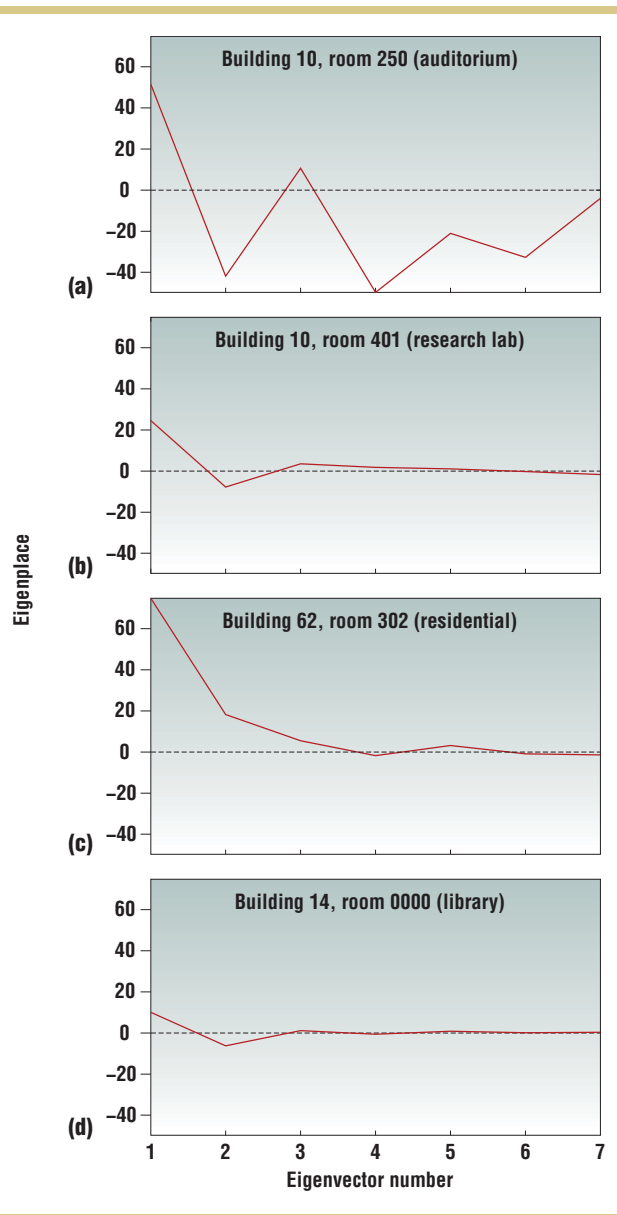


Figure 4. The eigenvectors' coefficients by location: (a) auditorium, (b) research lab, (c) residential, and (d) library. The first seven eigenvalues highlight the way that lesser values contribute almost nothing to the observed signal. However, the auditorium's extreme signal is much more difficult to fully capture with just four eigenvalues.

tion through walls and across floors. We had hoped that the Wi-Fi base stations' modest footprints would mean that they spanned fewer distinct uses and had correspondingly higher correlation between signature and function. However, abundant evidence indicates

that APs in a café might also serve adjacent classrooms or labs. And as the mixed results from the Stata Center suggest, coverage also varies with configuration because signals can propagate in unexpected ways. Nonetheless, correlating against only the use class of the room in which the AP is mounted still yields remarkable results using nothing more than aggregate wireless activity.

As Andres Sevtsuk and his colleagues detailed, there are important constraints on the activities that we can understand solely through network usage.⁶ At some places and times, such as during examinations or sporting events, network access is either banned outright or simply uncommon. We also can't account for Wi-Fi usage demographics, although evidence suggests that staff, graduates, and undergraduates use the network differently.^{11,12} However, this approach's power is that none of these issues is strictly relevant—we can search for similarities in network node usage without worrying about this difference's underlying drivers.

What's particularly interesting to us as built-environment researchers is that our method is a user-generated classification of space. Until recently, researchers have had difficulty investigating these aspects of human activity without extensive—and expensive—in-person studies, and this approach enables us to move toward a more nuanced vision of the environment as a

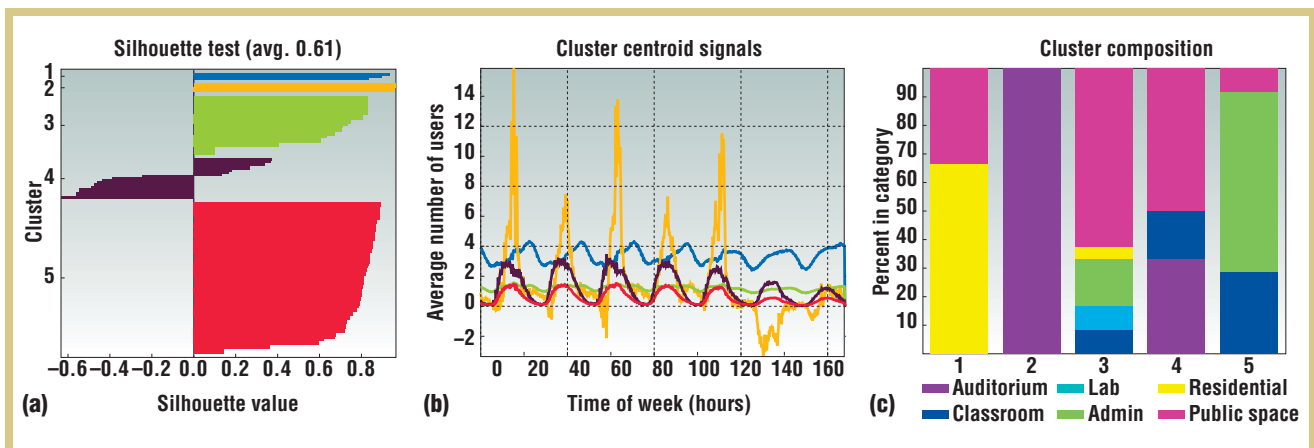


Figure 5. Results from clustering of the training data set. (a) The silhouette test shows the five k -means clusters (average s -value = 0.61). For instance, cluster 1 (blue) contains relatively few members and forms a fairly coherent grouping. (b) The clusters' average centroid signals. (c) Department of Facilities' classifications of access points by usage type.

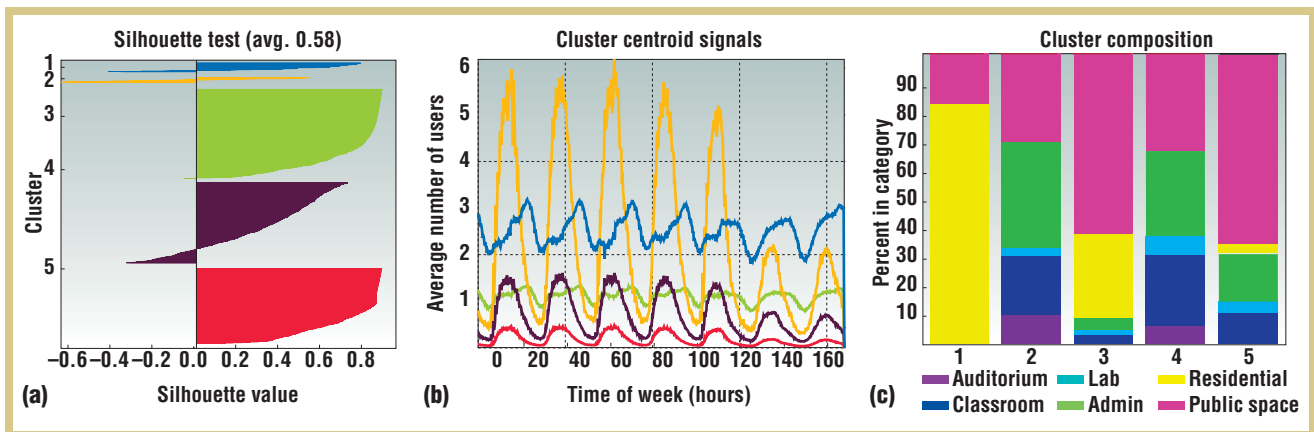



Figure 6. Results from clustering of the testing data set. (a) The silhouette test shows the five k -means clusters (average s -value = 0.58). (b) The clusters' average centroid signals. (c) Department of Facilities' classifications of access points by usage type. As we expected, within-group distances are somewhat larger, but the overall fit is still remarkably good.

dynamic system, not as a set of static, discrete spaces. Our approach has potentially valuable applications beyond the campus. For example, large advertising-supported systems, whether public or private, have had to balance targeted advertising's benefits against the privacy risks of snooping on individual users. Our approach offers an alternative that could be both anonymous and sensitive to activity context, including location, time of day, week, and year. 

ACKNOWLEDGMENTS

Jonathan Reades' research was supported by the International Balzan Prize Foundation.

REFERENCES

1. T. Henderson, D. Kotzand, and I. Abyzov, *The Changing Usage of a Mature Campus-wide Wireless Network*, tech. report TR2004-496, Computer Science Dept., Dartmouth College, 2004, pp. 1–19.
2. M. Kim and D. Kotz, "Classifying the Mobility of Users and the Popularity of Access Points," *Proc. Int'l Workshop Location- and Content-Awareness*, LNCS 3479, Springer, 2005, pp. 198–210.
3. M. Kim and D. Kotz, "Modeling Users' Mobility among WiFi Access Points," *Proc. Int'l Workshop Wireless Traffic Measurements and Modeling*, Usenix Assoc., 2005, pp. 19–24.
4. J.H. Kang et al., "Extracting Places from
- Traces of Locations," *Proc. 2nd ACM Int'l Workshop Wireless Mobile Applications and Services on WLAN*, ACM Press, 2004, pp. 110–118.
5. F. Dal Fiore, E. Beinart, and C. Ratti, "Do Mobile Users Move Differently? Exploring the Spatial Implications of Ubiquitous Connectivity at MIT Campus," *Proc. Geoinformatics Forum Salzburg*, A. Car, G. Griesebner, and J. Strobl, eds., 2008.
6. A. Sevtsuk et al., "Mapping the MIT Campus in Real Time Using WiFi," *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*, M. Foth, ed., IGI Global, 2008, pp. 326–338.
7. A. Sevtsuk and C. Ratti, "Urban Activity Dynamics," working paper, SENSEable

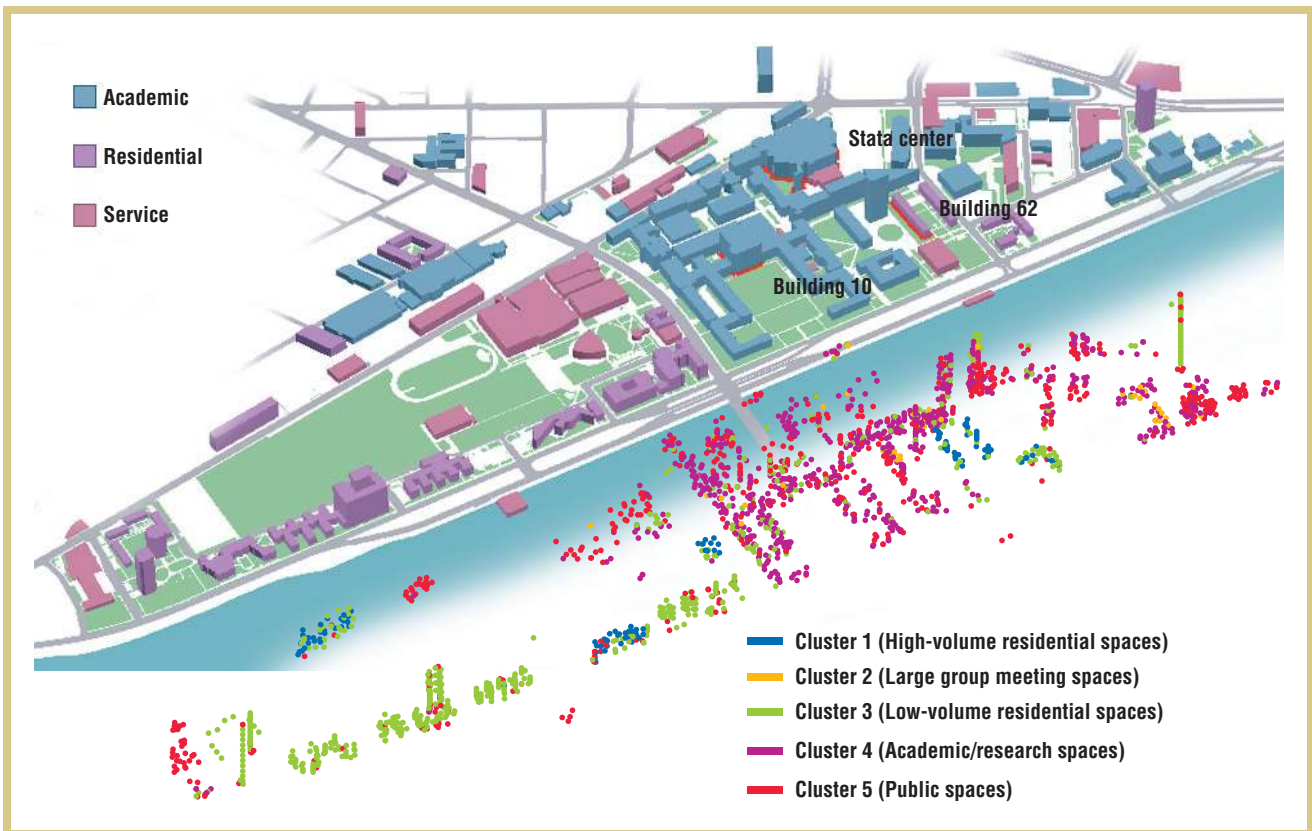


Figure 7. A 3D plot of campus clusters. Note the similarities between MIT's official building classification and the eigenplace analysis clusters. The Wi-Fi probes also pick up differences at a finer spatial scale in terms of usage, highlighting within-building usage differences.

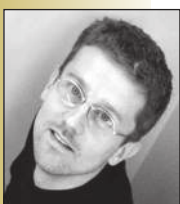
the AUTHORS



Francesco Calabrese is a postdoctoral associate at the Massachusetts Institute of Technology's SENSEable City Laboratory. His research interests include ubiquitous computing; analysis of urban dynamics through sensor networks; and analysis and design of distributed, hybrid, embedded control systems and computer-numerically-controlled machines. Calabrese has a PhD in computer and system engineering from the University of Naples Federico II. He's a member of the IEEE and the IEEE Control Systems Society. Contact him at fcalabre@mit.edu.



Jonathan Reades is a PhD candidate at University College London, where he is affiliated with the Centre for Advanced Spatial Analysis, part of the Bartlett School of Planning. His research interests include location theory and applying telecommunications data to the analysis of urban form and function. Reades holds a BA in comparative literature from Princeton University, and spent nearly 10 years working for a database mining and marketing firm that offered consultancy services to mobile network operators. He's a student member of the Royal Town Planning Institute and the Town and Country Planning Association. Contact him at j.reades@ucl.ac.uk.



Carlo Ratti is the director of the MIT SENSEable City Laboratory and an adjunct professor at Queensland University of Technology. He's also a founding partner (with Walter Nicolino) and director of the architectural firm *carloratti-associati*—Walter Nicolino & Carlo Ratti. Ratti has a PhD from the University of Cambridge. He's a member of the Ordine degli Ingegneri di Torino and the Association des Anciens Elèves de l'École Nationale des Ponts et Chaussées. Contact him at ratti@media.mit.edu.

City Laboratory, Massachusetts Inst. of Technology, 2008; <http://senseable.mit.edu/papers/pdf/SevtsukRatti-Activity-Dynamics-2008.pdf>.

8. J. Reades et al., "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing*, vol. 6, no. 3, 2007, pp. 30–38.
9. I.T. Jolliffe, *Principal Component Analysis*, Springer, 2002.
10. P. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics*, vol. 20, no. 1, 1987, pp. 53–65.
11. N. Eagle and A. Pentland, "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, 2009, pp. 1057–1066.
12. N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, 2006, pp. 255–268.