Evolutionary Clustering Analysis of Multiple Edge Set Networks used for Modeling Ivory Coast Mobile Phone Data and Sensemaking

Daniel B. Rajchwald, Thomas J. Klemas
Network Science Research Centre
Swansea University, Swansea, Wales
Emails: rajchwa1@gmail.com, tklemas@alum.mit.edu

***Abstract*—Static and evolutionary clustering approaches exist that enable dynamically adaptive cluster analysis of large networks. These techniques are typically based on any of the traditional techniques, such as k-means, spectral, Kerninghan-Lin, and other partitioning or clustering algorithms. In this paper, we utilize spectral clustering and k-means as the fundamental clustering mechanisms but combine adaptive and evolutionary clustering to capture problem dynamics. We apply our approach to analyze a complex, dynamic multiple edge set network that was used to model call data from the Ivory Coast compiled from France Telecom/Orange anonymized call records over a 5 month period. Our methods are used to identify important but non-evident structural groupings, resolve community clusters, develop insights based on the evolving structure and associated history, and to make sense of the raw data, the ultimate objective for Sensemaking technologies.**

Keywords–Sensemaking; adaptive clustering; spectral clustering; network theory; silhouette; k-means.

I. INTRODUCTION

Large data sets frequently contain patterns that are difficult to discern through observation alone. Data points in large data sets are often grouped in one or more dimensions based on similarities in data point values along those dimensions. However, many tools exist to group data based on proximity measures, such as Euclidean distance (where applicable), silhouette values, Saltines cosines, Pearson coefficients, or other measures of equivalence [1] that help evaluate similarity or dissimilarity between data points. Data clustering based on traditional algorithms, such as k-means, Spectral clustering, and Kerninghan-Lin, or hybrid combinations of these methods, can be a valuable tool to gain insight into many different types of data sets [1]. Once clusters are determined, new measurements can be classified more quickly based on proximity measures and parameters of the known clusters. However, over time, data groupings, clusters of data points, or even fundamental underlying network structure can evolve resulting in drift of parameters of the associated proximity measures. There has been significant study of cluster drift and related concepts of incremental and constrained clustering [2][3][4]. Evolutionary clustering techniques have been developed to capture cluster drift into clustering algorithms yet resist unduly perturbing clustering based on noise within the data by incorporating notions of expected smoothness in cluster parameters [2]. We adapt these concepts to accomplish evolutionary clustering analysis of a multiple edge set network used to model the Ivory Coast France Telecom/Orange call records.

The rest of this paper is organized as follows. Section II describes the technical details of adapting the evolutionary clustering algorithms for analysis of a multiple edge set network. Section III describes details of the data set and applying the algorithms to this data set. Section IV presents results of the multiple edge set network evolutionary clustering analysis. Section V presents our conclusions. The acknowledgement and references close the article.

II. TECHNICAL DETAILS

Now, we introduce our notation and review the basics of clustering. We model the social network derived from the call data by a graph G comprised of vertices V and edges E_1 and E_2 that represent subprefectures and the 2 types of connections between them, respectively.

$$G = (V, E_1, E_2) \tag{1}$$

The edges that connect vertex pairs represent calls, E_1, or travel, E_2, between those 2 paired subprefecture vertices. Even though the call and travel records identify the originating and terminating nodes in an edge, we construct an undirected graph model for simplicity. These cell towers are geographically distributed throughout the various subprefectures of the Ivory Coast, so there is an additional layer of mapping between the cell tower nodes and subprefecture nodes to which we applying the clustering analysis. A community S_i is comprised of a cluster of nodes, disjoint to every other community, because no vertex exists in more than one community.

$$V = \bigcup S_i, \forall_{i,j,i\neq j} S_i \cap S_j = \emptyset \tag{2}$$

To cluster the subprefectures into communities, we can assign each subprefecture, a, a feature vector, f_a and directly cluster the feature vectors into k clusters using the k-means algorithm. A more robust approach, [5] computes spectral decomposition

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{3}$$

of the $N \times N$ similarity matrix, $\boldsymbol{W}$ that is derived from the feature vectors, f_a,

$$W_{ij} = e^{\frac{-|(f_i - f_j)|^2}{2\sigma^2}} \tag{4}$$

and then clusters the row space of the eigenvectors, $\boldsymbol{U}$, corresponding to the largest k eigenvalues, by applying the k-means algorithm to the k-element rows of $[U_1 ... U_k]$ to compute k clusters. In [5], the parameter σ^2 determines the decay of

 ISBN: 978-1-61208-358-2

the affinity matrix values with distance in the feature space. In our implementation, we calculate σ^2 as the sample sum of the variances of each feature.

For evolutionary clustering, the silhouette metric is then used to determine the strength of community structure in an induced clustering, where the silhouette value [6] of one node or subprefecture, i, is defined as

$$silhouette\,(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{5}$$

where $a(i)$ is the average dissimilarity of i with the other subprefectures in its cluster and $b(i)$ is the minimum of the dissimilarities of i with all clusters that do not include i. Dissimilarity between 2 subprefectures is measured by the distance between their feature vectors. For instance, the dissimilarity between 2 feature vectors can simply be the Euclidean distance. The silhouette value for the entire clustering of nodes into k communities is simply the mean of the node silhouette values

$$silhouette(k) = mean_i(silhouette(i)) \tag{6}$$

Note that silhouette values range from -1 to 1, where 1 represents a strong community structure, -1 represents weak community structure, and 0 represents that the induced clustering is on the border with another viable clustering.

Then, we use a modified version of spectral clustering [2] to add temporal smoothness to clusterings. Before the modified version is discussed, we define some basic quantities. Given two subsets V_1 and V_2 of node set V, the association between the two subsets is $assoc(V_1, V_2) = \sum_{i \in V_1, j \in V_2} W(i,j)$ and the k-way average association between k clusters is $AA = \sum_{l=1}^{k} \frac{assoc(V_l, V_l)}{|V_l|}$. The modified spectral clustering [2] minimizes negated average association cost between two clusterings in adjacent time steps defined by

$$Cost_{NA} = \alpha NA_t|_{Z_t} + \beta NA_{t-1}|_{Z_t} \tag{7}$$

$$NA = Tr(W) - \sum_{l=1}^{k} \frac{assoc(V_l, V_l)}{|V_l|} = Tr(W) - Tr(Z^T W Z) \tag{8}$$

to obtain a clustering of the nodes at time t that is consistent with the network at time $t-1$. α and β, $\alpha + \beta = 1$, define the snapshot and temporal weights, respectively. Z_t is the $n \times k$ matrix that defines the partitioning at time t where $Z(i,j) = 1$ if and only if node i belongs to cluster j. Substituting equation 8 into equation 7 yields

$$Cost_{NA} = Tr(\alpha W_t + \beta W_{t-1}) - Tr(Z_t^T(\alpha W_t + \beta W_{t-1})Z_t) \tag{9}$$

Minimizing $Cost_{NA}$ is equivalent to maximizing $Tr(Z_t^T(\alpha W_t + \beta W_{t-1})Z_t)$ and optimizing Z_t turns out to be equivalent to applying spectral clustering to $W = \alpha W_t + \beta W_{t-1}$ [2], where W is the similarity matrix used in equation 5. Thus, by applying k-means to the rows of the matrix containing k eigenvectors corresponding to the top k eigenvalues of $W = \alpha W_t + \beta W_{t-1}$ yields the clustering at time t that maximizes both the snapshot and temporal quality.

III. Application to Data Set

A. Description of Data Set

This section of the paper is based on Blondel and Esch [7]. The data was organized into multiple sets. This research focused on Data Sets 1 and 2 in the Data For Development (D4D) collection. Data Set 1 consisted of antenna to antenna call records that include number of calls and duration of calls between any pairs of antennas, accumulated for each hour. Data Set 2 consists of records that identify cell phone tower indices for 500,000 randomly sampled callers but provided for only a 2 week duration. Data Set 3 consists of records that identify subprefecture indices for 50,000 randomly sampled callers for the entire 5 month duration of the D4D data. We decided to use Data Set 2 instead of Data Set 3 to model the traveler activity since the tower communication was recorded on a tower to tower basis. The data set also includes additional files that provide geographical location of antennas and subprefecture geographical center locations, enabling a mapping between antennas and the nearest subprefectures centers and thereby a graphical geographical depiction of result data.

B. Applying the Algorithms

We cluster the 255 subprefectures using temporal information with antenna call and/or cell phone user data. For example, if the feature vector was constructed entirely by antenna calls and time, then the feature vector for a would be defined as follows

$$f_a(t,b) = nCalls(t,a,b) \tag{10}$$

where $nCalls(t,a,b)$ represents the length of total cell phone tower communication between subprefectures a and b over a time period t. We then cluster all the feature vectors using spectral clustering as implemented by Ng and Jordan [5] (except we do not set the diagonals of the similarity matrix to 0) and the standard k-means algorithm. Given that k-means clustering does not account for noise and correlations in data, we quantify spectral clustering's effectiveness in clustering noisy and correlated data by comparing the performance of the two approaches. We implement the algorithms for cluster numbers, k, from 2 to 12 and compute the silhouette value of each clustering. The upper cluster number 12 was empirically determined from silhouette values that yield low numbers beyond 10 clusters. Once an optimum clustering is obtained, one can make inferences about relationships between the clustering features and established Ivory Coast information such as geographical, cultural, and political facts.

To compare the similarity of two clusterings, C_1 and C_2, over the same network, we first define the similarity score of a node, i, to be

$$SimilarityScore(i) = \frac{|C_1(i) \bigcap C_2(i)|}{|C_1(i) \bigcup C_2(i)|} \tag{11}$$

where $C_1(i)$ and $C_2(i)$ denote i's community in clusterings C_1 and C_2, respectively. Taking the mean of the similarity score over all nodes in the network yields the similarity score of the two clusterings, $SimilarityScore(C_1, C_2)$.

IV. Results

A. Clustering on Antenna Communication Edge Set Network

Cluster analysis was accomplished using feature vectors representing antenna communication between subprefectures

 ISBN: 978-1-61208-358-2

over 1 week from February 3, 2012 to February 9, 2012. We used two different sets of feature vectors. We defined the first set of feature vectors to be the cumulative activites of subprefectures over the 1 week period. For this section, we define activity to be antenna communication. Thus, a feature vector of a single subprefecture, a, would be a 255 dimensional vector, g_a, where $g_a(b)$ is the total length of calls beween subprefectures a and b accumulated over the entire week. We defined the second set of feature vectors to be the dynamic or evolutionary activity of the subprefectures over the 1 week period with a time interval of 24 hours. Thus, the feature vector for subprefecture a would be f_a where $f_a(b, n)$ is the total length of calls between subprefectures a and b on day n. Although the Ivory Coast regions have different populations, there is no population data on individual subprefectures so we were not able to normalize the feature vectors by population. Clustering was implemented using spectral clustering and k-means on the two sets of feature vectors. Figure 1 shows the result of spectral clustering on the dynamic subprefecture activities using 3 clusters.

We then applied adaptive spectral clustering to each of the 7 days from February 3, 2012 to February 9, 2012 as adopted from [2], for cluster sizes k = 2 to 12. A feature vector for a subprefecture, a, for a single day, n, would be defined as $h_a(b) = f_a(b, n)$ for f_a defined above. We used a snapshot cost, α, of 0.8 and a temporal cost, β, of 0.2. We computed the similarity score for each day with respect to both the cumulative and dynamic antenna communication activity clusters (through spectral clustering) over the February 3 to 9 interval. The similarity score of a day, n, was computed by averaging the similarity score of n's 11 clusters from k = 2 to 12 with either the evolutionary or cumulative clustering for February 3 to 12. The evolutionary and cumulative clusterings for the week serve as a common average to compare to each day. The results are plotted in Figure 2. Note that the day similarity scores are all relatively high and curves show oscillatory patterns.

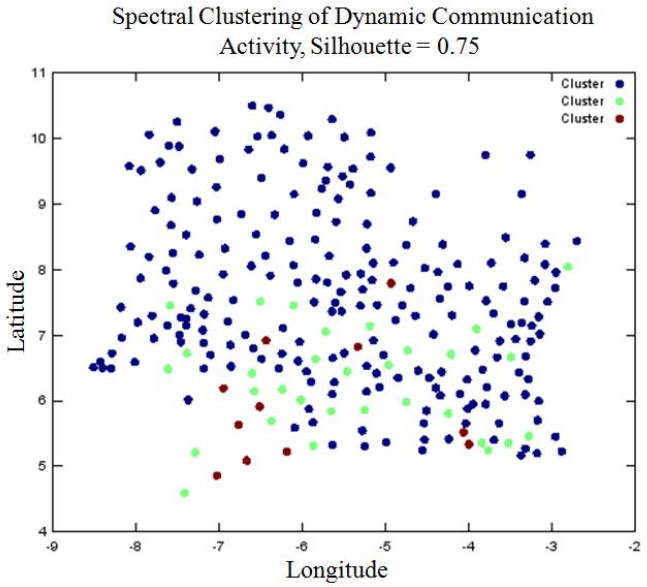

Figure 1. Example of Spectral Clustering using Subprefecture Communication Data

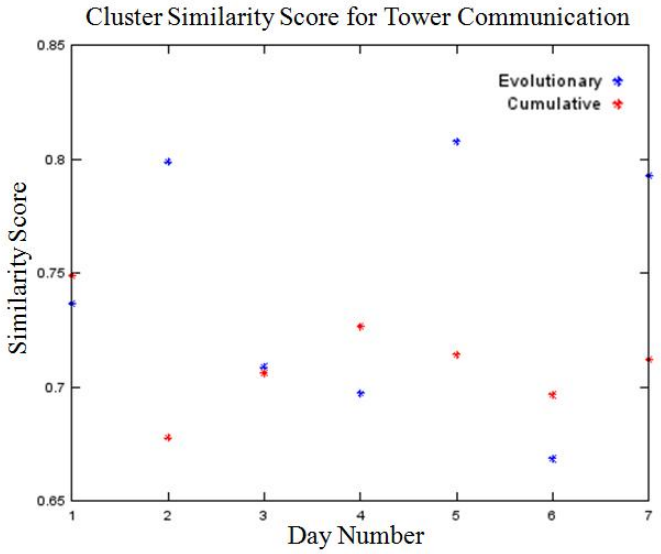

Figure 2. Similarity between Single Day Clusterings and Average Clusterings for the Week (Networks formed from Communication Data)

B. Clustering on Travel Edge Set Network

Cluster analysis was also achieved using the edge set corresponding to travel between subprefectures using spectral clustering and k-means. The second data set provides the location and times of cell phone users throughout the Ivory Coast. By geolocating the cell phone users by subprefectures, one can track when and where they travel between subprefectures. The cumulative and dynamic feature vectors were also formed from traveler data (D4D Data Set 2) between February 3, 2012 and February 9, 2012 (the dynamic travel vectors also with a time increment of 24 hours). The result of applying k-means when k = 3 to the dynamic traveler data can be seen in Figure 3. Despite the nice appearance of the 3 tight clusters, the clustering had a low silhouette score of -0.26.

Over all 12 spectral clusterings, the one with k = 2 clusters yielded the highest silhouette value in cases of antenna communication activity, as described in the last section, and traveler activity. Both k = 2 clusterings isolated the red subprefecture are seen in Figure 4. This subprefecture corresponds to Abidjan's location, the largest city and economic center of the Ivory Coast (355 of the 1031 cell phone towers were mapped to this subprefecture). Abidjan's prominant role would explain why its cell phone tower communication and traveler data are very different from other subprefectures. This does not imply that the other 254 subprefecture are similar, just that none of them are similar to Abidjan. In all clustering algorithms we used, we ran k-means numerous times with different initial random centroid placements to ensure Abidjan is a true singlet cluster. The silhouette scores of the remaining clusterings will be discussed in the next section. Figure 5 shows the similarity score for each day of the week using traveler activity in the feature vectors. The similarity scores are all higher than corresponding antenna communication similarity scores shown in Figure 2, indicating that travel does not seem to change as much as antenna communication from day to day. Both evolutionary and cumulative curves follow the same oscillatory pattern, unlike in Figure 2, implying that cumulative and dynamic traveler behavior are more consistent than the

 ISBN: 978-1-61208-358-2

corresponding communication behavior.

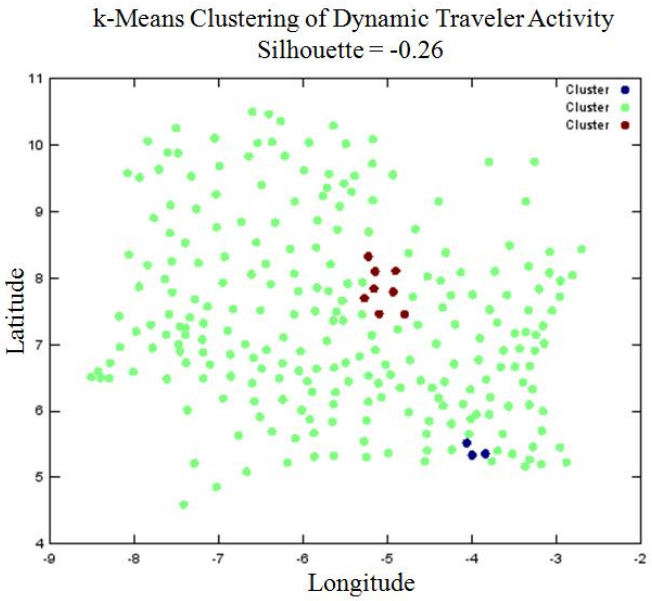

Figure 3. Example of k-means using Subprefecture Travel Data

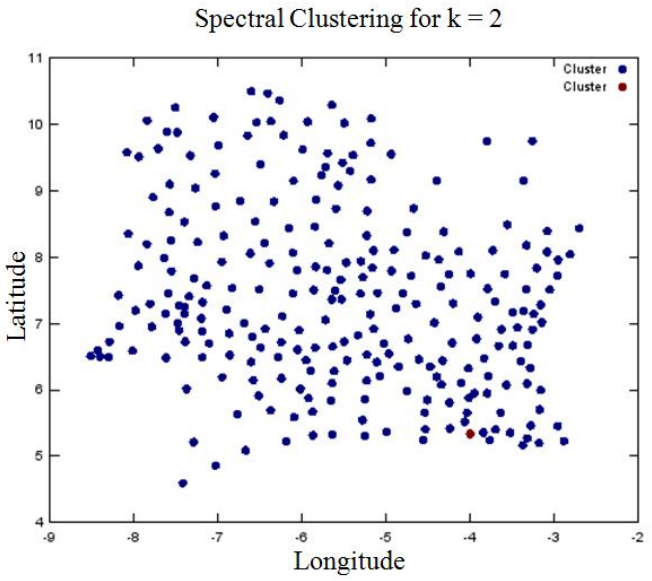

Figure 4. Spectral Clustering when k = 2 in cases of Subprefecture Communication and Travel Data

C. Multiple Edge Set Clustering

We concatenated the antenna communication and traveler feature vectors in the previous two sections to see the effect of our clustering methods on a network with more than one edge type. In Figure 6, we see the similarity scores computed for February 3 to February 9 using combined tower communication and traveler features. Both evolutionary and cumulative curves show less noisy patterns than in Figures 5 and 2 and we see a clear dip and minimum for both curves at Day 4. In Figures 7 and 8, we plot the silhouette values for each number of clusters for each combination of clustering

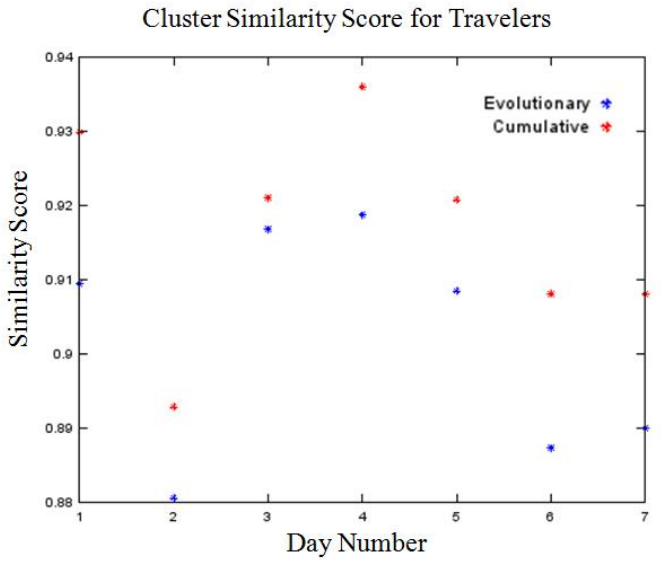

Figure 5. Similarity between Single Day Clusterings and Average Clusterings for the Week (Networks formed from Traveler Data)

algorithm and feature type. In Figure 7, we cluster cumulative activity and in Figure 8, we cluster the dynamic activity, both from February 3 to 9. Both plots are very similar, showing there is little difference between the strength of community structure between cumulative and evolutionary activity. There is a dramatic increase in silhouette values, the green curves, for the travel data from k-means to spectral clustering in Figures 7 and 8. The same is not true for the communication features and the combined communication and travel features (the red and blue curves). For both the red and blue points, spectral clustering silhouette values are very close to k-means silhouette values though there is marginal improvement when the number of clusters is more than 6. The improvement of spectral clustering over k-means depends substantially on the geometry of the feature values [5]; so, it is likely the case that the geometry of the antenna communication feature vectors is more conducive to k-means than the traveler feature vectors.

V. Conclusion

We applied clustering techniques to antenna communication and traveler data from February 3 to 9, 2012 between 255 Ivory Coast subprefectures. The optimum clustering for all feature and clustering algorithm combinations occurs when the number of clusters is 2 due to the unique central position of Abidjan (see Figure 4). While the cluster similarities scores were relatively high throughout the week in all cases, there was a smoother pattern seen in the case when communication and traveler features are combined though more work needs to be done to verify the cause of this. The consistency of the community structure over time can also be seen through the proximity between the strengths of the evolutionary and cumulative community structures (see Figures 7 and 8). Spectral clustering dramatically improved the community structure over k-mean clustering in the traveler feature space. However, the community structure over the combined traveler/communication feature space is only marginally better on average than that over the communication feature space. By adapting dynamic clustering techniques for networks with multiple edge sets,

 ISBN: 978-1-61208-358-2

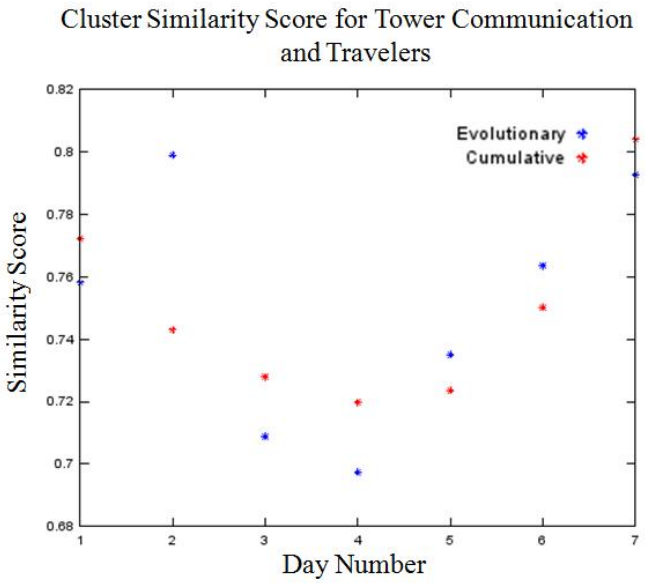

Figure 6. Similarity between Single Day Clusterings and Average Clusterings for 2/3 to 2/9 (Networks formed from Communication and Traveler Data)

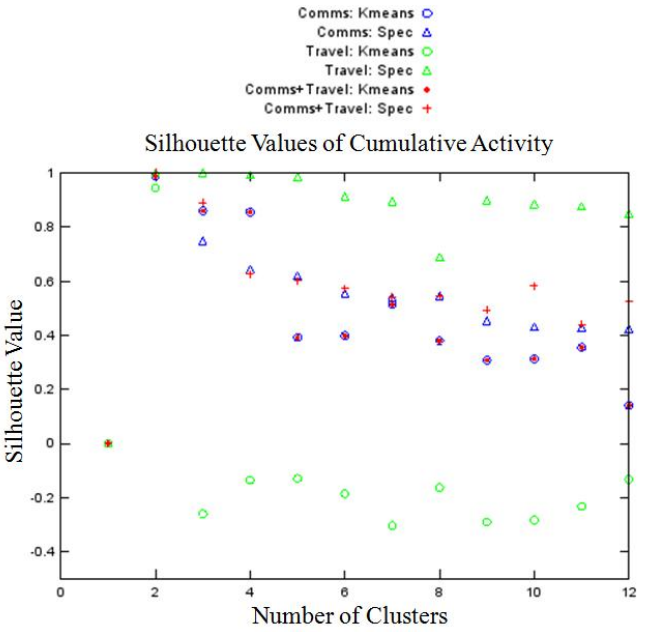

Figure 7. Silhouette Values of Cumulative Activity (2/3 to 2/9) Clusterings over Different Data Features and Algorithms

we were able to make sense of key spatial and temporal network attributes and propose new questions about clustering in heterogeneous networks.

ACKNOWLEDGMENT

The authors would like to thank the Sensemaking/PACOM Fellowship and Swansea University's Network Science Research Center for providing inspiration that led to this research and the opportunity to analyze this data set. Finally, we would like to thank Dr. Steve Chan, Jan-Kees Buenen, and Stef van den Elzenhave for advice regarding this paper.

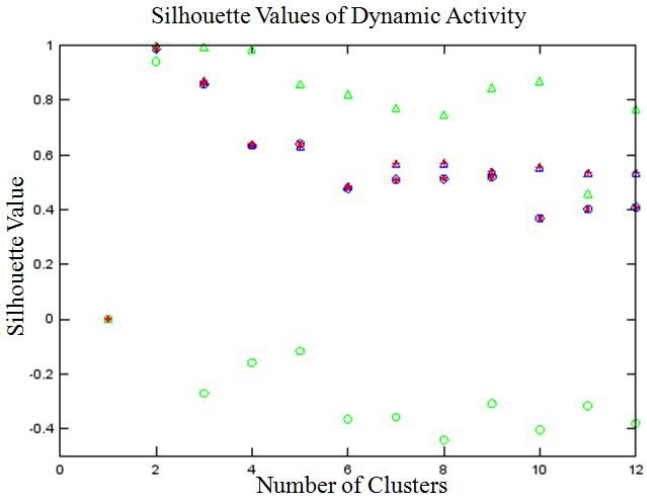

Figure 8. Silhouette Values of Dynamic Activity (2/3 to 2/9) Clusterings over Different Data Features and Algorithms

REFERENCES

[1] M. Newman, Networks, An Introduction. Oxford: Oxford University Press, 2010.

[2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng, "Evolutionary spectral clustering by incorporating temporal smoothness," in KDD Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2007, pp. 1–5.

[3] P. Grindrod and D. Higham, "Evolving graphs: Dynamical models, inverse problems, and propagation," in Proceedings of the Royal Society A, 2009, pp. 753–770.

[4] H. Jo, R. Pan, and K. Kaski, "Emergence of bursts and communities in evolving weighted networks," PLOS ONE, 2011, pp. 1–3.

[5] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in Neural Information Processing Systems (NIPS), vol. 14, 2002, pp. 1–6.

[6] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Computational and Applied Mathematics, vol. 20, 1987, pp. 53–65.

[7] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, and E. Huens, "Data for development: The d4d challenge on mobile phone data," arXiv:1210.0137v1 [cs.CY], Sep 2012, pp. 5–9.

 ISBN: 978-1-61208-358-2