Clustering and Visualization of Fuzzy Communities In Social Networks

Timothy C. Havens
Department of Electrical and Computer Engineering
Department of Computer Science
Michigan Technological University,
Houghton, MI 49931 USA
thavens@mtu.edu

James C. Bezdek, Christopher Leckie, Jeffery Chan, Wei Liu, James Bailey, Kotagiri Ramamohanarao, Marimuthu Palaniswami
Department of Computer and Information Systems
Department of Electrical and Electronic Engineering
University of Melbourne
Victoria 3010, Australia

Abstract— **We discuss a new formulation of a fuzzy validity index that generalizes the Newman-Girvan (NG) modularity function. The NG function serves as a cluster validity functional in community detection studies. The input data is an undirected graph G = (V, E) that represents a social network. Clusters in V correspond to socially similar substructures in the network. We compare our fuzzy modularity to an existing modularity function using the well-studied Karate Club data set.**

Keywords— fuzzy communities; community detection; modularity; fuzzy modularity; specVAT; clique discovery

I. INTRODUCTION

Suppose $O=\{o_1,\ldots,o_n\}$ denotes a set of n objects, usually, but not restricted to, humans (karate students, monks, southern women, etc.). Let $R=[r_{ij}]$ be a matrix of relational values on $O \times O$, r_{ij} being the relation between o_i and o_j. A common form of R arises as dissimilarity data, say $D = [d_{ij}]$, where d_{ij} is the pair wise dissimilarity between object vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbb{R}^p$, $d_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|$. In this case D is a symmetric matrix of distances. But for other types of (dis)similarity data, $d_{ij} = d(o_i,o_j)$ may not be symmetric, $d_{ij} \neq d_{ji}$. For example, Sampson's monastery data [1] is of this type. Breiger et al. [2] give the relationship from Bonhaven to Ambrose the value 2 in Sampson's data, but the value from Ambrose to Bonhaven in the opposite direction is 1. According to Wasserman and Faust [3] this is the most common form of social network data.

The Wasserman and Faust text is arguably the "bible" for social network analysis (18th printing, 2009), and yet, it does not mention fuzzy models of social networks! This is probably due to the well known disconnect between various communities of scholars working in related but uncommunicative fields. Selected readings in the literature from various groups indicate that this is probably quite accidental, most likely due to a lack of time to explore what may be essentially similar approaches advanced by disparate groups of researchers.

But many recent papers do exhibit fuzzy or possibilistic clusters in social networks. We will begin with a short review of the evolution of fuzzy models in social networks that culminates with current work about overlapping (fuzzy) communities in social networks. Then we will develop a new measure of fuzzy modularity for community detection, and compare it to an existing one using Zachary's Karate Club [4] data set.

II. FUZZY MODELS FOR SOCIAL NETWORKS

Social network analysis usually begins with a crisp (meaning not fuzzy, probabilistic or possibilistic) graph-theoretic representation of the social network, say G = (V, E, W), where V is the vertex set, E is the edge set, and W is the set of edge weights. Different social situations are realized by graphs with various properties: directed or not, weighted or not, connected or not, complete or not, and so on. In this note, G is undirected and weighted. Clusters (cliques, subtrees, etc.) in G (subsets of vertices in V) represent groups of individuals that are somehow related to each other more closely than to the individuals in the other clusters.

Any weighted graph can be thought of as a (possibly un-normalized) fuzzy graph, or a fuzzy similarity relation on pairs of nodes, first discussed by Zadeh in [5]. The earliest work on the use of fuzzy relations for social network analysis was Blin [6], who introduced the idea of using fuzzy relations in group decision theory. Bezdek et al. [7-9] collected data from small groups of students in communications classes, and developed models based on reciprocal fuzzy relations that quantified notions such as distance to consensus.

An idea that is gaining traction in social network analysis is the notion of overlapping communities in social networks [10]. Communities are defined as groups of densely interconnected nodes that are only loosely connected to the rest of the network in [11]. There is no clustering algorithm in [11]. Instead, overlapping clusters are seen visually as off-diagonal content in co-appearance images of the connection data. The model in [12] finds fuzzy communities by multicut spectral clustering. Clustering is done by both hard/fuzzy c-means (HCM/FCM, [13]) and validation is done with an index called fuzzy modularity by the authors.

So far, we have not described any method for finding a fuzzy c-partition of V, but once we have a set CPs, we have a means for assessing the quality of each candidate in it, namely Q_g. Brandes et al. [16] review "an array of heuristic algorithms that have been proposed to optimize modularity based on greedy agglomeration, spectral division, simulated annealing and extremal optimization, to name but a few prominent examples." None of the references given in [16] uses the explicit formulation for Q_h shown in (7). We conjecture here, but leave to another study [15], the possibility that imbedding Q_h in the more general setting afforded by Q_g will lead to a new, possibly better way, to maximize this popular index.

Several other formulas that are also called fuzzy modularity appear in the literature [12, 17]. We are interested here in the formulation due to Zhang et al. [12]. Their fuzzy version of (4) begins by partitioning V with spectral clustering applied to G using FCM once the eigenvector representation of G is selected. After a fuzzy c-partition $U \in M_{fcn}$ is found this way, they convert it to a possibilistic c-partition $U \in M_{pcn}$ of V as follows: they choose a threshold λ, (presumably $0 < \lambda < 1$), and extract from the k-th column of $U \in M_{fcn}$ the index set $V_k = \{i \mid u_{ik} > \lambda; 1 \le i \le c\}$. For each vertex i in V_k, the value u_{ik} in the fuzzy c-partition is replaced by a 1. After a pass over all n columns of U is completed, the remaining (non-1) memberships are set to 0. Doing this for k = 1 to n results in the conversion $U \in M_{fcn} \Rightarrow U_\lambda \in M_{pcn}$. Figure 1 is an example of the conversion procedure that illustrates the conversion for $\lambda = 0.10$ and $\lambda = 0.20$.

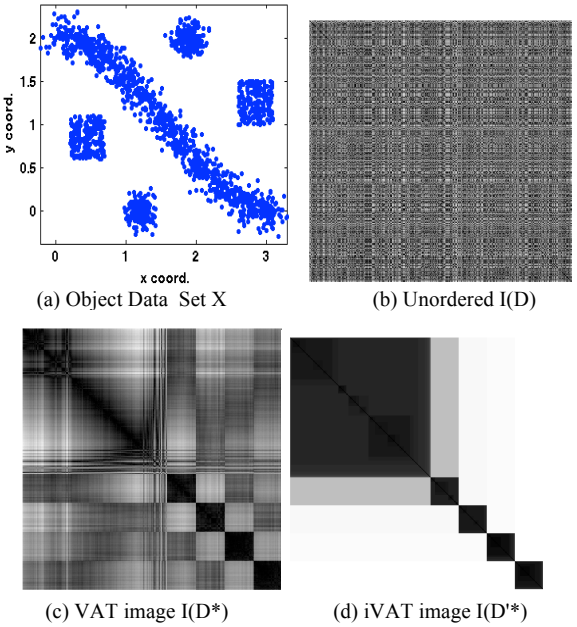
$$U = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{c} p_1 \quad p_2 \quad p_3 \quad p_4 \quad p_5 \\ \begin{bmatrix} 0.02 & 0.12 & 0.40 & 0.44 & 0.91 \\ 0.08 & 0.18 & 0.22 & 0.56 & 0.05 \\ 0.90 & 0.70 & 0.38 & 0.00 & 0.04 \end{bmatrix} \end{array}$$

$$U_{0.1} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad U_{0.2} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Figure 1. Illustration of Zhang et al.'s Conversion

Imagine that FCM is applied to the spectral data from a network of 5 people $\{p_k\}$ at c = 3, and terminates at the fuzzy 3-partition U shown in Figure 1. Then Zhang et al.'s conversion yields the possibilistic 3-partitions $U_{0.1}$ and $U_{0.2}$ shown below U in the figure. Zhang et al. interpret "fuzzy communities" in $U_{0.1}$ as follows. Persons '2' and '3' belong to all three groups; '4' belongs to groups 1 and 2, while '1' is only in group 3 and '5' is only in group 1.

Now consider the second partition shown in Figure 1. When λ is increased from 0.1 to 0.2, joint membership in fuzzy communities is more stringent. Now only person '3' belongs



(a) Object Data Set X (b) Unordered I(D)

(c) VAT image I(D*) (d) iVAT image I(D'*)

Figure 2. VAT/iVAT images of Boxes and Stripe

to all three groups, '4' belongs to groups 1 and 2, '5' belongs to group 1, while '1' and '2' are in just group 3. Thus, the joint membership of an individual in various communities is a function of the threshold λ. Zhang et al. do not specify the range of λ, but it must be $0 < \lambda < 1$, for otherwise the bounds of this conversion procedure would be $[\mathbf{1}]_{cxn}$ at $\lambda = 0$ and $[\mathbf{0}]_{cxn}$ at $\lambda = 1$. There are (infinitely) many candidate partitions available from this procedure because we can apply this process to candidates generated by FCM at each c = 2, 3, ... n-1; and within each c, for any $0 < \lambda < 1$. Zhang et al. choose a "best" possibilistic U by maximizing their version of the fuzzy modularity index, which is defined as follows:

$$V_k = \{i \mid u_{ik} > \lambda; 1 \le i \le c\}; \quad k = 1, \ldots, n \quad , \tag{8a}$$

$$S_z(V_k, V_k) = \left(\sum_{i,j \in V_k} \frac{u_{ik} + u_{jk}}{2} \right) w_{ij} \quad ; \tag{8b}$$

$$S_z(V_k, V) = S_z(V_k, V_k) + \left(\sum_{\substack{i \in V_k \\ j \in V \setminus V_k}} \frac{u_{ik} + (1 - u_{jk})}{2} \right) w_{ij} \; ; \tag{8c}$$

$$Q_z = \sum_{k=1}^{c} \left(\frac{S_z(V_k, V_k)}{S(V,V)} - \left(\frac{S_z(V_k, V)}{S(V,V)} \right)^2 \right) . \tag{8d}$$

The values $\{u_{ik}\}$ appearing in (8b) and (8c) are from the fuzzy c-partition before possibilistic conversion. Clearly, (8b) and (8c) reduce to $S(V_k,V_k)$ and $S(V,V_k)$, respectively, for crisp partitions (assuming $0 < \lambda < 1$). Hence, $Q_z = Q_h$ for crisp partitions. However, we believe that Q_z has theoretical problems, which we outline in [15]. Here, we are content to compare the analysis of two real data sets using the indices Q_z and Q_g.

IV. VISUAL CLUSTER TENDENCY ASSESSMENT

Let D be a set of square or rectangular dissimilarity data. The idea of visually analyzing the rows and/or columns of D to reveal structural relationships between individuals associated with D began with Loua [18] in 1873. The first reordered dissimilarity image (RDI) of a square data matrix appears in Czekanoski [19]. The methods for constructing and using RDIs in various applications have subsequently grown almost without bounds. Wilkinson and Friendly [20] survey contemporary methods using this idea in bioinformatics, where the reordered image is called a "cluster heat map." These authors state that this method has appeared in more than 4,000 papers in the last decade. Liiv [21] gives a very useful survey of seriation methods for social network analysis.

The visual assessment of tendency (VAT, [22]) model reorders symmetric, square D to D* using the indices of a minimal spanning tree on D, and then displays a heatmap image I(D*) of D* (often a gray-scale image). The basic rationale for VAT is that if an object tends to cluster with other objects, then it should also be part of a submatrix of "similarly small" values corresponding to those objects. These submatrices are seen as dark blocks along the diagonal of the VAT image I(D*). Contrast can be improved by setting the diagonal to the minimum of the off-diagonal values. Zhang et al. [23] discuss the use of VAT in a product called RoleVAT, a role engineering tool for role based access control. Improved VAT (iVAT, [24]) transforms D to D' using geodesic distances to replace the input distances, followed by VAT reordering of D' to D'*. We will also discuss a new adaptation of specVAT, a member of the VAT family related to spectral clustering [25].

Figure 2 illustrates VAT / iVAT using the 2D object data set X called Boxes and Stripe in [24]. View 2(a) scatterplots X, which is converted to the symmetric matrix D using the Euclidean norm, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$. Figure 2(b) is the image I(D). Figure 2(c) is the VAT reordered image I(D*), and Figure 2(d) is the iVAT reordered image I(D'*). Most observers would agree that there are c = 5 pretty distinct clusters in Boxes and Stripe. This substructure is not evident in I(D). The VAT image I(D*) does a better job at highlighting the structure, but the upper left block along the diagonal, which corresponds to the stripe cluster in X, is not so clear. More generally, the image in view 2(c) lacks clarity—there is not much contrast between the on and off diagonal blocks. Replacing Euclidean distances in D by geodesic distances in D' prior to VAT reordering with iVAT recursion renders the substructure in X quite nicely, as seen in Figure 2(d).

However, we will see that iVAT doesn't work as well for the Karate Club data set because social network data in the form of the graph G = (V, E, W) don't respond well to iVAT reordering. In brief, the conversion of G to a distance matrix D is fraught with problems—we describe this problem more in detail in the next section. We now turn to describing the Karate Club Data and analyzing the use of various visual clustering tools, including iVAT and specVAT, and fuzzy modularity.

V. NUMERICAL EXAMPLE: KARATE CLUB DATA

The Karate Club data is an undirected graph $G = (V, E, W)$ with 34 vertices that show links between the 34 members of a university karate club collected by Zachary in 1977 [4]. Edge weight w_{ij} indicates the relative strength of the association between individuals i and j (number of situations in and outside the club in which interactions occurred). The maximum value in W is 7 for the edge between members 26 and 32.

The Karate Club data is a favorite amongst social network-ists, because the evolution of the relationship between pairs of members in the Karate club—which was known and recorded by Zachary—provides a sort of "ground truth" for various social network analyses.

Zachary used these data and an information flow model of network conflict resolution to explain the split of this group into two factions (the squares and circles) following disputes among the members. The principals in the split were the karate instructor (vertex 1) and the president of the club (vertex 34).

Figure 3 shows Zachary's karate club network as depicted by Newman and Girvan [14]. Square nodes represent the instructor's faction and circular nodes the presidents's faction. The original belief was that this network should decompose well into two clusters as shown in Figure 3, because its members did bifurcate, following either the president or the instructor. However, various discussions of this data in the literature disagree.

Figure 4(a) is the image of the matrix $D = [W$ (the diagonal of D is also set to 0 following this transformation) where W is the matrix of edge weights for the graph G. Figure 4(b) is the iVAT image of D. View 4(b) suggests that the Karate club has three pretty tight clusters (the red blocks), overlain by a weaker and larger orange block. The orange block is inside an even larger yellow block. Five individuals at the bottom and one at the top are isolated pixels in the iVAT image, indicating

Figure 3. Karate club social network [14]

(a) Data image $I(D)$ (b) iVAT image $I(D'_k)$

Figure 4. iVAT image of Karate club data

non-association with the other 28 members of the club. However, the implications of this representation of the data are wrong. Essentially, the transformation of W to the distance matrix D considers the edges with non-valued weight $w_{ij} = 0$ to have finite distance $d_{ij} = 7$, where, we argue, that these distances should be infinite (i.e., the absence of a path between i and j).

We examined other transformations of W into D, such as artificially increasing the distances corresponding to zero-weight edges, but did not achieve a pleasing result. Hence, we now turn to a spectral method for visualizing cluster tendency.

To alleviate the problem that we see with using VAT / iVAT with these data, we will use specVAT [25], which displays a dissimilarity image of the spectral components of a normalized weight matrix. specVAT first computes the top k eigenvectors of the eigenvalue problem

$$Lx = \lambda x, \qquad (9)$$

where $L = M^{-1/2} W M^{-1/2}$ and M is the n x n diagonal matrix with the vector $m = W1_n$ on the diagonal. Then each of the vectors are normalized to the unit hypersphere by l_2 normalization. Finally, the distance matrix D is computed by $d_{ij} = [||(x_i - x_j)||]$. VAT (and iVAT) can then applied to D to visualize the clustering tendency of W.

c = 2 c = 3 c = 4

(a) New formulation of specVAT using Eq.(10)

c = 2 c = 3 c = 4

(b) Old formulation of specVAT using Eq.(9)

Figure 5. specVAT images of Karate Club data

We generalize specVAT by computing the eigenvectors with the generalized eigenvalue problem

$$Wx = \lambda Mx. \qquad (10)$$

It can easily be shown that (10) is equivalent to (9), but for most eigensolvers, (10) is a more stable instantiation because small elements on M do not induce numerical stability issues.

Figure 5 compares the use of (9) and (10) in creating the specVAT images using c = 2, 3, and 4 eigenvectors for the Karate Club data. It is somewhat clear from the specVAT images in view 5(a) that the Karate Club data have 3 clusters. There seem to be three dark blocks for c = 2 and c = 3 images. At c = 4, the image begins to break down which indicates that there should not be cluster structure at c = 4. In the specVAT images, shown in view 5(b), the cluster structure is less apparent. Interestingly, Figure 5 shows that although (9) and (10) are mathematically equivalent, their instantiation in the eigensolver can prove to drastically alter the results.

Although the specVAT images in Figure 5(a) seem to suggest 3 clusters, we wanted to test how the iVAT geodesic distance transform could be applied to improve the visualization. We applied the transformation to the distance matrix computed by specVAT. We call the formulation of specVAT using (10) with iVAT, specieVAT for spectral improved eigensolver VAT. These images are shown in Figure 6. It is now very clear, by the images in view 6(a), that specieVAT reinforces the popular viewpoint that the Karate Club data have 3 clusters. However, the views in 6(b), made by applying the iVAT transformation to the old formulation of specVAT, do not seem to suggest any visually pleasing (relative to the precedent) cluster structure in the data set. We see this as further evidence that specVAT using (10) is superior to the original formulation, albeit mathematically the same. Furthermore, this shows that using the iVAT transformation with the new formulation of specVAT improves the visualization.

c = 2 c = 3 c = 4

(a) specieVAT using Eq.(10)

c = 2 c = 3 c = 4

(b) specVAT using Eq.(9) with iVAT

Figure 6. specieVAT and specVAT + iVAT images of Karate Club da

belief; however, we would say that number of clusters in this data set is uncertain. If we simply add up the columns of Table I, c = 4 is the clear choice. But if we examine the plot in Fig. 9, Q_g has the sharpest peak, occurring at c = 3. This is a prime example of the cluster validity conundrum—which partition do I choose and which validity measure do I trust? There is no perfect answer to this question, but this experiment does point to a underlying issue in Z_{XQ}—namely, how does one choose "? In this case, it would be easy, in hind sight, to say that one should choose " = 0.25—this value of " chooses the preferred c = 3. However, what works for this data set may not work for every data set; hence, we prefer the parameter-free generalization Q_g.

VI. Conclusions

Our generalization of Newman's modularity function at (7) represents a step forward in the problem of finding fuzzy communities in graph data. First, it is a direct generalization of the underlying theory of graph modularity: namely, providing a measure of the probability of a graph structure relative to a null hypothesis. For more on this discussion, please refer to [15]. Second, it is a parameter-free instantiation of fuzzy modularity: the only of its kind to date. Last, we showed that on the well-studied Karate Club data set that our fuzzy modularity index performs comparably to Newman's crisp modularity function and to Zhang et. al's fuzzy index.

The second point of progress in this paper is the reformulation of specVAT into the new algorithm, specieVAT, for determining the number of clusters in graph-based data. First, we showed that specVAT can be reformulated as a generalized eigenvalue problem at (10). This formulation is mathematically equivalent to the original specVAT, but is numerically more stable for most, if not all, eigensolvers. Second, we showed how the iVAT distance transform could be applied to specVAT to improve the tendency assessment visualization.

In the future, we will examine how our fuzzy modularity at (7) can be directly maximized. Initial work in this direction indicates that (7) can be posed as a generalized eigenvalue problem much like (10). Our current efforts are focused on establishing this maximization problem within the existing spectral clustering framework and on stabilizing the numerical issues involved in reaching the solution. Second, we will examine how fuzzy modularity can be generalized as a cluster validity index (or quality function) for asymmetric adjacency matrices or directed graphs. Finally, we will do a comprehensive comparison of various validity indices and clustering algorithms on a wide assortment of graph-based and social network data. As always, we believe that there is no free lunch in the clustering game and that the best attacks at these problems will involve a merger of clustering tools.

References

[1] S.F. Sampson, "A novitiate in a period of change: an experimental and case study of social relationships. Unpublished Doctoral Dissertation. 1968.

[2] R. Breiger, S. Boorman, P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," Journal of Mathematical Psychology, vol. 12, 1975, pp. 328-383.

[3] S. Wasserman and K. Faust, Social Network Analysis. Cambridge: Cambridge University Press, 1994.

[4] W. Zachary, "An information flow model for conflict and fission in small groups. Journal of Anthropological Research, vol. 33, 1977, pp. 452-473.

[5] L. A. Zadeh, "Similarity relations and fuzzy orderings," Information Sciences, vol. 3, 1971, pp. 177-200.

[6] J. M. Blin, "Fuzzy relations in group decision theory," J. Cybernetics, 1974, pp. 17-22.

[7] J. C. Bezdek, B. Spillman, B. and R. Spillman, "A Fuzzy Relation Space for Group Decision Theory," Fuzzy Sets and Systems, vol. 1, 1978, pp. 255-268.

[8] J. C. Bezdek, B. Spillman and R. Spillman, "Fuzzy relation spaces for group decision theory: An application," Fuzzy Sets and Systems, vol. 2, 1979, pp. 5-14.

[9] B. Spillman, J. Bezdek and R. Spillman, "Development of an Instrument for the Dynamic Measurement of Consensus," Comm. Mono., vol. 46, 1979, pp. 1-12.

[10] S. Fortunato, "Community detection in graphs," Phys. Rep., vol. 486, 2010, pp. 75-174.

[11] J. Reichardt and S. Bornholdt, "Detecting fuzzy community structures in complex networks with a Potts model," Stat. Mech, 2004.

[12] S. Zhang, R.S. Wang and X.S Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering," Stat. Mechanics and its Appl., vol. 374(1), 2007, pp. 483-490.

[13] J.C. Bezdek, J.M. Keller, R. Krishnapuram, and N.R. Pal, Fuzzy models and algorithms for Pattern Recognition and Image Processing. Springer: NY, 1999.

[14] M. E. J. Newman and M. Girvan, Phys. Rev. E 69(2), 2004.

[15] T. C. Havens, J. C. Bezdek, C. Leckie, K. Ramamohanarao, and M. Palaniswami, "A soft modularity function for detecting fuzzy communities in social networks," IEEE Trans. Fuzzy Systems, 2013, doi:10.1109/TFUZZ.2013.2245145

[16] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski and D. Wagner, "On Modularity Clustering," IEEE TKDE, vol. 20(2), 2008, pp. 172-188.

[17] J. Liu, "Fuzzy modularity and fuzzy community structure in networks," Eur. Phys. J. B, vol. 77, 2010, pp. 547-557.

[18] T. Loua Atlas statistique de la population de Paris, J. Dejey, 1873.

[19] J. Czekanowski, "Zur differential-diagnose der Neandertalgruppe," Korrespondenzblatt der Deutsche Gesellschaft für Anthropologie, Ethnologie und Urgeschichte, vol. 40, 1909, pp. 44-47.

[20] L. Wilkinson and M. Friendly, "The history of the cluster heat map," The American Statistician, vol. 63(2), 2009, pp. 179-184.

[21] I. Liiv, "Seriation and Matrix Reordering Methods: An Historical Overview," Stat. Anal. and Data Mining, vol. 3(2), 2010, pp. 70-91.

[22] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," Proc. IJCNN, 2002, pp. 2225-2230.

[23] D.Zhang, K. Ramamohanarao, S. Versteg and R. Zhang, "RoleVAT: Visual Assessment of Practical Need Role Based Access Control," Proc. IEEE Conf. on Security Apps, 2009, pp. 13-22.

[24] T. C. Havens and J. C. Bezdek, "An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm," IEEE TKDE, vol. 24(5), 2012, pp. 813-822.

[25] L. Wang, X. Geng, J.C. Bezdek, C. Leckie and K. Ramamohanarao, "Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning" IEEE TKDE, vol. 22(10), 2010, pp. 1401-1414.

[26] D. Verma and M. Meila, "A comparison of spectral clustering algorithms," UW CSE Technical Report, 03-05-01, 2003.