involve

a journal of mathematics

Seriation algorithms for determining the evolution of *The Star Husband Tale*

Crista Arangala, J. Todd Lee and Cheryl Borden

msp

2014 vol. 7, no. 1

INVOLVE 7:1 (2014)
dx.doi.org/10.2140/involve.2014.7.1

Seriation algorithms for determining the evolution of *The Star Husband Tale*

Crista Arangala, J. Todd Lee and Cheryl Borden

(Communicated by Kenneth S. Berenhaut)

We give an introduction to seriation techniques and apply such techniques to the North American folklore tale known as the *Star Husband Tale*. In particular, a spectral algorithm with imposed clustering is applied, with significant results that support the algorithm's effectiveness.

1. Introduction

In the field of archeology, many researchers investigate whether objects can be chronologically ordered based strictly on their physical characteristics, in a process known as seriation. Typically, one tries to arrange artifacts from numerous sites in sequential order.

A variety of seriation techniques are in common use; historical reviews can be found in [Lyman et al. 1998; O'Brien and Lyman 2002]. One common technique, known as frequency seriation, is based on the relative frequency of artifact types. Other seriation techniques, such as occurrence and phyletic seriation, are based on similar characteristics between artifacts. The idea behind them is that one can use the presence or absence of certain characteristics, or attributes, in particular digs in order to order the artifacts chronologically. These seriation techniques, introduced in [Petrie 1899], are based on a binary incidence matrix (called the Petrie matrix) and attempt to minimize dissimilarities between digs by ordering them appropriately. (Another class of seriation techniques, which will not concern us, involves what's called phylogenetic trees. See [Buneman 1971; Huson and Bryant 2006] for information.)

Despite their popularity in archeology, seriation techniques have not been widely used in studying the geographical spread of stories and folklore. In this paper, we use a dissimilarity approach to seriate a well-known North American Indian folk tale, the *Star Husband Tale* (see Section 2). We do this both by brute force and by using an elegant spectral algorithm from [Atkins et al. 1999]. We show that,

MSC2000: 15-04, 15A18.
Keywords: seriation, linear algebra, folk tales.

based strictly on the content dissimilarities among versions, one can track the tale's progression in a way that matches the geographical proximity of the tribes telling these tales.

Organization of the paper. Section 2 describes the tale and mentions some prior studies. A basic seriation technique is described in Section 3 and results from its application to a small subset of the data are analyzed in Section 5. To be able to study a larger data set, we discuss in Section 4 a spectral algorithm from [Atkins et al. 1999]. As shown in Section 6, this algorithm is able to order and cluster successfully a larger data set consisting of eighty-six versions of our tale.

2. The *Star Husband Tale*

The basic form of the *Star Husband Tale* [Young 1978] tells of two girls who are sleeping out in the open during the night. While outside, they see two stars and each girl makes a wish to be married to a star. When they awake, both have been transported to the heavens and are married to the stars as they wished. One of the star husbands is a young man and the other is an older man. Heedless of a warning they've received, the girls at one point start digging in the heavens and make a hole through which they can see their old homes below. Overcome with homesickness, they eventually lower themselves down to earth using a rope.

Dundes [1965] discusses various narrative elements peculiar to eighty-six versions of the *Star Husband Tale* coming from 44 tribes throughout North America. These tribes are grouped into nine geographical zones: Eskimo, Mackenzie, North Pacific, California, Plateau, Plains, Southeast, Southwest, and Woodlands [Carroll 1979]. Thompson chose those characteristics that occur most frequently as the principal tale elements. These principal tale elements are then collated into archetypes and subarchetypes [Rich 1971]. A sample of the principal tale elements followed by their archetypes is presented in Table 1.

Table 2 gives a sample of the records (tribes) and the traits that are present (or absent) in their versions of the tale.

In total, 86 versions of the *Star Husband Tale* and a total of 135 traits (archetypes and subarchetypes) are included in the study.

3. Seriation

In this section we express the seriation problem mathematically: how to list a set of objects so as to minimize the sum of the dissimilarities between consecutive objects. This is the traveling salesman problem: our objects (versions of the tale) correspond to cities, and the measure of dissimilarities corresponds to distances. However we can make a simplifying assumption (which often holds only approximately) that makes the problem easier than the general traveling salesman problem.

Trait A: Number of women		Trait B: Introductory action	
A1	One	B1	Trait not present
A2	Two	B2	Wish for star husband
A3	Two at first, then one	B3	Pursuit of porcupine
A4	More than two	B4	Miscellaneous
Trait D: Method of ascent		**Trait H: Taboo broken in Upper World**	
D1	Not indicated	H1	No taboo broken
D2	Stretching tree	H2	Digging or disturbing ground
D3	Translation during sleep	H3	Moving a large rock
D4	Carried through the air	H4	Looking somewhere
D5	Carried in a basket	H5	Shooting a meadow lark
D6	Carried by whirlwind	H6	Making noise before an animal sings
D7	Carried by a feather		

Table 1. Sample of traits and their archetypes. Some archetypes for traits B, D, and H are further subdivided (H1, H1a...).

Tale 1	Eskimo	Smith Sound	A1	B3a	D3	H2
Tale 2	Eskimo	Kodiak 1	A3	B1	D3,4	H2
Tale 3	California	Patwin	A1	B3a	D6	H1
Tale 4	California	Washo 1	A2	B1	D2	H1a
Tale 5	North Pacific	Snuqualmi 1	A2	B1	D2	H1
Tale 6	North Pacific	Snuqualmi 2	A2	B1	D2	H1
Tale 7	Plains	Sarsi	A3	B1	D3	H1a
Tale 8	Plains	Blackfoot 1	A3	B1	D3,7	H1a

Table 2. Classification of a sampling of tales with respect to the traits listed in Table 1. More than one archetype or subarchetype can be present for a given trait in a tribe's version of the tale.

The first step is to express the information in an *incidence matrix*. For the data in Table 2, this matrix is

$$A = \begin{array}{r|ccccccccccccc|}
 & \text{A1} & \text{A2} & \text{A3} & \text{B1} & \text{B3a} & \text{D2} & \text{D3} & \text{D4} & \text{D6} & \text{D7} & \text{H1} & \text{H1a} & \text{H2} \\
\text{Tale 1} & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{Tale 2} & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
\text{Tale 3} & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
\text{Tale 4} & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\text{Tale 5} & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\text{Tale 6} & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\text{Tale 7} & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
\text{Tale 8} & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0
\end{array}.$$

Each row corresponds to a tale, and each column to a trait that the tale may or may not possess. The measure of dissimilarity between tales is the number of places where the corresponding rows differ.

If the rows were written in the order in which the tale evolved, it would be natural to expect 1s to cluster together in each column: traits may be introduced or removed as the tale is handed down, but a given trait is unlikely to jump in and out of existence several times. We formalize this with the following concept:

Definition. A binary matrix A is called a *Petrie matrix*, or *P-matrix*, if each column contains a sequence of 0s, followed by a sequence of 1s, followed again by a sequence of 0s. (In each column, it is possible for any of these sequences to have length zero.)

Thus a P-matrix is characterized by the absence of *embedded* 0s (that is, 0s that have 1s above and below them in the same column). Equivalently, all 1s in a column are consecutive.

In our application, saying that the incidence matrix is a P-matrix means that once a trait is present in the tale it may remain in the tale throughout its progression; but if the trait then disappears from the tale, it will not reappear in later renditions.

Definition. A matrix A is called *pre-P*, or pre-Petrie, if there is a row-permutation matrix Σ such that ΣA is Petrie. (From a permutation σ we obtain a permutation matrix Σ by setting $\Sigma(i, \sigma(i)) = 1$ for each row index i, and setting other entries equal to 0.)

The *consecutive ones problem* consists in rearranging the rows of an incidence matrix so it becomes a P-matrix — corresponding, in our case, to sorting the tales into a temporal order consistent with the changes in traits. If the matrix is pre-P the problem is solvable (by definition!) and an appropriate permutation matrix can be found quickly by any of several efficient algorithms. However, in applications, it is often the case that the incidence matrix is not pre-P, just "almost" so. In that case, the problem becomes more complex and a solution is not guaranteed to exist [Dundes 1965]; nonetheless one can look for a permutation that brings the incidence matrix into a form as close as possible to a P-matrix.

Dissimilarity. Our first approach is brute force: we test all possible permutations and choose one that gives a result closest to a P-matrix. Because of exponential growth, we are limited with this approach to small data sets; but at least we can avoid having to compare the rows of A itself each time. Instead, we introduce the *similarity matrix* $S = AA^T$, whose rows and columns both correspond to tales. Its name is due to the fact that the off-diagonal entries of S express how many 1s two rows of A have in common — that is, how many traits two tales share. (The diagonal entries in S show the number of traits possessed by each tale. It is easy to see that S is symmetric and nonnegative.)

For instance, taking again the data in Table 2, the similarity matrix is

$$S = AA^T = \begin{pmatrix} 4 & 2 & 2 & 0 & 0 & 0 & 1 & 1 \\ 2 & 5 & 0 & 1 & 1 & 1 & 3 & 3 \\ 2 & 0 & 4 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 4 & 3 & 3 & 2 & 2 \\ 0 & 1 & 1 & 3 & 4 & 4 & 1 & 1 \\ 0 & 1 & 1 & 3 & 4 & 4 & 1 & 1 \\ 1 & 3 & 0 & 2 & 1 & 1 & 4 & 4 \\ 1 & 3 & 0 & 2 & 1 & 1 & 4 & 5 \end{pmatrix}.$$

Even more convenient to use is the *dissimilarity matrix* D, defined by setting $D(i, j) = n - S(i, j)$ for $1 \leq i, j \leq m$. Here n is the number of columns of A (possible traits) and m the number of rows (tales). For our running example, with $m = 8$ and $n = 13$, the dissimilarity matrix is

$$D = \begin{pmatrix} 9 & 11 & 11 & 13 & 13 & 13 & 12 & 12 \\ 11 & 8 & 13 & 12 & 12 & 12 & 10 & 10 \\ 11 & 13 & 9 & 13 & 12 & 12 & 13 & 13 \\ 13 & 12 & 13 & 9 & 10 & 10 & 11 & 11 \\ 13 & 12 & 12 & 10 & 9 & 9 & 12 & 12 \\ 13 & 12 & 12 & 10 & 9 & 9 & 12 & 12 \\ 12 & 10 & 13 & 11 & 12 & 12 & 9 & 9 \\ 12 & 10 & 13 & 11 & 12 & 12 & 9 & 8 \end{pmatrix}.$$

Note that the entry $D(i, i+1)$ gives the number of changes (dissimilarities) from the i-th row to the next, so the quantity that concerns us is the sum $\sum_{i=1}^{m-1} D(i, i+1)$, which we call the *total dissimilarity of* A. For our running example this number is $11 + 13 + 13 + 10 + 9 + 12 + 9 = 77$.

Once we apply a permutation σ to the rows of A we are looking at the quantity

$$L(\sigma) := \sum_{i=1}^{m-1} D(\sigma(i), \sigma(i+1)).$$

We call $L(\sigma)$ the *total dissimilarity of A permuted by* σ. We can also view it as the sum of the $(i, i+1)$ entries of the conjugate matrix $\Sigma D \Sigma^{-1}$, where Σ is the permutation matrix. Indeed, under the action of σ, the incidence matrix A becomes ΣA, so S becomes $\Sigma A(\Sigma A)^T = \Sigma A A^T \Sigma^T = \Sigma S \Sigma^{-1}$, since for a permutation matrix, transposing is the same as inverting. Similarly, Σ turns D into $\Sigma D \Sigma^{-1}$.

Our goal now is to *minimize* the total dissimilarity over all permutations. When a dissimilarity-minimizing permutation is applied to A, the result will be as close to a P-matrix as possible, and the following criterion says whether or not it is in fact a P-matrix:

Theorem 1 [Shuchat 1984]. *Let A be an $m \times n$ incidence matrix, with similarity matrix $S = AA^T$ and dissimilarity matrix D. For any row permutation Σ, the total dissimilarity $L(\Sigma)$ satisfies*

$$L(\Sigma) \geq \mathrm{trace}(D) = mn - \mathrm{trace}(S).$$

Further, A is a pre-P matrix if and only if equality is attained for some Σ, in which case ΣA is a Petrie matrix.

This is the translation of [Shuchat 1984, Theorem 1] to our setup, which differs from Shuchat's in that his matrices include a dummy row.

For our running example one dissimilarity-minimizing permutation is $\sigma = 78213654$ (this means the entries $\Sigma(1,7)$, $\Sigma(2,8)$, ..., $\Sigma(8,4)$ equal 1). Upon its application the dissimilarity matrix becomes

$$\Sigma D \Sigma^{-1} = \begin{pmatrix} 8 & 9 & 12 & 12 & 11 & 13 & 10 & 12 \\ 9 & 9 & 12 & 12 & 11 & 13 & 10 & 12 \\ 12 & 12 & 9 & 9 & 10 & 12 & 12 & 13 \\ 12 & 12 & 9 & 9 & 10 & 12 & 12 & 13 \\ 11 & 11 & 10 & 10 & 9 & 13 & 12 & 13 \\ 13 & 13 & 12 & 12 & 13 & 9 & 13 & 11 \\ 10 & 10 & 12 & 12 & 12 & 13 & 8 & 11 \\ 12 & 12 & 13 & 13 & 13 & 11 & 11 & 9 \end{pmatrix},$$

and $L(\Sigma) = 72$ is the minimum total dissimilarity. Since this is strictly more than $mn - \mathrm{tr}(S) = 70$, we conclude that A is not a pre-P matrix.

The permutation 78213654 encodes a possible reconstruction of the evolution of *Star Husband Tale* based on dissimilarities between traits. It suggests this scenario:[1] The earliest tale is 7, from the Plains Sarsi tribe. From the Plains tribes the tale went to the Eskimo tribes, followed by California and Northern Pacific tribes. Note that even with just a few traits included in these tales, this method tends to appropriately group the Plains tribes together, holding true to their geographic location. One can see the location of the tribes in Figure 1.

We used Mathematica version 8 to generate all permutations that minimize the total dissimilarity; with just 8 tribes and 13 traits this took approximately 248 seconds of CPU time. (Mathematica 8 has a built-in function that finds one shortest tour in the traveling salesman problem; this could be applied to the dissimilarity matrix to find a minimizing permutation.)

[1] By construction, the reverse permutation, 45631287, is also minimizing, so the reverse order would be equally possible: Tale 7 the most recent, etc. Many other minimizing permutations exist.

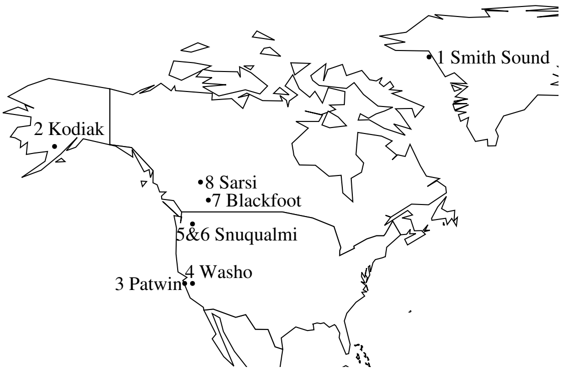

Figure 1. Locations of tribes from Table 2.

The consecutive ones problem. We return to the formulation given on page 4, where we mentioned that a P-matrix is one having no embedded 0s. Equivalently, for a P-matrix the distance between the first and last 1s in each column is 1 less than the number of 1s in that column (the distance is $k-1$ when there are k consecutive 1s). Equivalently, for a P-matrix the sum of these distances over all columns is simply the total number of 1s minus the number of columns. And obviously, for *any* binary matrix the sum must be at least as great as this difference.

Since the total number of 1s in A is the sum of diagonal elements of $S = AA^T$, we have proved the following:

Theorem 2 [Shuchat 1984]. *Let A be an $m \times n$ incidence matrix, and for each column j, let $r_j(A)$ be the difference between the row index of the last* 1 *in column j and that of the first* 1 *in the same column. Given a row-permutation matrix Σ, define the* 1*-content*

$$R(\Sigma) = \sum_{j=1}^{n} r_j(\Sigma A),$$

where of course $r_j(\Sigma A)$ is the corresponding difference for the permuted matrix ΣA. Then

$$R(\Sigma) \geq \operatorname{tr}(S) - n.$$

The matrix A is a pre-P matrix if and only if equality is attained for some Σ, in which case ΣA is a Petrie matrix.

For the permutation matrix A of page 3, we have $R(\text{identity}) = \sum_{j=1}^{n} r_j(A) = 35$. The minimum 1-content $R(\Sigma)$ using the 8 tribes and 13 tales from Table 2 turns out to be 26, again showing that the incidence matrix A is not a pre-P matrix. An example of a permutation that can be used to obtain this minimum 1-content is 54876132.

These give the chronological order of evolution of *Star Husband Tale* based on the number of embedded 0s within tales. For example, the minimum dissimilarity permutation 54876132 orders the progression of the tale starting with Tale 5, which came from the North Pacific Snuqualmi 1 tribe. Based on this analysis, of this small data set, the tale may have bounced between North Pacific tribes and Plains tribes before moving to the Eskimo and California tribes. Mathematica version 8 generated all permutations that minimize the number of embedded 0s for the matrix A in 61 seconds of CPU time. This is four times faster than with the dissimilarity method (page 6); but a similar computation for the full problem — 86 tribes with 135 traits — is impossible using the brute-force approach. In the next section, we discuss an algorithm that is computationally more practical for larger data sets. A comparison of all three algorithms is presented in Section 5.

4. A spectral algorithm for seriation

Atkins et al. [1999] gave an algorithm, based on eigenvalues and eigenvectors, for finding a permutation matrix Σ such that ΣA is a P-matrix. It assumes that the original matrix A is a pre-P matrix, but it degrades gracefully in the absence of that condition (that is, its results are not greatly affected if the matrix is almost a pre-P matrix.)

We start with some definitions.

Definition. A matrix $S \in \mathbb{R}^{m\times m}$ is *reducible* if there exists a permutation matrix Σ such that

$$\Sigma S \Sigma^{-1} = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix},$$

where $B \in \mathbb{R}^{r\times r}$, $D \in \mathbb{R}^{(m-r)\times(m-r)}$, and $0 < r < m$. If no such permutation exists, S is called *irreducible*.

Definition. Given an $m \times m$ symmetric matrix S and a diagonal matrix D such that $D(i,i) = \sum_{j=1}^{m} S(i,j)$ for $1 \le i \le m$, the *Laplacian* of S is $L = D - S$. It is easy to see that $e = (1, \dots, 1)$ is an eigenvector of L, with eigenvalue 0. The minimum eigenvalue of L with an eigenvector orthogonal to e is called the *Fiedler value*, and a corresponding eigenvector is a *Fiedler vector*.

Definition. A square matrix S is called a *Robinson matrix* [1951], or *R-matrix*, if

$$S(i,j) \le S(i,k) \text{ for } j < k < i \quad \text{and} \quad S(i,j) \ge S(i,k) \text{ for } i < j < k.$$

If there is a permutation matrix Σ such that $\Sigma S \Sigma^{-1}$ is an R-matrix, S is called a *pre-R matrix*.

Theorem 3 [Atkins et al. 1999]. *Any R-matrix has a monotone Fiedler vector.*

Theorem 4 [Atkins et al. 1999]. *Let S be a pre-R matrix with a simple Fiedler value and a Fiedler vector with no repeated values. Let Σ_1 and Σ_2, respectively be the permutations induced by sorting the values in the Fiedler vector in increasing and decreasing order. Then $\Sigma_1 S \Sigma_1^{-1}$ and $\Sigma_2 S \Sigma_2^{-1}$ are R-matrices and no other permutations of S produce R-matrices.*

For the similarity matrix S of our running example (page 5), is irreducible with simple Fiedler value approximately equal to 3.517 and Fiedler vector

$$(-1.674, 0.722, -3.800, 1.238, 0.757, 0.757, 1, 1).$$

Since the Fiedler vector is not monotonic, by Theorem 3, S is not an R-matrix. Under the assumption that S is a pre-R matrix, one can find the permutation that puts the Fiedler vector in increasing order; however Theorem 4 cannot be applied due to the occurrence of repeated entries in the Fiedler vector. The following two theorems prove helpful if the similarity matrix is reducible or has a Fiedler vector with repeated values.

Lemma 5 [Atkins et al. 1999]. *Let S_k be the irreducible blocks of a pre-R matrix A and let Σ_k be permutations that make these blocks become R-matrices. Then any permutation obtained by concatenating the Σ_k will make A become an R-matrix.*

Theorem 6 [Atkins et al. 1999]. *Let S be a pre-R matrix with a simple Fiedler value and Fiedler vector x. Suppose that there is some repeated value β in x and define I, J and K to be the indices for which*

- *$x_i < \beta$ for all $i \in I$,*
- *$x_i = \beta$ for all $i \in J$,*
- *$x_i > \beta$ for all $i \in K$.*

Then ΣS is an R-matrix if and only if Σ or its reversal can be expressed as $(\Sigma_i, \Sigma_j, \Sigma_k)$, where Σ_j is an R-matrix ordering for the submatrix $S(J, J)$ of S induced by J and Σ_i and Σ_k are the restrictions of some R-matrix ordering for S to I and K respectively.

Applying Theorem 6 to our running example, the spectral algorithm provides the permutation ordering 48765213 for the tales, under the assumption that S is a pre-P matrix. This algorithm is much less time-consuming than the seriation techniques in Section 3; it took 0.062 seconds of CPU time in Mathematica 8, and properly grouped Plains and North Pacific tribes together.

5. Seriation results: the Woodlands region

In order to show the reader the strength of seriation while still staying within current computing capacity, we chose to limit *the Star Husband* tribes to the Woodlands

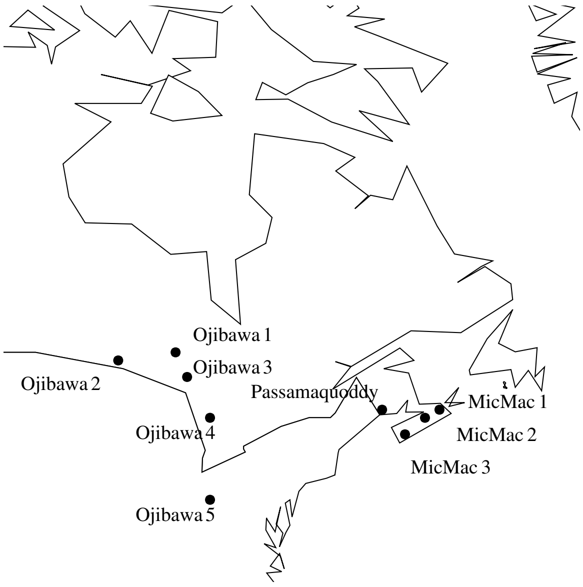

Figure 2. A map of the Woodland area tribe locations.

area, comprised of the Ojibwa, Micmac, and Passamaquoddy tribes, which contains 9 versions of the *Star Husband Tale* with 30 traits. A geographical map of the Woodlands area can be found in Figure 2.

In creating the incidence matrix for the folklore story, we used nine versions of the tale, numbered as follows:

1. Ojibwa 1
2. Ojibwa 2
3. Ojibwa 3
4. Ojibwa 4
5. Ojibwa 5
6. Micmac 1
7. Micmac 2
8. Micmac 3
9. Passamaquoddy

Note that if a trait within the nine tribes' tales was the same, the trait was eliminated from the incidence matrix all together. From the results, the minimum total dissimilarity L from all permutations is 192, while the maximum is 223. The minimum 1-content R is 86, while the maximum is 169. The range of total dissimilarities versus 1-contents can be visualized in Figure 3. Ideally, the best permutation would be the one with both the minimum L and minimum R. The permutations that satisfy this criterion are

$$687549312, \; 798651432, \; 798651423, \; 423561798, \; 312459687, \; 312459678.$$

In the majority of these permutations, the Ojibwa (tribes 1, 2, 3, and 4) and Mic Mac tribes (6, 7, and 8) are grouped together respectively based on their tales' characteristics. Also note that Ojibwa 4 and 5 as well as the Passamaquody tales

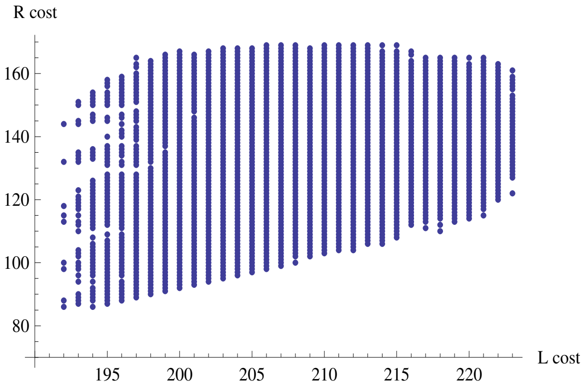

Figure 3. Cost functions length $L(\Sigma)$ versus 1-content $R(\Sigma)$ for all permutation matrices Σ.

are most often the transitional tales in the seriation. This is most likely due to the geographic central proximately of these tribes to the neighboring tribes. These are significant results as the evolution of the *Star Husband Tale* based strictly on the presence or absence of traits in the tales matches the geographic locality of these tribes as well.

As mentioned in the previous section, we introduce the spectral algorithm to find the evolution of the *Star Husband Tale* as an alternative to the seriation techniques that require the generation of all permutations of tribes. Note that both of the seriation techniques presented in Section 3 would need to make computations with $9!$ permutations as applied to the Woodlands tribe while the spectral algorithm looks strictly at the eigenvalues and eigenvectors of the similarity matrix.

Using the spectral algorithm technique to order the tales from the Woodlands area takes significantly less time then the traditional seriation techniques. This algorithm produces an ordering of

$$8, 7, 6, 9, 3, 5, 1, 4, 2,$$

grouping the Ojibwa tales together and Mic Mac with Passamaquody tales which corresponds to the geographically locations of these tribes as well. In addition, the ordering puts tale 3, Ojibwa 3, closest to the Passamaquody tale. Although geographically Ojibwa 4 is closer to the Passamaquody tribe, Ojibwa 3 comes in a close second.

6. Seriation results: eighty-six tribes

This spectral algorithm also does a reasonable job in ordering the entire eighty-six versions of *Star Husband Tale* as well, something the other seriation techniques can

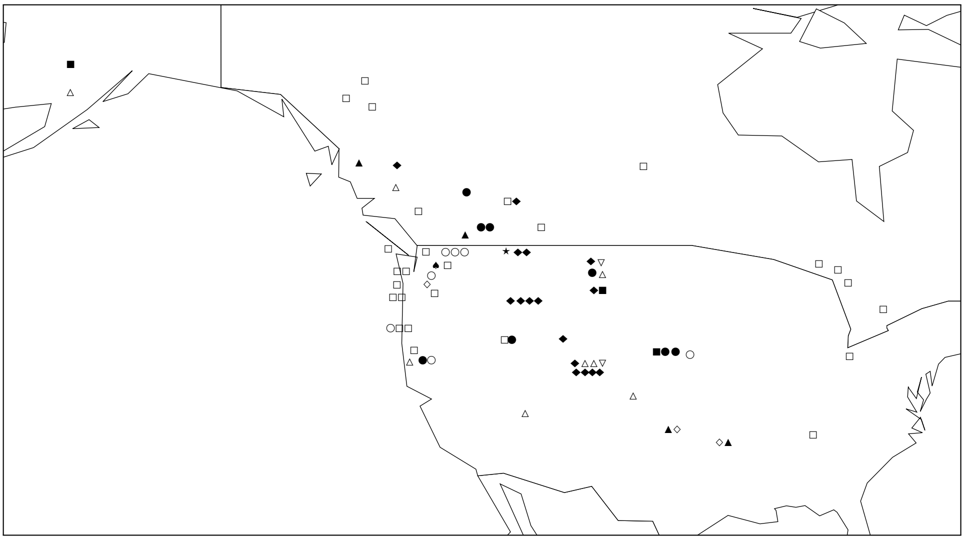

Figure 4. The geographic location and cluster based on spectral algorithm with imposed clustering. Order in which the tales represented by each symbol occurs in the seriation: □ = 1, ○ = 2, ♠ = 3, ◇ = 4, ★ = 5, ▲ = 6, ● = 7, ■ = 8, ◆ = 9, ▽ = 10, △ = 11.

not achieve computationally. The ordering produced with the eighty-six version data set reveals that *the Star Husband Tale* originated somewhere in the Plains region, possibly with the Cree tribe, and stayed in and around the Northwest border of the United States and Canada before spreading south and east to the California and Woodlands regions respectively. This ordering falls in line with Thompson's analysis [Dundes 1965], which claims that the tale did in fact originate in the Plains region.

Although the seriation techniques deal specifically with ordering and do not have a natural imposed clustering, *grouping*, if one did wish to cluster the seriated data note that tales within clusters should be similar. We impose a clustering by calculating the Hamming distance between adjacent tales in the spectral algorithm results. When the Hamming distance varies significantly a new cluster is created.

Figure 4 shows all eighty-six tale locations based on the spectral algorithm and imposed clustering. With both a seriation, ordering, and clustering, grouping, of the data one can analyze the progression of *the Star Husband Tale* between clusters. The significance of the results presented in Figure 4 is that the algorithm which produces a clustering based on characteristics of the tales also matches up geographically with the locations of the tribes.

The spectral algorithm with imposed clustering produces a first cluster of 31 tales. This cluster may be the most interesting to analyze. One can see two significant subclusters from this first cluster, one on the western coast and one along the northeastern border of the United States and Canada, and a large region in central Canada containing only one tribe, also contained in this cluster. It is highly possible that tribes from both subclusters shared common hunting grounds located in the plains region of Canada and thus producing similar versions of the tale.

We have presented here just a few algorithms for ordering *the Star Husband Tale*. With this particular data set the spectral algorithm was very successful; however all of the algorithms described assume that the data matrix is a pre-P matrix. Without this attribute, results could degrade quickly. One might consider applying clustering techniques to a similar data set. For this particular data set both agglomerative clustering and k-means clustering were explored but the results were much less attractive than those produced by the spectral algorithm with imposed clustering.

References

[Atkins et al. 1999] J. E. Atkins, E. G. Boman, and B. Hendrickson, "A spectral algorithm for seriation and the consecutive ones problem", *SIAM J. Comput.* **28**:1 (1999), 297–310. MR 99j:68049 Zbl 0930.05064

[Buneman 1971] P. Buneman, "The recovery of trees from measures of dissimilarity", pp. 387–395 in *Mathematics in the archaeological and historical sciences*, edited by F. R. Hodson et al., Edinburgh University Press, Edinburgh, 1971.

[Carroll 1979] M. P. Carroll, "A new look at Freud on myth: reanalyzing the star-husband tale", *Ethos* **7**:3 (1979), 189–205.

[Dundes 1965] A. Dundes (editor), *Introduction to Stith Thompson's "The star husband tale"*, pp. 414–415, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[Huson and Bryant 2006] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies", *Mol. Biol. Evol.* **23**:2 (2006), 254–267.

[Lyman et al. 1998] R. L. Lyman, S. Wolverton, and M. J. O'Brien, "Seriation, superposition, and interdigitation: a history of Americanist graphic depictions of culture change", *Amer. Antiquity* **63**:2 (1998), 239–261.

[O'Brien and Lyman 2002] M. J. O'Brien and R. L. Lyman, *Seriation, stratigraphy, and index fossils: the backbone of archaeological dating*, Kluwer Academic, New York, 2002.

[Petrie 1899] W. M. F. Petrie, "Sequences in prehistoric remains", *J. Anthropol. Inst.* **29**:3–4 (1899), 295–301.

[Rich 1971] G. W. Rich, "Rethinking the 'star-husbands'", *J. Amer. Folklore* **84**:334 (1971), 436–441.

[Robinson 1951] W. S. Robinson, "A method for chronologically ordering archaeological deposits", *Amer. Antiquity* **16**:4 (1951), 293–301.

[Shuchat 1984] A. Shuchat, "Matrix and network models in archaeology", *Math. Mag.* **57**:1 (1984), 3–14. MR 86a:00017 Zbl 0532.90097

[Young 1978] F. W. Young, "Folktales and social structure: a comparison of three analyses of the star-husband tale", *J. Amer. Folklore* **91**:360 (1978), 691–699.

Received: 2010-02-26 Revised: 2011-10-27 Accepted: 2011-12-03

ccoles@elon.edu *Department of Mathematics and Statistics, Elon University,*
Elon, NC 27244, United States

tlee@elon.edu *Department of Mathematics and Statistics, Elon University,*
Elon, NC 27244, United States

cborden2@elon.edu *Elon University, Elon, NC 27244, United States*

mathematical sciences publishers

msp

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US $120/year for the electronic version, and $165/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2014 vol. 7 no. 1

1944-4176(2014)7:1;1-8