CS 224W Final Report
Characterizing the Urban Form with Persistence-Based Clustering on Graphs

Rohan Aras, Alex Nutkiewicz, Andrew Sonta

December 10, 2018

1 Introduction

Cities have the great task of having to provide various services to its people: water, energy, telecommunications, and clean air, among others. However, given the speed at which cities are growing and how population is urbanizing, it can be difficult for city and infrastructure planners to correctly size the services it needs for its citizens. One of the key solutions to this problem is to design modular infrastructure, servicing some of the natural boundaries or neighborhoods of rapidly growing cities. To implement this solution, a key challenge remains: how can we define the naturally occurring spatial boundaries of cities?

To answer this question, we begin by reviewing various methods used to explore the structure of shapes and graphs. Specifically, we look at how Heat Kernel Signatures (HKS), derived from thermodynamic theory, can describe similar neighborhoods of points within a manifold at multiple scales (e.g., locally within a shape, globally across a shape). Additionally, we explore Persistence-Based Clustering (PBC), when done in conjunction with HKS, and how it can show shapes being broken down into meaningful components. Given these observations, we explore in this paper:

- How can we apply Persistence-Based Clustering with a Heat Kernel Signature onto graphs?
- How can we identify the natural boundaries of cities based on these methods and a dataset of their components?
- How can we interpret the components of cities that are more topologically persistent than others?

By utilizing HKS and PBC, we hope to provide a method to interpreting the natural structure of cities.

2 Related Work

There is a long standing body of work that suggests that the built up area of cities can be modeled as fractals. These fractal assumptions can be used to explore the relationship between the perimeter and area of built up regions. Specifically, [TTVF11] employs a method called Minkowski dilation, where repeated "clustering" of buildings can be used to achieve a "crucial distance threshold" – at which point a morphological boundary (aka "urban envelope") can be defined. This method was applied to three French cities – Besançon, Belfort, and Montbèliard - and three Belgian cities - Namur, Liège, and Charleroi. This technique was able to show how a higher homogeneity of the urban landscape in Belgium relative to France reflected a higher level of urban density. However, these methods do not provide a means for interpreting smaller structural patterns within the built environment, which becomes an interesting topic of study when considering that cities are heterogeneous environments (i.e., cities are made up of unique neighborhoods).

However, there are other methods in shape theory used to understand their inherent structure, including one based on the geodesic distance between points [HSKK01] and another based on creating increasingly smooth interpretations of a shape [LG05]. However, many of these previously developed point signatures are sensitive to noise, are very computationally expensive, or are only able to create global heat signatures.

Thus, these methods cannot perform multi-scale comparisons of neighborhoods of points within a single shape. [SOG09] instead bases its point signature on the concept of heat diffusion between points on the surface of a shape. The idea of their Heat Kernel Signature (HKS), described in more detail in Section 3.1, is based on the concept of heat diffusing to progressively large neighborhoods of points, where time becomes a natural way to describe the shape of points around a given point. Because of this concept, detailed, highly local shape features can be observed through the behavior of heat diffusion of a shorter period of time, while summaries of a shape in large neighborhoods of points can be assessed through the behavior of heat diffusion over a longer period of time.

The purpose of Persistence-Based Clustering (PBC) is to be able to segment a shape into a smaller number of meaningful components. This area of work is in general related to watershed methods - an analogy to physical topography in which certain regions are split based on watersheds, metaphorically referring to physical watersheds that separate drainage basins. Guibas et al. [GSO+10] outlines some issues with existing work on mesh segmentation. In particular, they discuss the problems with the use of curvature as the watershed function, which is not robust enough for meaningful shape segmentation. Additionally, current segmentation methods do not come with the guarantee of quality in reconstructed segmentation, nor with segmentation stability. As discussed above, the HKS method [SOG09] addresses some of these issues. However, the use of HKS on manifolds *per se* does not complete the process of image segmentation, and so Guibas et al. [GSO+10] novelly introduces the concept of Persistence-Based Clustering (PBC) to be used in tangent with HKS.

PBC is focused on recovering basins of attraction of a function (such as the HKS) on a space. It reveals births and deaths of components of the space that are fully connected based on this function over time. Inspecting these births and deaths over time through a persistence diagram (PD), the user is able to determine visually the stability of different segments of the space based on a tuning parameter.

While various fractal-based methods have been used to study the structure of cities, they lack the ability to understand them at a more granular scale (e.g., neighborhood-level). Therefore, our project aims to use Heat Kernel Signatures and Persistence-Based Clustering to demonstrate how heat diffusion can describe similar neighborhoods of points within a manifold at multiple scales (e.g., locally within a shape, globally across a shape). [Bai07] shows how HKS can be applied to graphs instead of manifold/shapes: the Laplacian Matrix replaces the Laplace-Beltrami operator. However, Persistence-Based Clustering using HKS has not been extended to graphs – one of the contributions we hope to make as part of this work.

3 Methods

As discussed prior, the goal of this work is to see if the combination of Heat Kernel Signature and Persistence-Based Clustering can help us learn the naturally occurring spatial boundaries of cities.

3.1 Heat Kernel Signature

First introduced in [SOG09], the Heat Kernel Signature (HKS) attempts to capture information about the neighborhood of a point on a graph by recording the dissipation of heat from that point to the rest of the points in the shape in a set amount of time t. Mathematically, this concept is described by the equation:

$$k_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x) \phi_i(y) \tag{1}$$

where λ_i and ϕ_i are the i^{th} eigenvalue and the i^{th} eigenvector of the laplacian, respectively. The authors of [SOG09] argue that this, relative to many other shape analysis signatures, is more computationally efficient and is able to capture information about neighborhoods of a given point at multiple scales (from local to global) by modifying the t parameter.

The authors take this idea of a heat kernel and its inherent benefits related to creating point signatures for shapes and collections of shapes. However, the authors simplify the heat kernel by restricting it to the temporal domain, allowing for a more concise and easily commensurable method for understanding repeated structures within the same shape and across a collection for shapes. The HKS is defined as:

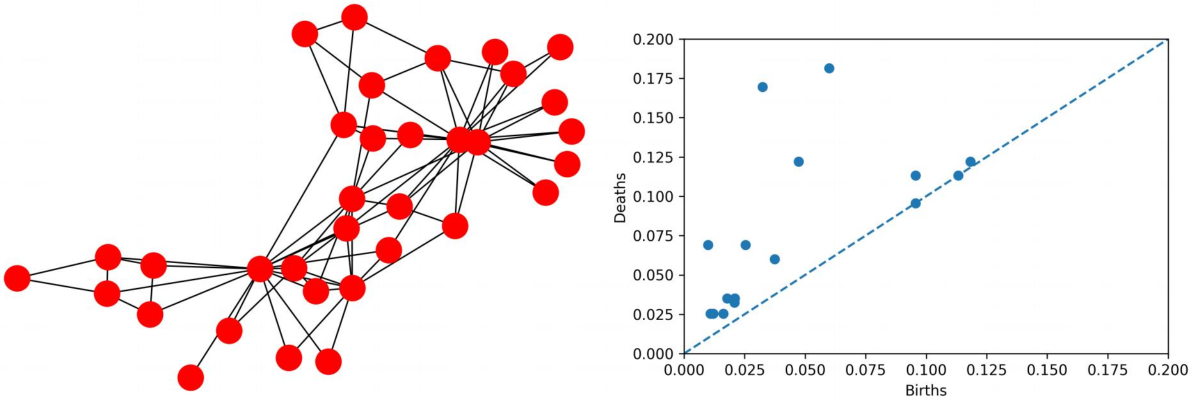


Figure 1: Karate network and computed PD based on HKS

$$HKS(x) : \mathbb{R}^+ \to \mathbb{R}, HKS(x,t) = k_t(x,x) \tag{2}$$

Where the HKS is a function over the temporal domain only. One of the main points the paper discusses is the Informative Theorem, which concludes that despite restricting the HKS to the temporal domain and removing the spatial domain from the heat kernel, HKS, defined by $k_t(x,x)$, is still able to maintain all the information necessary for describing a point signature. We employ this method in our implementation of the HKS for characterizing the natural structure of cities. However, calculating all of the eigenvalues and vectors of the Laplcian of a graph with several thousand nodes is quite expensive. Thus, as is discussed by [GSO$^+$10], we use a smaller subset of eigenpairs as there is an exponential decay in the influence of individual eigenvalues.

3.2 Persistence-Based Clustering

PBC operates over a space $\mathbb{X}$ with an associated function f and recovers *basins of attraction*, as discussed in [GSO$^+$10]. In our case, the space is a graph of buildings, and the function is the heat kernel signature value applied at each node in the graph. This algorithm can intuitively be thought of as analogous to defining where mountains begin and end in relation to one another as is done in mountaineering [EM97]. For example, we could define every peak (local maximum over the function of height above sea level) as its own mountain. However, clearly not every peak should qualify as a mountain—many mountains may have multiple summits but only one should characterize the mountain. PBC effectively merges nearby summits together if their difference in prominence (the amount one has to go downhill from one summit before going uphill toward the next) is large enough.

The algorithms for computing the PD and the actual clusters are the same. We set a hyperparameter (τ) to infinity when finding the PD and to a user-specified value when computing clusters. This is because the PD effectively tries to find all possible clusters over f. The inputs to the PBC algorithm are a graph G and a function f. In our case, f is the HKS function described above. We compute f for all nodes in the graph, and then we iterate through the nodes in decreasing order of f. We find the 1-hop neighborhood of each node x and find the local maximum. If x is a local maximum, we create a component and assign x to itself: $C(x) = x$. If x is not a local maximum, we assign it to a neighboring component. If the node is connected to two or more existing components, we merge the two components if they are not τ-persistent. In order to merge components with maxima x_1 and x_2 such that $f(x_1) < f(x_2)$, we set $C(x_1) = x_2$. When we merge components, we output the pair $(f(x_1), f(x))$, and these points become the values in the PD.

We demonstrate the PBC algorithm on a small test graph: Zachary's Karate Club, because it is well known that the network can be naturally defined by two large communities corresponding to the split of the club into two separate clubs. The PD produced through the algorithm produces two components that seem to persist longer than the others, as shown in the top left of the plot in Figure 1. We would expect

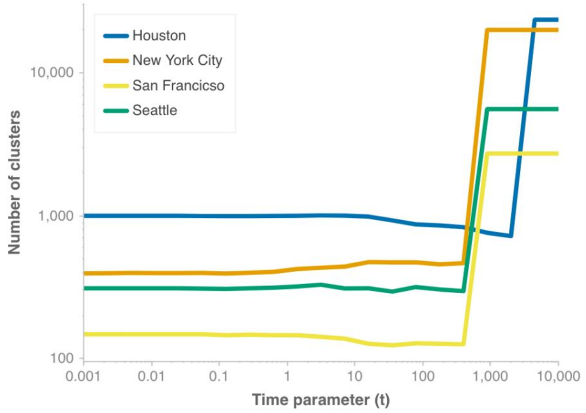


Figure 2: Number of clusters as t varies, with $\tau = 0$.

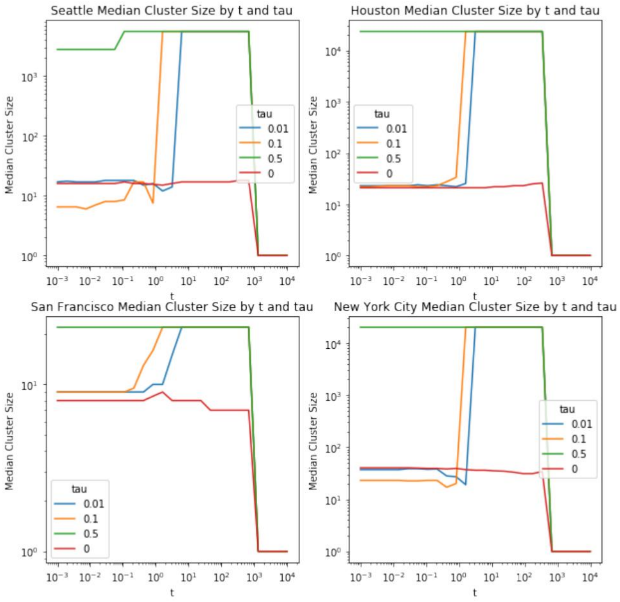


Figure 3: Median number size of clusters as t varies, for $\tau = 0, 0.01, 0.1, 0.5$.

the clusters formed through these components to correspond to the natural clusters that exist in the Karate Club Network.

4 Data and Results

4.1 Data and Graph Construction

Our dataset consists of 2017 tax parcel data from four cities: New York City, San Francisco, Houston, and Seattle. All four datasets have the shapefiles for each parcel as well its street address and a land use classification. These classification schemes are not consistent across the different cities, though we try to use categories (e.g. commercial) that have rough correspondents in all four. We restrict our analysis to only the commercial buildings for a couple of reasons: first, it reduces the size of each graph and therefore the computational resources needed for the HKS+PBC analysis. Second, it is a classification scheme that is quite similar across the different cities, allowing us to compare the cities on similar terms.

In order to generate a graph for each of these cities, we first calculate the latitude-longitude centroid of each parcel of each city. These coordinates are converted to the Universal Transverse Mercator (UTM) coordinate system so that calculations can be done in planar space. We calculate the distance between every pair of centroids in a single city. With this information, each node is defined as a single parcel centroid with k edges connecting it to the k nearest additional centroids (aka, parcels) in planar space. Additionally, each edge is weighted as the distance between the two centroids. k is chosen such that the graph for a given city has a single connected component, while keeping k as small as possible. When we do this for our three cities, we find that $k_{NYC} = 34$, $k_{SF} = 14$, $k_{Houston} = 12$, and $k_{Seattle} = 14$.

4.2 Physical Interpretation of Key Parameters

The two key parameters for our HKS+PBC clustering analysis are the time parameter, t, and the persistence parameter, τ. The time parameter t is the amount of time that heat is allowed to dissipate from one node to another in the HKS algorithm. Increasing t would mean comparing a given node to other nodes farther away in the graph. On the other hand, τ allows us to see how topologically persistent the clusters found through PBC are. Setting τ to 0 allows us to see all clusters.

We demonstrate how the number of clusters vary in each of the four cities as t changes (when $\tau = 0$), as shown in Figure 2. We can see that at a certain point, the number of clusters sharply increases toward the number of buildings in the graph, suggesting that each building is within its own cluster. The threshold at which this happens seems to correlate with the size of k, rather than the size of the graph. Our physical interpretation of this phenomenon is that, with very large t, each node is essentially compared to every other node in the graph, and therefore is placed within its own cluster. Interestingly, for t smaller than this threshold value, the number of clusters seems to be fairly constant, indicating that the clustering process is not very sensitive to the t parameter.

As an example for how we can compare similar types of neighborhood structure across a city, we use the island borough of Manhattan in New York City. In Figure 4 we plot the results of calculating the Heat Kernel Signature on every node in each graph (using $k = 8$ for this experiment). We choose node 1832 as a point of comparison. Node 1832 is located near the southern end of Midtown in the dense corridor of commercial buildings near the (east-west) center of the island. As we can see, for different values of t the geography of the nodes that are similar to 1832 change. In particular, for small values of t we see that other dense clusters of buildings are highlighted more than their surroundings.

To understand how the varying levels of τ affect the components generated by Persistence-Based Clustering, in Figure 5 we plot the clusters of commercial buildings in Manhattan for HKS values produced at a given value of t for two values of τ. $\tau = 0$ is chosen to show the base level clusters produced by the method before aggregation based on persistence. The second figure on the right shows a small amount of aggregation. From the plot on the right we can begin to see that certain parts of the graph with different topologies are being left in their own clusters, while the "generic" structure of the graph is aggregated into its own single component. By comparing these two levels of clustering, we can see that a higher value of τ reveals small clusters of buildings in Midtown and Lower Manhattan that are unique from others across the borough because they have not yet been aggregated into a larger cluster at this stage.

In Figure 3 we see that for every city, for values of τ bounded by 0 and 0.5, we see that there appear to only be two non-degenerate cluster sizes (where each building being in it's own cluster is the degenerate case). At one scale, for each city, there appears to be consistently between 10^1 and 10^2 buildings per cluster. At the other scale, every building in dataset for a given city is included in the cluster—there appears to be no middle ground.

Finally, as discussed in earlier in this section, we found that in comparing the number of building clusters of each city against the time parameter, this process is not very sensitive to the t parameter. So, we wanted to see the tradeoffs between τ and t and how they each affect the number of clusters in a city. Figure 7 shows the change in number of building clusters for each of the four cities based on both t and τ parameters. In studying the results, one can see that around the same time parameter t, each city sees a similar drop off in number of clusters. Depending on the selected τ value, many small building clusters will quickly grow into a large one that spans nearly the entire city. This confirms the earlier idea of there being two "scales" to a city: because of the rapid drop off in number of clusters, this model is able to cluster buildings at both the neighborhood and city scales.

5 Conclusion and Future Work

In this study, we introduced Persistence-Based Clustering and Heat Kernel Signatures to graphs in order to understand the patterns that define the natural boundaries of cities. We explored this method on four US cities: New York, San Francisco, Houston, and Seattle. Graphs for each city were constructed by defining nodes as buildings and edges based on the k-nearest buildings to each node. Using other measurements of distance that have more meaning to how people actually use cities would probably be more useful. For one, citizens of cities living in cities have barriers that make euclidean distance a crude approximation. It would be more useful to measure the travel time distance between buildings.

In doing this analysis, we learned that cities exist at multiple scales: both local, or "neighborhood," scales as well as more global, or "city," scales. When clustering techniques are applied to datasets describing urban buildings, patterns emerge showing pockets of unique urban forms within entire cities. With the ability to better understand the underlying structure of cities, planners, designers, and engineers will be better able to design future infrastructure to accommodate a rapidly urbanizing world.

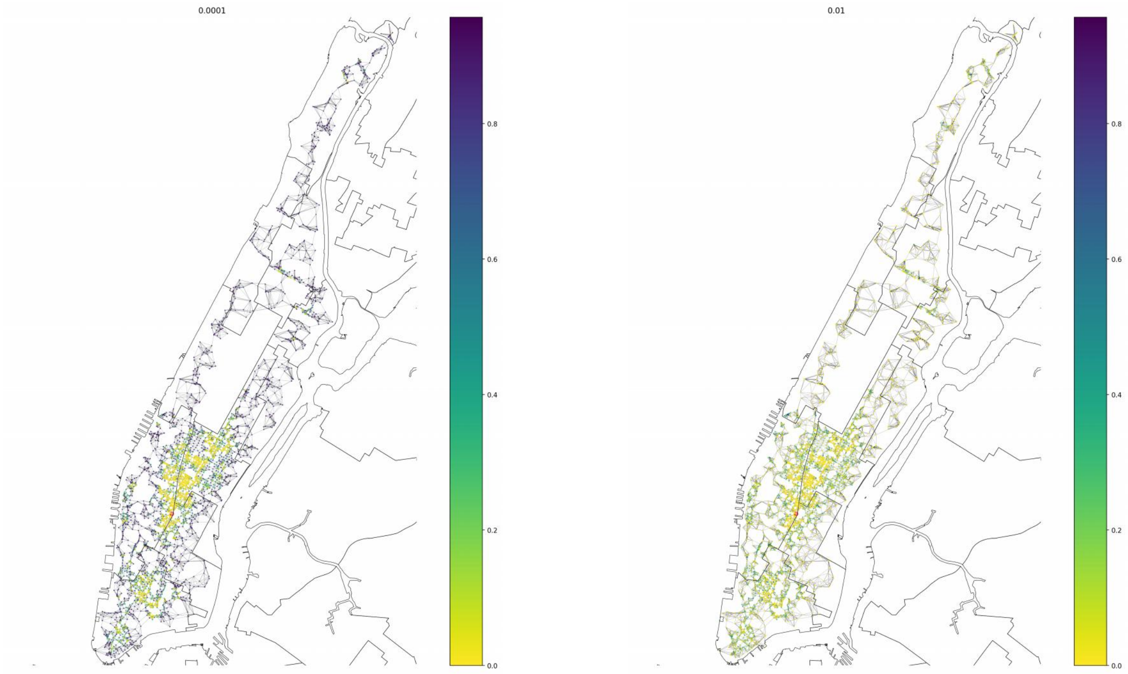
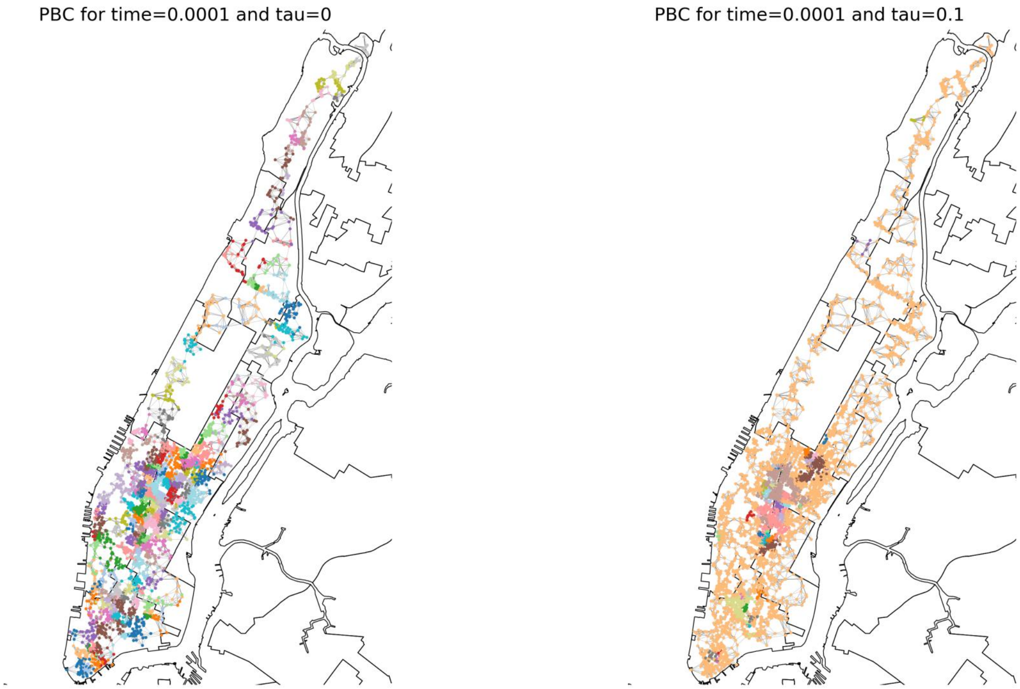


Figure 4: Comparison of HKS values to node 1832 (marked in red) for t=0.0001 and 0.01. Yellow nodes are similar while blue nodes are less similar.


PBC for time=0.0001 and tau=0

PBC for time=0.0001 and tau=0.1

Figure 5: Levels of Persistence-Based Clustering in Manhattan subset dataset at t=0.0001. The orange line denotes tau=0 and tau=0.1 respectively.

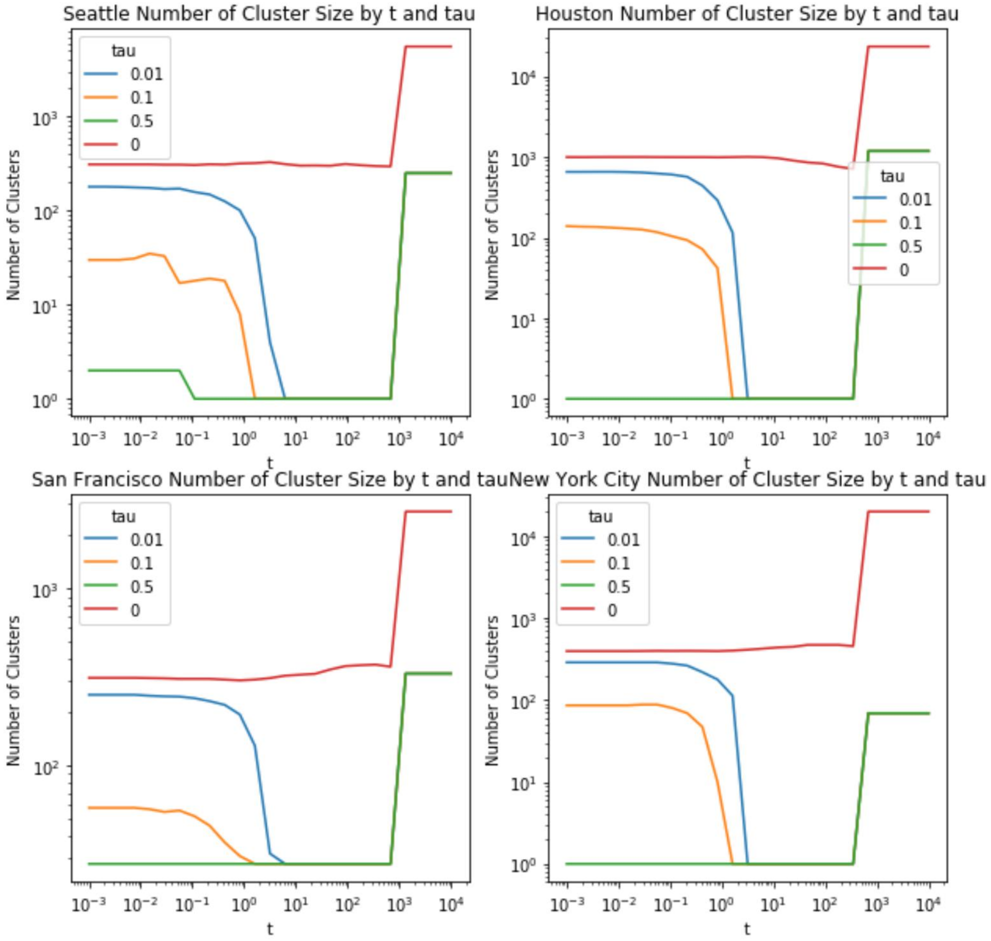


Figure 6: Tradeoffs between t and τ parameters in determining the number of clusters across each city.

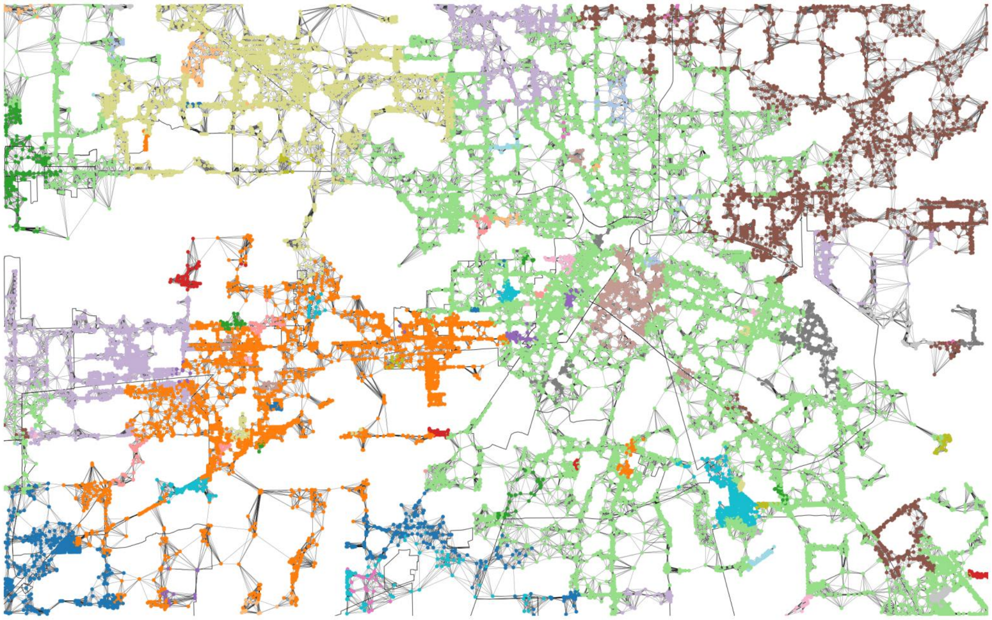


Figure 7: Clusters in Houston at $t = 0.056$ and $\tau = 0.1$. Note that several unique features such as downtown (light brown) are clearly differentiated from their surroundings.

6 Link to repository

https://github.com/rohanaras/CS224w-Building-HKS

7 Individual Contributions

- Rohan: Collected data (though this was for earlier research), problem formulation, lit review, wrote/debugged HKS algorithm, helped write PBC algorithm, exploratory+final analysis and plots, helped write up report.
- Alex: Problem formulation, wrote report, ran analyses and generated plots/figures, designed poster
- Andrew: Problem formulation, report writing, wrote/debugged PBC algorithm, ran exploratory tests

References

[Bai07] Xiao Bai. *Heat Kernel Analysis On Graphs*. PhD thesis, University of York, 2007.

[EM97] Herbert Edelsbrunner and Dmitriy Morozov. PERSISTENT HOMOLOGY. In *Handbook of Discrete and Computational Geometry*, chapter 26. 1997.

[GSO$^+$10] Leonidas J Guibas, Primoz Skraba, Maks Ovsjanikov, Frédéric Chazal, and Leonidas Guibas. Persistence-based Segmentation of Deformable Shapes. 2010.

[HSKK01] Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura, and Tosiyasu L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pages 203–212, New York, New York, USA, 2001. ACM Press.

[LG05] Xinju Li and Igor Guskov. Multi-scale features for approximate alignment of point-based surfaces. In *Proceedings of the third Eurographics symposium on Geometry processing*, page 236, Vienna, 2005. Eurographics Association.

[SOG09] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multiscale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, pages 1383–1392, Berlin, 2009. Eurographics Association.

[TTVF11] Cécile Tannier, Isabelle Thomas, Gilles Vuidel, and Pierre Frankhauser. A Fractal Approach to Identifying Urban Boundaries. *Geographical Analysis*, 43(2):211–227, 4 2011.