Abstract—The clustering method described in this paper is meant to aid the researcher who, for any of a variety of reasons, has clustered a data set a large number of times and is now faced with the problem of reconciling many different clustering results into a single, robust clustering solution.

In this paper a clustering method is developed that takes the results of these many data clusterings, properly arranges them to form a similarity matrix, and then uses the variable aggregation ideas of Herbert Simon and Albert Ando to provide a single solution. Once the germane parts of Simon-Ando theory are reviewed, an algorithm is developed and tested.

I. Problem Description

For a variety of reasons, we may cluster a data set multiple times. We may use a clustering method (e.g. $k$-means or non-negative matrix factorization (NMF)) that does not return a unique solution and wish to see a variety of such solutions. We may be unsure of which, among the many clustering algorithms, is best suited to the application, and so decide to cluster the data with a number of them. We may be unsure of $k$, the number of clusters to be found, and so decide to repeatedly run an algorithm using a range of reasonable values. Lastly, we may be overwhelmed by the large number of parameter values most commercially available clustering software allow and decide to repeatedly run the algorithm, each time with a different parameter set. Note that these reasons are not incompatible as we may cluster a data set many times for any combination of these reasons. The clustering problem becomes one of reconciling these multiple results to obtain a final answer.

One approach for solving this problem is to determine a clustering that is as close as possible to all the clusterings already found. This is an optimization problem known as median partition which is NP-complete. A number of heuristics for the median partition problem exist and a comparison of their performance can be found in [1].

Another approach that has been studied is a consensus clustering framework based on variational Bayes Mixture of Gaussians [2].

A third approach involves storing all the clustering results in matrix form. For example, the result from each run of a clustering algorithm is stored in an adjacency matrix $A$, where $A_{ij} = 1$ if data set elements $i$ and $j$ are in the same cluster and $A_{ij} = 0$ otherwise. The collection of these adjacency matrices can be used to represent the connections between the original data as a hypergraph, and clusters can be discovered through hypergraph partitioning algorithms [3].

Alternatively, we can create a matrix $S$, the sum of all these adjacency matrices, and then aim to cluster the original data based on the contents of $S$ (often each element of $S$ is divided by the number of clustering runs to create a matrix with entries in the interval $[0, 1]$). Methods that have been used for this problem include clustering $S$ using single-link hierarchical clustering after zeroing out entries below a certain threshold [4].

A new approach introduced in this paper is based on the observation that the structure of $S$, as constructed above, is that of a nearly completely decomposable matrix, and thus the results of Simon-Ando theory are applicable. Simon-Ando
theory assumes that we have knowledge of the structure of a system (i.e. the clusters), which allows us to make predictions about the behavior of the system over time. The key insight for creating this new clustering method is that we can consider the problem from the opposite direction, that is, we can look at the behavior of the system over time and use it to determine the clusters.

II. OVERVIEW OF SIMON-ANDO THEORY

Simon and Ando [5] developed the theory for the behavior of the vector \( x_t \) defined by

\[
x_t = x_{t-1} P. \tag{1}
\]

\( P \) is a nearly completely decomposable matrix defined as

\[
\begin{pmatrix}
P_{11} & P_{12} & \ldots & P_{1k} \\
P_{21} & P_{22} & \ldots & P_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
P_{k1} & P_{k2} & \ldots & P_{kk}
\end{pmatrix},
\]

where the diagonal blocks are square and the off-diagonal blocks \( P_{ij} \) contain sufficiently small entries.

When considering the long-term behavior of this system, it is often helpful to think of equation (1) by its equivalent expression

\[
x_t = x_0 P^t, \tag{2}
\]

since (2) emphasizes the fact that changes in \( x_t \) are the result of changes in powers of \( P \).

Though they use differing vocabulary, both Cortois [6] and Meyer [7] clarify one of the key conclusions of Simon-Ando theory, namely that as \( t \) increases, \( x_t \) travels through four stages:

1) Short-term dynamics - at first the relatively large terms in diagonal blocks \( P_{ii} \) cause large changes in \( x_t \).
2) Short-term equilibrium - a “settling down” period after the initial effects of the main diagonal blocks but before the effects of the off-diagonal blocks are seen.
3) Long-term dynamics - changes to \( x_t \) influenced by the off-diagonal blocks.
4) Long-term equilibrium - \( x_t \) becomes stable and all changes induced by \( P \) are complete.

For clustering purposes, the key stage is short-term equilibrium.

III. CLUSTERING USING SIMON-ANDO THEORY

For reasons that will appear later in this section, to use Simon-Ando theory as a clustering method, matrix \( P \) must be doubly stochastic. The theory put forth by Sinkhorn and Knopp [8] provides the framework for converting the similarity matrix \( S \) into \( P \) through the creation of diagonal matrices \( D_1 \) and \( D_2 \) such that \( P = D_1 S D_2 \). Although there are theoretical conditions that must be met to assure conversion is possible, this has not been an issue while running test cases. However, to cover any eventuality, the algorithm developed here will respond to non-convergence of the Sinkhorn-Knopp process by adding a small positive constant (one one-hundredth times the number of clusterings being used) to each element of \( S \). By doing so we assure that the similarity matrix has total support [8], a necessary and sufficient condition for the existence of \( D_1 \) and \( D_2 \).

Given \( x_0 \), a random initial probability vector, Markov theory assures us that if \( P \) is \( n \times n \), then over time \( x_t \) will approach the uniform probability vector \((1/n, 1/n, \ldots, 1/n)\). In the language of Simon-Ando, the uniform probability vector is a sign that the system has reached long-term equilibrium. Knowing when long-term equilibrium is reached is important, since it makes certain that the clustering algorithm has a fixed criteria for when it has gone too far.

In practice, the algorithm rarely iterates equation (1) until long-term equilibrium occurs, as it almost always recognizes that \( x_t \) is in short-term equilibrium. This state is identifiable since a subset of the elements of \( x_t \) are all approximately equal to a certain value, while a disjoint subset is approximately equal to a different value, continuing similarly up to the number of clusters.

IV. METHOD

This Simon-Ando inspired clustering method is implemented through the following steps:

1) The data set is clustered 100 times using the multiplicative update version of the non-negative matrix factorization algorithm [9].
fixed value of $k$ was used for all 100 runs of the NMF code.

2) The similarity matrix $S$ is constructed where the entry $S_{ij}$ is equal to the number of times data elements $i$ and $j$ are in the same cluster ($0 \leq S_{ij} \leq 100$).

3) The Sinkhorn-Knopp algorithm is used to convert $S$ into a doubly stochastic matrix $P$.

4) An initial probability vector $x_0$ is created randomly.

5) The calculation $x_{t+1} = x_t P$ is repeated until a stopping criteria is met. At this time in the algorithm’s development different stopping criteria are being tested.

6) The entries in the final $x_t$ are put into $k$ clusters using hierarchical clustering. This is equivalent to sorting the elements of $x_t$ based on the $k-1$ largest gaps in the sorted list. Since hierarchical clustering is susceptible to outliers it sometimes creates singleton clusters. When this happens the algorithm creates a new random initial probability vector and repeats steps 5 and 6.

V. RESULTS

The algorithm was tested on a DNA microarray data set containing gene expression levels for 5000 genes in bone marrow samples from 38 leukemia patients [10]. The patients had known diagnoses of either acute myelogenous leukemia (AML) or acute lymphoblastic leukemia (ALL). In addition, ALL has T and B cell subtypes, making this data set an excellent test case for either $k = 2$ (AML, ALL) or $k = 3$ (AML, ALL-T, ALL-B). In this data set, samples with the same diagnosis are grouped together. The clustering algorithm only groups like samples and cannot label each group with its diagnosis. Thus it is assumed that after clustering, a person familiar with the data set would label each cluster with a diagnosis.

Table I and II summarize the result that the clusters found by this new stochastic method, by the well-known Fielder method and by clustering $S$ using NMF are identical for both $k = 2$ and $k = 3$. This statement is made with the acknowledgment that when using the Fielder method to find three clusters, the $k = 2$ solution is examined, a cluster removed from the data set and the remaining cluster split in two. This intervention can certainly lead to a user “guiding” the method to a “correct” answer.

<table>
<thead>
<tr>
<th>Type</th>
<th>Cluster Members Found</th>
<th>Correct Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>1-5, 7-27, 29</td>
<td>1-27</td>
</tr>
<tr>
<td>AML</td>
<td>6, 28, 30-38</td>
<td>28-38</td>
</tr>
</tbody>
</table>

Number Mis-clustered: 2

<table>
<thead>
<tr>
<th>Type</th>
<th>Cluster Members Found</th>
<th>Correct Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL-B</td>
<td>1-9, 11-19, 29</td>
<td>1-19</td>
</tr>
<tr>
<td>ALL-T</td>
<td>10, 20-27</td>
<td>20-27</td>
</tr>
<tr>
<td>AML</td>
<td>28, 30-38</td>
<td>28-38</td>
</tr>
</tbody>
</table>

Number Mis-clustered: 2

We consider an additional test case to determine if these methods are in fact equivalent.

Using data from [11], an $88 \times 48$ data matrix was created. Each column represents one of the contiguous U. S. states and each row represents a presidential candidate in each of the elections from 1912 through 2008. The matrix entries are the number of votes that candidate received in that state.

Clustering the resulting $S$ with $k = 2$ yields the same result using either Fielder or NMF. With few exceptions, one cluster contains the Northeastern and Midwestern states, while the other cluster contains states of the South and West. The two clusters can be seen in Figure 1.

For this clustering, the new stochastic clustering method disagrees with the other two methods in one case: it places Minnesota with the other Midwestern states.

Figure 2 shows how the results for NMF and the stochastic method differ when $k = 3$. NMF formed its new cluster with a few states from each of its previous two clusters, while the stochastic method’s new cluster is a subset of one of its previous clusters. No result is shown for the Fielder method because of the previously mentioned ad hoc approach when $k = 3$. 
using clustering algorithms other than NMF or ensembles of algorithms to create S. Other topics for further investigation include using spectral properties of $P$ to select an appropriate $k$, refining the algorithm’s stopping criteria, and choosing good initial probability vectors.

REFERENCES


VI. CONCLUSION

The Simon-Ando inspired clustering method described in this paper was designed with the researcher confronted with multiple clustering results in mind. The method uses all the available input in a natural way to arrive at a final clustering.

Future work will compare the effectiveness of