SIRIUS: Dual, Symmetric, Interactive Dimension Reductions

Michelle Dowling, John Wenskovitch, *Student Member, IEEE*, J.T. Fry, Scotland Leman, Leanna House, and Chris North

Fig. 1. The initial, interactive symmetric dual projections of a multidimensional dataset using SIRIUS. Observations (animals) are projected in the left panel, while attributes (animal characteristics) are projected in the right panel. Both panels project similar items closer together based on a weighted high-dimensional distance function in which the weights reflect a conceptual notion of "importance." These weights are reflected by the node sizes and opacities in the opposing panel. For example, *Quadrupedal* has a higher weight in the left projection of animals, and *Tiger* has a slightly higher weight in the right projection of characteristics.

Abstract—Much research has been done regarding how to visualize and interact with observations and attributes of high-dimensional data for exploratory data analysis. From the analyst's perceptual and cognitive perspective, current visualization approaches typically treat the observations of the high-dimensional dataset very differently from the attributes. Often, the attributes are treated as inputs (e.g., sliders), and observations as outputs (e.g., projection plots), thus emphasizing investigation of the observations. However, there are many cases in which analysts wish to investigate both the observations *and* the attributes of the dataset, suggesting a symmetry between how analysts think about attributes and observations. To address this, we define SIRIUS (Symmetric Interactive Representations In a Unified System), a symmetric, dual projection technique to support exploratory data analysis of high-dimensional data. We provide an example implementation of SIRIUS and demonstrate how this symmetry affords additional insights.

Index Terms—Dimension reduction, semantic interaction, exploratory data analysis, observation projection, attribute projection

1 Introduction

Visualizing and interacting with high-dimensional data for exploratory data analysis is an open research area with many facets to explore. In this paper, we focus on visual analytics techniques for high-dimensional data exploration that use dimension reduction to project 2D scatterplots of the data. Many existing techniques and interactions therein focus on **observation-centric** tasks that reveal relationships between observations as defined by their attributes[1], such as clustering tasks [3]. For example, with the animal dataset used throughout this paper [29], analysts could investigate questions such as "Which attributes separate the *Tiger* and *Wolf* from the *Blue Whale* and *Dolphin*?" or "What other animals are similar to those animal groups?" Projection methods often define weight parameters on the attributes that enable analysts to assign different levels of **importance** to each attribute, thus enabling exploration of alternative observation projections [7, 43].

Likewise, there are complementary **attribute-centric** tasks that reveal relationships between attributes as defined by their observations, such as correlation tasks [3]. For example, a follow-up question might be "What other attributes are correlated with the attributes that separate these two groups?" This question is more difficult to answer with observation-centric projections. Thus, other kinds of visualizations are often used, such as linked distribution plots and dynamic queries [1, 31, 34, 46, 47] or correlation matrices [44], creating asymmetry in how observation-centric and attribute-centric tasks are supported.

A natural **symmetry** between observation-centric and attribute-centric tasks in high-dimensional data can be defined as equivalent tasks on the data table or data matrix and its transpose (which swaps the observations and attributes). For example, this symmetry often arises in visualizations for text analytics. With a vector space model matrix, documents can be projected in terms of their word usage [2, 9, 13, 14]. Alternatively, with the matrix transpose, words can be projected in terms of their usage in documents [6, 11, 37, 40].

We propose that this task symmetry between observations and attributes reflects a symmetry in the cognition of multidimensional data, and therefore is better supported by a symmetry between how the observations and attributes are visualized and interacted with. Such a symmetry would give analysts equal power to investigate both observations and attributes, using the same visual representations and interactions for both. Additionally, previous work has shown cognitive bias towards symmetric stimuli, as well as an association between asymmetry and disgust [16]. Based on this, we assert that asymmetric visualization and interaction design should increase cognitive load in comparison to symmetric designs as they require analysts to simultaneously interpret different approaches to observations and attributes.

To address this need for symmetry, we define a **symmetric dual projection** technique in which the projection of observations is defined by the attributes, and the projection of attributes is defined by the observations. We further define **interactions that connect** the observations

- *Michelle Dowling, John Wenskovitch, and Chris North are with the Virginia Tech Department of Computer Science. E-mail: dowlingm — jw87 — north@cs.vt.edu.*
- *J.T. Fry, Scotland Leman, and Leanna House are with the Virginia Tech Department of Statistics. E-mail: fryjt1 — leman — lhouse@vt.edu.*

[1]In this paper, we define *observations* to be the items in a dataset and the *attributes* as properties or dimensions of those items.

and attributes between the symmetric projections, resulting in a manipulation of the projection of observations influencing the projection of the attributes and vice versa. These interactions reflect a notion of a deep connection between the observations and the attributes that mirrors the analyst's notion of how observations and attributes are interconnected.

Specifically, our contributions in this work are:

1. Defining a technique for symmetric, interconnected projections of observations and attributes to directly address the lack of symmetry in current visual analytics techniques (detailed in Sect. 3).
2. Creating an implemented instantiation of SIRIUS using Weighted Multidimensional Scaling (WMDS) (described in Sect. 4).
3. Demonstrating how SIRIUS allows analysts to gain insight on both observation-centric and attribute-centric tasks (shown in Sect. 5).

2 Related Work

Here, we provide a brief survey of interactive visual analytics techniques for exploratory data analysis with high-dimensional data to highlight a lack of connection and symmetry between observations and attributes. We directly address this limitation via SIRIUS. In the following discussion, we refer to visualizations as simplistic if they do not incorporate many dimensions, and interactions as simplistic if they result in a trivial interpretation to change a mathematical model used to process or visualize the data.

2.1 Displaying and Interacting with Attributes

The attributes of high-dimensional data are visualized using a variety of techniques, ranging from simplistic (e.g., a raw data table or data matrix [7]) to more complex and informative (e.g., MDS projections [8, 50] or PCA (Principal Component Analysis) projections [6,20,33,50,58, 59]). In most cases, attribute visualizations implement a more simplistic technique like visual encodings such as color [8, 19, 40, 41, 45, 57, 58], size [2, 8, 40], or labels [2, 6, 23, 40]. Another method is to show the attribute values for observations along a one-dimensional line [19, 43]. Individual axes in parallel coordinates [19, 25, 54] create a similar visualization of attributes. As the complexity of the visualizations grow, we begin to see visualizations capable of conveying more information, such as scatterplots [49], histograms [44], heatmaps [7, 54], and polar coordinates [40, 55, 56]. Some visualizations implement a specific method for conveying information with this level of complexity, such as the arrows used by Brown [6], representing attributes as "magnetic" nodes that pull on observation nodes [35, 57], and the "Axis Rainbow" in AxiSketcher [28]. The most complex examples of visualizations for attributes include the aforementioned MDS and PCA projections.

Interactions with the attributes tend to also be more simplistic. For techniques that map an individual attribute to an axis, the axis can be enlarged, shrunk, or rotated to alter how the given attribute influences the visualization [19, 22, 43]. Another method for altering the visualization of the attributes is to change the color mapping [2]. To see data associated with a particular attribute, analysts can sometimes hover over or click on nodes [6, 8]. Brushing and linking can also be used to highlight attributes [49, 58]. Searching mechanisms allow new attributes to be added to the visualization [40], while sorting enable analysts to easily find specific attributes easily [19]. Some techniques also support clustering of attributes [40, 58]. With respect to attributes, there are few examples of more complex interactions, such as dragging the attribute nodes [35, 40, 57], the focus and context interactions described by Turkay et al. [49], the update features described by Brown [6], and altering the attribute values in the aster plot in AxiSketcher [28]. These types of interactions adjust the underlying visualization mechanisms to update the visualization itself based on the analyst's interaction.

2.2 Displaying and Interacting with Observations

Visualization techniques used to display the observations of high-dimensional data also range widely in complexity, but they tend to be more complex in comparison to those used to display the attributes. The least complex of these are raw data [7], color [8, 23, 44, 58], size [8], and lists [40]. Increasing in complexity, we again see visualizations like heatmaps [44] as well as frequency plots [25], dendrograms [44], and scatterplots [19, 21, 28, 49]. Many visualizations for observations attempt to incorporate all the attributes explicitly in visualizations [19, 22, 25, 35, 54, 57]. Examples of implicitly including all attributes can be seen in scatterplot-like projections of the data (e.g., MDS projections [8, 43], PCA projections [7, 25, 49, 58], t-SNE [23], and force-directed layouts [2, 52]).

Similarly, the interactions on the observations tend towards more complex interactions. Simplistic interactions include hovering or clicking on representations of observations to see the associated data [6–8, 22, 25, 35, 43, 52] and altering how color is mapped in the visualization [8, 41, 44]. Common, but still simple, interactions such as filtering [22, 35, 44, 54, 58], searching [6, 54], and brushing and linking [6,7,44,49,58] are often included. However, many visualizations for the observations enable direct manipulation of the visualization itself, such as how the attributes are used for the axes [19, 22, 44], manipulating the projection to alter an underlying mathematical model [7, 21, 43], selecting the clustering algorithm used or at what level clustering occurs [44, 58], manipulating how different attributes influence the visualization of the observations [2, 35, 43, 52, 57], drawing lines through the visualization to redefine the axes [28], and using a lens to separate groups of observations [23].

2.3 Projecting Attributes and Observations

Although at first glance it may appear that techniques such as the Data Context Map [8], Dimension Projection Matrix/Tree [58], and the visualization defined by Turkay et al. [49] provide symmetry in how observations and attributes are visualized and interacted with, there are important differences in their visualization methods, interaction methods, or both. In the Data Context Map [8], observations and attributes are plotted in the same MDS projection. Such a projection enables insights regarding similarity-based relationships between observations and attributes while contextualizing the projection of the observations. However, the tradeoff is that this technique necessarily distorts either the projection of the observations, the attributes, or both. This distortion is caused by the fact that each additional piece of data plotted in an MDS projection influences the projection of all other data. Therefore, the observations and/or attributes can appear more or less similar than what they actually are. Furthermore, the interactions for the Data Context Map focus on drawing contours around observations based on ranges of attribute values; there is no interaction to draw contours based on ranges of observations.

As for the Dimension Projection Matrix/Tree [58], both observations and attributes are visualized using PCA projections. However, the observation projection can be split into multiple projections based on specific subsets of attributes, whereas the attribute projection always remains a single projection. Additionally, the projections of the observations are given colored axes based on their corresponding subset of attributes; such information about the observations is not provided in the attribute projection. Thus, while this visualization and the interactions therein enable exploration of how attribute subspaces affect projections of the observations, the obvious tradeoff in this technique is that such exploration of observation subspaces is not supported.

Lastly, the visualization by Turkay et al. [49] provides three scatterplots: one for observations using one attribute for each axis, an observation projection using two principal components, and one visualizing attributes using their mean and standard deviation. Despite observations and attributes being displayed in separate scatterplots, the manner in which each portrays information is inherently different, meaning these projections do not have a strong symmetry. Additionally, interactions include brushing and linking between observations selected in the first scatrerplot and associated data in the other two scatterplots as well as focus and context interactions based on selected attributes in the attribute scatterplot. Thus, the interactions for observations are very different than those for attributes, creating further disparity between observations and attributes. Therefore, while this visualization technique enables deep exploration of the observations, the tradeoff is that such exploration for attributes is not well-supported.

Table 1. A description of the commonly used variables and functions in the equations throughout this paper.

O	A	The original data matrix (observations; O) and its transpose (attributes; $A = O^T$), with the columns for each matrix normalized
n	p	The number of observations n or attributes p, as represented by the number of rows in O or A, respectively
O_i	A_i	The high-dimensional data for the i^{th} observation (O_i) or i^{th} attribute (A_i), as represented by the i^{th} row of O or A, respectively
$\hat{O}$	$\hat{A}$	The dimensionally reduced matrices derived from O and A. In SIRIUS, the dimensionally reduced space is 2D, enabling easy projection onto a computer screen
W_O	W_A	The observation weights (W_O) or attribute weights (W_A), each represented as a single vector
W_{O_i}	W_{A_i}	The i^{th} observation weight (W_{O_i}) or attribute weight (W_{A_i})
$wDist_O$	$wDist_A$	A matrix of high-dimensional pairwise weighted distances between observations ($wDist_O$) or attributes ($wDist_A$)
$hDist_O$	$hDist_A$	The weighted high-dimensional distance function to calculate similarities between observations ($hDist_O$) or attributes ($hDist_A$). In SIRIUS, we use weighted Euclidean distance for each of these distance functions.
$lDist$		The low-dimensional distance function to calculate the pairwise distances between rows of a dimensionally reduced matrix. The same function is used for both observations and attributes, which further promotes symmetry in the presentation, interaction, and interpretation of both projections. In SIRIUS, we use 2D Euclidean distance.

Given the above discussion, attributes are generally treated very differently than observations, yet many tasks that analysts have regarding attributes are symmetrically similar to those regarding observations. Therefore, there is an opportunity to explore a new part of the design space of semantic interaction in visual analytics in which observations and attributes are displayed and interacted with in a symmetric and interconnected manner. Thus, we propose a new, symmetric exploratory data analysis technique for visualizing and interacting with both observations and attributes of high-dimensional data called SIRIUS, detailed in the next section. An example implementation is described in Sect. 4.

3 SIRIUS: A Technique for Symmetric, Interactive Projections of Observations and Attributes

To enable exploratory data analysis with high-dimensional data using symmetric visualization and interaction techniques between the attributes and observations, we designed a new technique called SIRIUS (Symmetric Interactive Representations In a Unified System). We assert that in SIRIUS the analyst must be able to:

- Goal 1: View similarity-based relationships between observations and similarity-based relationships between attributes of high-dimensional data.
- Goal 2: Explore different projections of the data by altering the importance of specific observations or attributes.
- Goal 3: Understand how importances of observations affect the importances of attributes and vice versa.

Each of these goals are further described in the following subsections. To clearly exemplify these concepts, we use a subset of the animal dataset from Lampert et al. [29] containing 13 observations and 13 attributes. These observations and attributes were selected to provide a clear and intuitive example (e.g., by excluding attributes like *Newworld*) while ensuring variance (e.g., by only including *Horse* and not *Zebra*). More complex examples are presented in Sect. 5. Variables and functions are defined in Table 1.

3.1 Goal 1: Visualize Similarity-Based Relationships

This first goal combines the tasks of seeing similarities between observations and seeing similarities between attributes. For example, in the animal dataset where the observations are animals, the *German Shepherd* is similar to the *Wolf* but not very similar to the *Elephant*. These similarities between the observations can be visually represented via a projection method. There are many different methods of doing so, including but not limited to PCA [20, 39, 53], t-SNE [51], and MDS [26, 27, 48]. We generalize the lower-dimension projection of the observations to be the output of the function $project_O$:

$$\hat{O} = project_O(wDist_O)$$

Similarly, an additional projection method should enable analysts to see similarities between attributes. For example, in the animal dataset, the attribute *Strength* is more similar to the attribute *Size* than to *Grazer*. We generalize the lower-dimension projection of the attributes to be the output of a function, $project_A$, as follows:

$$\hat{A} = project_A(wDist_A)$$

As noted in our evaluation of the Data Context Map [8] in Sect. 2.3, visualizing the observations and the attributes in a single projection necessarily distorts the projection of the similarities between observations, attributes, or both. Thus, the observations and attributes must be projected into separate spaces to maintain an accurate representation of their similarities. This means we need two similarity-based projections: one for observations ($project_O$) and one for attributes ($project_A$).

Furthermore, to reduce confusion between the two projections, we propose that $project_O$ and $project_A$ should produce projections that are understood in the same manner by the analyst. This is best reflected by a symmetry in the manner in which observations and attributes are visualized (and later interacted with). The easiest way to accomplish this is to use the same projection method for both $project_O$ and $project_A$ (e.g., MDS), but they can differ if the analyst perceives them in the same manner (e.g., MDS and PCA).

3.2 Goal 2: Explore Different Projections

Exploring different projections stems from the need to gain new insights based on domain knowledge or a hypothesis, or for general exploratory analysis. These insights can be gained by redefining the similarity between observations or attributes, which implies interaction that alters at least one projection. Such interaction can be accomplished by either altering the parameters that generate the high-dimensional pairwise distances or by directly manipulating the projection.

3.2.1 Explore Projections of Observations

To exemplify what is meant by each of these interaction methods, we first focus on $project_O$. From a projection of the observations, the analyst may want to understand how placing more importance on different attributes affects the similarities between the observations. For example, the analyst may want to investigate how animals differ based on their *Water* attribute. By placing more importance on this one attribute, animals like *Otter* should be reprojected much closer to the *Dolphin* and *Blue Whale* and farther away from the *Siamese Cat*.

Alternatively, the analyst may want to see how altering similarities between observations influences the level of importance that should be placed on each attribute. For example, if the analyst drags nodes for *Dolphin* and *Blue Whale* in one corner of the animal projection (to denote their desired similarity) and *Elephant* to the opposite corner (to denote its desired difference from the other two animals), the technique should reflect a higher level of importance for attributes that describe the differences between these two groups. In this case, the *Walks* and *Grazer* attributes describe the differences between these two groups, implying they should be given higher levels of importance.

In both of these types of interactions within the observation projection, the importances of attributes are altered. This reflects the fact that

observations are understood based on their attributes. Given that the importance of attributes can be represented by weights on the attributes, the similarity between two observations, O_i and O_j, can be generalized with the following weighted high-dimensional distance function:

$$wDist_{O_{i,j}} = hDist_O(W_A, O_i, O_j)$$

There are many weighted distance functions that could be used here, such as weighted variants of Euclidean distance, Manhattan distance, cosine distance, Gower distance [18], Pearson coefficient [38], and Bray-Curtis dissimilarity metric [5, 6]. Which distance function to use is often determined by the tasks supported and data used, however it must be compatible with the chosen projection method and desired outcome of the projection itself. For example, while PCA can be used to accomplish Goal 1, it emphasizes variance rather than strictly pairwise distances. Thus, using a weighted Euclidean distance with PCA may not produce the desired results.

3.2.2 Explore Projections of Attributes

To maintain the desired symmetry with the observations, the same interactions are enabled in the projection of the attributes. Thus, the analyst can alter the importance of a specific observation to understand how this affects the similarities between attributes. For example, increasing the importance for *Cow* should result in attributes like *Walks*, *Size*, and *Strength* being placed closer together but far away from *Stripes*. Additionally, the analyst can alter the similarities between attributes to understand how the importances of observations are affected. For instance, if the analyst drags nodes for *Grazer* and *Size* to one corner of the projection and *Water* to another, *Horse* describes the differences between these two groups of attributes and should therefore receive a higher weight to denote its increased importance.

In both of these types of interactions within the attribute projection, the importances of observations are altered. This reflects the fact that attributes are understood based on the observations. Since the importance of observations can be represented by weights on the observations, the similarity between two attributes, A_i and A_j, can be generalized with the following weighted high-dimensional distance function:

$$wDist_{A_{i,j}} = hDist_A(W_O, A_i, A_j)$$

Note the symmetry between this equation and the equation for weighted high-dimensional distances between observations.

3.3 Goal 3: Relate Importances to Each Other

As stated previously, the analyst understands the observations based on their attributes and vice versa. This hints at a connectedness between the observations and the attributes themselves. In the equations in the previous subsection, this connectedness is initiated by the importance of the attributes affecting the similarity of the observations and the importance of the observations affecting the similarity of the attributes. However, the notion of interconnectedness goes beyond these equations: attributes that are given more importance indicate which observations should be given more importance and vice versa.

As a common example of this, a keyword search for relevant documents results in those keywords (i.e., attributes of documents) being given a high level of importance. This means that documents that are better described by those keywords are, in turn, more important as well, and hence should be returned in the query results. This example implies that observations that have higher values (e.g., keyword frequencies) for an attribute are more important, and vice versa.

Using the animal dataset again to exemplify this, increasing the importance of the *Water* attribute should result in animals like *Dolphin*, *Blue Whale*, and *Otter* also being given a high level of importance in addition to reprojecting their nodes closer together. *Siamese Cat*, on the other hand, should have a low level of importance as it is not well described by the *Water* attribute. However, these updated importances for the different animals also denote which attributes should be considered important, beyond the single *Water* attribute that was interacted with. This means that given important animals like *Dolphin*, *Blue Whale*, and *Otter*, attributes like *Speed*, *Active*, and *Smart* are also more important than other attributes that do not describe these animals as well.

Therefore, the importance of attributes should affect the importance of observations and vice versa to reflect the analyst's notion of the interconnectedness between the attributes and the observations. This can be accomplished with the following equations, where Imp_O and Imp_A compute the importance for one observation or attribute, respectively:

$$W_{O_i} = Imp_O(O_i, W_A)$$
$$W_{A_i} = Imp_A(A_i, W_O)$$

This interconnectedness between the importances of the observations and the importances of the attributes enables a new visual analytics technique that not treats the observations and the attributes in a symmetric manner while enabling rich interaction between both projections.

Given this interconnectedness in SIRIUS, the weights that reflect the importances of observations and attributes are crucial components of the technique. They are used to project the data via weighted high-dimensional distance functions to explore different projections and to relate the importances of observations and attributes to each other. We demonstrate how this can be accomplished in Sect. 4.

4 An Implementation of SIRIUS

The generalized SIRIUS technique above supports a design space of possible projection methods, distance functions, and interaction methods that can be inserted. We present a particular implementation of SIRIUS that addresses the goals of Sect. 3 in the following manner:

1. Using WMDS for both $project_O$ and $project_A$, and weighted Euclidean distance for $hDist_O$ and $hDist_A$, to create both a projection of the observations and a projection of the attributes.
2. Using the notions of parametric interactions (which we call PaI) and observation-level interactions described by Self et al. [43] to manipulate the weights W_O and W_A of the weighted Euclidean distance function. Since we use the concept of observation-level interactions in both the observation projection and the attribute projection, we have renamed observation-level interactions as projection interactions (PrI) to reduce any potential confusion[2].
3. Relating the importances of observations and attributes to each other by defining Imp_O and Imp_A as a dot product between the original data matrix or its transpose and a set of attribute weights or observations weights (respectively). Ultimately, these equations for importance result in interconnecting both projections by using an interaction in one projection to alter both projections.

We made these particular design choices based on previous research in visualizing and interacting with high-dimensional data [4, 13, 15, 43, 52]. However, these are not strict constraints; any distance function, projection method, or interaction method that properly address the goals defined in Sect. 3 may be used in place of our choices here.

The following subsections describe how we accomplished each goal in detail, using the same subset of the animal dataset from the previous section to exemplify each concept.

4.1 Goal 1: Visualize Similarity-Based Relationships

4.1.1 Observation Projection

In SIRIUS, the left projection is designed to depict similarities between observations. To visualize these similarities, we use weighted Euclidean distance in WMDS [15], as defined by the following equation:

$$\hat{O} = \underset{\hat{O}_1,\ldots,\hat{O}_n}{\arg\min} \sum_{i=1}^{n-1} \sum_{j>i}^{n} \left(lDist(\hat{O}_i, \hat{O}_j) - hDist_O(W_A, O_i, O_j) \right)^2 \quad (1)$$

Before the data can be projected, some preprocessing must occur. To overcome any potential distortions in the projection caused by attribute values on different scales, O is z-score normalized prior to visualization.

[2] Briefly, PaI refers to direct alteration of a model parameter via some control mechanism, such as a slider or textbox. An analyst may update that parameter with a precise value. In contrast, PrI *learns* a set of model parameters based on *an interpretation* of analyst alteration of the projection itself.

Fig. 2. An initial projection of a subset of the animal dataset using SIRIUS, which maps "importance" to node size and opacity to provide a deeper semantic connection between observations and attributes. This allows analysts to determine at a glance which animals best describe the attribute projection (from the observation panel) and which attributes best describe the animal projection (from the attribute panel).

Additionally, a set of attribute weights, W_A, that reflect the importances of each of the attributes must be defined before the high-dimensional distance matrix can be calculated. W_A has two constraints: $0 \leq W_{A_i} \leq 1$ and $\sum_i W_{A_i} = 1$. Thus, weights are interpreted as proportions of the analyst's interest in each attribute (i.e., its level of importance). Although this is a minor point in the initialization of SIRIUS, it greatly affects the interaction methods, as discussed in the following subsections. For now, we will say that W_A is initialized by determining the importances of the attributes, with details discussed in Sect. 4.3. The initial observation projection is shown in the left panel in Fig. 2, which accurately shows that *German Shepherd* and *Wolf* are more similar to each other than to the *Elephant* as desired.

4.1.2 Attribute Projection

Given the desire for symmetry between the observations and the attributes, a second projection is used to depict the similarities between the attributes. It also uses weighted Euclidean distance in WMDS, as defined in the following equation:

$$\hat{A} = \underset{\hat{A}_1,\ldots,\hat{A}_p}{\arg\min} \sum_{i=1}^{p-1} \sum_{j>i}^{p} \left(lDist(\hat{A}_i,\hat{A}_j) - hDist_A(W_O,A_i,A_j)\right)^2 \quad (2)$$

Again, the data must be preprocessed before it can be projected. This includes z-score normalizing A, as well as initializing an observation weight vector, W_O. These weights reflect the importances of each observation and must hold to the same two constraints as W_A. We again leave detailed discussion for how we initialize this set of weights for Sect. 4.3. This initial projection is shown in the right panel of Fig. 2, which demonstrates that *Strength* and *Size* are more similar to each other than to *Grazer*, as desired.

4.2 Goal 2: Explore Different Projections

In SIRIUS, we use weighted Euclidean distance to define both the similarities between observations and the similarities between attributes. Thus, exploration of different projections is accomplished by manipulating the associated weights, thereby allowing the use of PaI and PrI as described by Self et al. [43] to enable rich interactions.

4.2.1 Parametric Interaction (PaI)

To explore how attribute importances affect similarities between observations, PaI is enabled via an "Importance" slider, which is accessible by clicking a node in the attribute projection. The analyst can alter the attribute weight (i.e., importance) by manipulating the slider. During this interaction, all attribute weights are re-normalized to adhere to the previously described sum-to-1 and 0-to-1 constraints. The change in attribute weights is reflected in updates to the size and opacity of the attribute nodes, which visually reflects their importance. All observations are then reprojected using Equation 1 with the updated attribute weights to show the effect of the analyst's interaction. An example of PaI on the *Water* attribute is depicted in Fig. 3-A, which pulls *Otter* closer to the *Dolphin* and *Blue Whale* than to the *Siamese Cat*.

Symmetrically, PaI is also used to explore how observation importance affects similarities between attributes via the same "Importance" slider. The analyst can click an observation and adjust the slider to alter the weight (i.e., importance) for the given observation, and all observation weights are re-normalized. The size and opacity of the observation nodes are updated to reflect their new weights. All attributes are then reprojected using the updated observation weights in Equation 2. An example of PaI on the *Cow* observation is shown in Fig. 3-B, which correctly results in the *Walks*, *Size*, and *Strength* attributes being placed close together but far apart from the *Stripes* attribute to reflect their similarity in "cow-ness."

4.2.2 Projection Interaction (PrI)

To explore how observation similarities affect attribute importances, analysts can use PrI in the observation projection. This is accomplished by directly manipulating the observation projection via clicking and dragging observation nodes of interest to redefine their relative similarities. Once the analyst is done manipulating the projection, an "Update Layout" button above the observation panel is clicked. This triggers a semi-supervised re-learning of the attribute weights using *only* the observation nodes the analyst interacted with, $\hat{O}^*$, in the following optimization, essentially inverting the WMDS process in Equation 1:

$$W_A = \underset{W_{A_1},\ldots,W_{A_p}}{\arg\min} \sum_{i\in\hat{O}^*} \sum_{j\in\hat{O}^*} \left(lDist(\hat{O}_i^*,\hat{O}_j^*) - hDist_O(W_A,O_i,O_j)\right)^2 \quad (3)$$

This optimization must also adhere to the sum to 1 and 0 to 1 constraints for W_A. From the new attribute weights, the attribute node sizes and opacities are updated to reflect these new levels of importance, and Equation 1 is re-executed to reproject all observations. For example, as depicted in Fig. 3-C, dragging the nodes for *Dolphin* and *Blue Whale* to one corner of the projection and *Elephant* to the opposite corner, results in an increase in the importances of the *Walks* and *Grazer* attributes, which distinguish these two groups of animals.

Symmetrically, the projection of the attributes permits exploring how attribute similarities affect observation importances. This is accomplished via PrI by dragging attribute nodes and clicking the "Update Layout" button above the attribute panel. This triggers a very similar algorithm that effectively inverts the WMDS process in Equation 2 using *only* the attribute nodes the analyst interacted with, $\hat{A}^*$ and following the same 0 to 1 and sum to 1 constraints for W_O:

$$W_O = \underset{W_{O_1},\ldots,W_{O_n}}{\arg\min} \sum_{i\in\hat{A}^*} \sum_{j\in\hat{A}^*} \left(lDist(\hat{A}_i^*,\hat{A}_j^*) - hDist_A(W_O,A_i,A_j)\right)^2 \quad (4)$$

Using the new observation weights, the observation node sizes and opacities are updated to reflect these new levels of importance, and Equation 2 is re-executed to reproject all attributes. To demonstrate, Fig. 3-D shows that dragging the nodes for *Grazer* and *Size* far away from *Water* results in an increase in the importance for *Horse*.

4.3 Goal 3: Relate Importances to Each Other

As SIRIUS has been described thus far, it is somewhat similar to Andromeda [43]. The main difference is that instead of listing the attributes, an attribute projection is provided alongside the an observation projection. However, we now introduce two equations to interconnect observation importances and attribute importances: $W_{O_i} = O_i \bullet W_A$ and $W_{A_i} = A_i \bullet W_O$. These equations can be more generally expressed as:

$$W_O = O \bullet W_A \quad (5)$$

$$W_A = A \bullet W_O \quad (6)$$

Both of these equations are used in initializing the projections as well as both PaI and PrI, as shown in Fig. 4, thereby interconnecting the projections of the observations and attributes together. This interconnectedness between observation importances and attribute importances has crucial implications in revealing new relationships and affording additional insights by providing methods to alter node size, opacity, and position for both observations and attributes after any interaction. Thus, analysts are afforded insights such as correlations between observations or between attributes at a glance during any point of analysis.

Our use of these equations is loosely based on simple approaches to relevance computations in information retrieval and recommender

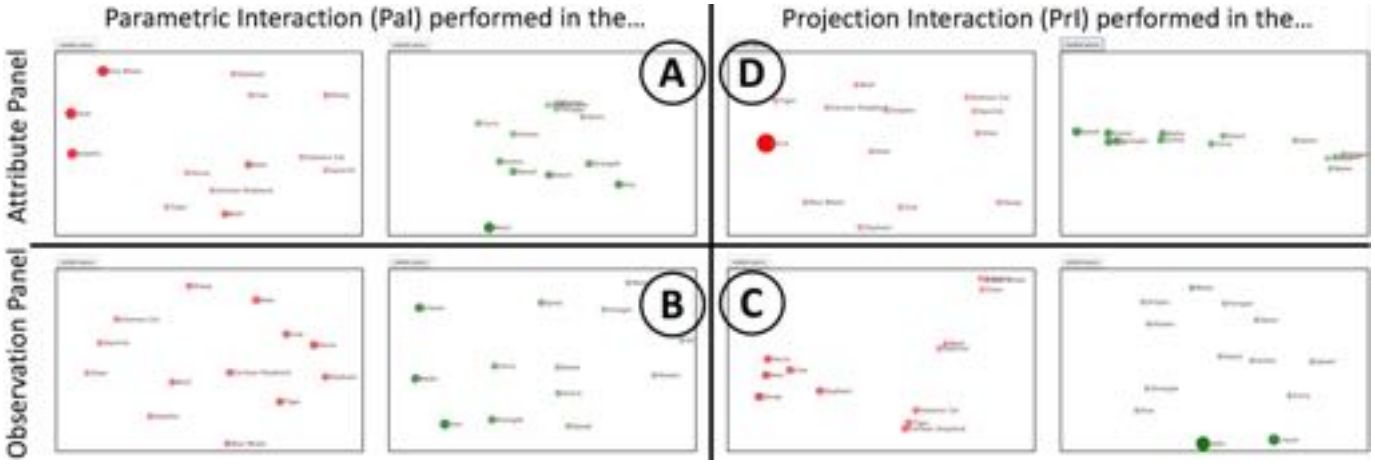

Fig. 3. The results of two examples of PaI and two examples of PrI described in Sect. 4 with Fig. 2 as the initial projection of the data and continuing to map "importance" to node size and opacity: **A** PaI performed on the *Water* attribute; **B** PaI performed on the *Cow* observation; **C** PrI performed by dragging the *Dolphin* and *Blue Whale* observations into one corner and the *Elephant* observation into the opposite corner; and **D** PrI performed by dragging the *Grazer* and *Size* attributes into one corner and the *Water* attribute into the opposite corner.

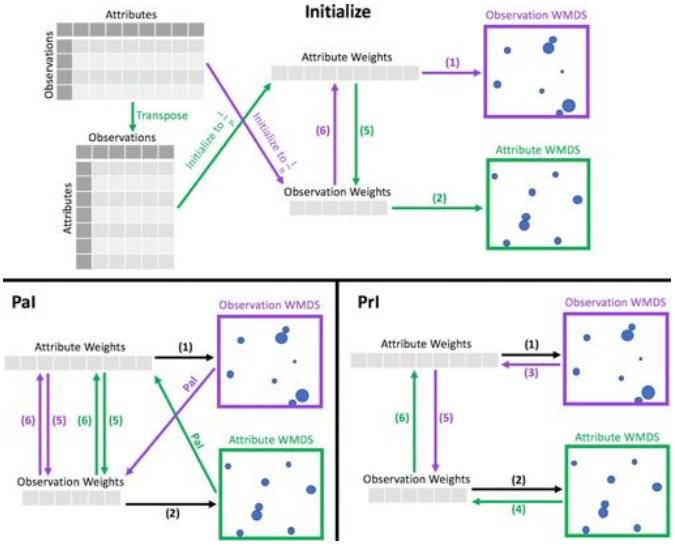

Fig. 4. A flowchart depicting how Equation 5 and Equation 6 are used in conjunction with Equations 1–4 on initialization or when PaI or PrI occur. Arrows and their associated equation numbers are colored based on whether they are used for the observation panel (purple), attribute panel (green), or both (black). Note that Equation 5 and Equation 6 are both used in PaI, whereas only one of these equations is used in PrI.

systems, such as the HITS (Hubs and Authorities on the Internet) Algorithm [24], which underlies Google's PageRank query technique [36]. However, a major difference is that HITS iterates over these two equations until convergence, whereas our implementation of SIRIUS only iterates once to enable explorations of alternative projections via PaI and PrI. However, it might be interesting to iterate until convergence during initialization.

Interconnecting the observation importances and the attribute importances is first seen when initializing each of the projections. For the projection of the observations, an observation weight vector is first initialized so that each observation has a weight of $1/n$. This reflects an equal level of importance for each observation while maintaining the sum-to-1 and 0-to-1 constraints for the weight vector. Then, Equation 6 is used to determine the attribute importances. However, this equation is not constrained as the attribute weights are. Therefore, to use these attribute importance values for the attribute weights in Equation 1, they are normalized to sum to 1. Similarly, the projection of the attributes is initialized by first generating a set of attribute weights in which each weight is $1/p$. Then, Equation 5 is used and normalized to sum to 1 to generate the observation weights for Equation 2.

After projecting the data using Equation 1 and Equation 2, as depicted in Fig. 2, the node sizes and opacities can be interpreted as visualizations of (left) which observations are most important or best describe the differences between the attributes and (right) which attributes are most important or best describe the differences between the observations. For this initial projection, the manner in which the importances of the observations and attributes are determined result in emphasizing items that are most "popular" (i.e., have the highest sum across the dataset, as explained in Sect. 5.1.1). Additionally, these projections show similarities (i.e., correlations) between attributes or similarities between observations.

These equations for importances are also used during each interaction. For example, when the analyst performs PaI on an attribute, a new set of observation importances is calculated using Equation 5. These importance values are used to update observation node size and opacity. Then, the observation importances are used to calculate a new set of attribute importances using Equation 6. This results in an update to attribute node size and opacity. Additionally, both observations and attributes are reprojected via Equation 1 and Equation 2 (respectively) after normalizing both new sets of importances to sum to 1. Thus, PaI results in node size, opacity, and position being updated in both projections. This effect is demonstrated in Fig. 3-A and Fig. 3-B.

Similarly, PrI in the observation panel produces a new set of attribute weights. These weights are used in Equation 1 to reproject the observations and in Equation 5 to determine new sizes and opacities for the observation nodes. Then, the new observation importances are used in Equation 6 to update the attribute node size and opacity. To update the attribute node positions, the new observation importances are normalized to sum to 1 and used in Equation 2. Thus, PrI also results in node size, opacity, and position being updated in both projections. This effect is depicted in Fig. 3-C and Fig. 3-D.

In the above use of dot products to relate observation importances to attribute importances and vice versa, there is an implied assumption that higher data values represent more importance. For example, we assume that a higher importance for the *Water* attribute indicates that animals that have a high value for the *Water* attribute are more important than animals with a low value. While this assumption is appropriate for some applications, such as in text analytics where values represent word occurrences, it may be less appropriate in other applications, such as wanting to emphasize both extrema. However, the technique for relating observation and attribute importances as described in Sect. 3 is

purposefully generic to support a variety of mathematical definitions for relating importances, including one which might generate higher importance values for animals with extreme low *Water* values as well as those with extreme high *Water* values.

When interpreting the dual projections, it is important to understand that, while the node sizes and opacities in one projection describe the spatial layout of the other projection, the projections themselves do not map onto each other. That is, since the projections represent separate high-dimensional spaces, the spatial positions of nodes in one projection do not specifically relate to the node positions in the other. Attempting to align all spatial positions would create a projection similar to the Data Context Map [8], which necessarily distorts the similarities between the observations, attributes, or both, as discussed in Sect. 2.3. Although one of the projections can be rotated and reflected to better match the layout in the other projection (e.g., rotate the attribute panel so that the *Water* and *Grazer* attributes are roughly in the same positions as the animals that are higher in those attributes), the potential tradeoff is that this may lead analysts to conclude that the two projections *can* be mapped onto each other. An inaccurate conclusion such as this can lead to significant misinterpretations of how the projections relate to each other. For these reasons, we do not attempt any such alignment between the two projections.

5 Examples of Data Analysis with SIRIUS

To demonstrate how Equations 1–6 coalesce to enable exploratory data analysis with diverse high-dimensional datasets, we provide examples of animal data, intelligence analysis data, and breast cancer data in our implementation of SIRIUS. These examples show that SIRIUS can be used to explore quantitative data, textual data, and large datasets, respectively. A demonstration video for each of these examples is available at `https://youtu.be/TzBjImkrbDU`.

Since Equations 1–6 rely on strictly numerical data, some of the example datasets had to be altered to change categorical attributes to numerical representations. Additionally, any rows that contained missing data were removed. Such issues could be better addressed through the use of alternative distance functions, such as Gower distance [18].

As we move away from more intuitive datasets like the one by Lampert et al. [29], it is important to note that raw data for a selected observation or attribute is displayed alongside the "Importance" slider. For numerical datasets, this raw data is expressed as a simple list of key-value pairs. For text datasets, selecting a document instead displays the associated raw text. Thus, analysts can readily explore the entire dataset without having to reference spreadsheets or outside tools to interpret the projections.

5.1 An Animal Dataset

5.1.1 The Full Dataset

Given the initial projection of the entire animal dataset by Lampert et al. [29] depicted in Fig. 1, we can already begin gaining insights about the dataset. For example, while there are no strongly distinguished groups or clusters of animals, more water-dwelling animals appear in the upper right whereas more land-dwelling animals appear in the lower left. Despite the hypothesis that the differences between animals are best described by their *Water* attribute, the size and opacity of the attribute nodes indicate that *Quadrupedal* is the correct answer. Since this is the initial projection of the attributes, this also means more animals have a high value for *Quadrupedal* than for any other attribute. Therefore, many animals in the dataset are *Quadrupedal* and that this attribute is the most "popular" attribute in the dataset. In addition to these insights, the projection of the attributes shows that there are strong correlations between certain attributes, such as *Quadrupedal* and *Furry*, since they are projected closely together. Thus, an animal that has a high value for *Quadrupedal* is likely to also have a high value for *Furry*. Similarly, *Grazer* and *Hooves* are somewhat correlated, but since they are projected on the opposite side of the attribute panel, they are not correlated with *Quadrupedal* and *Furry*.

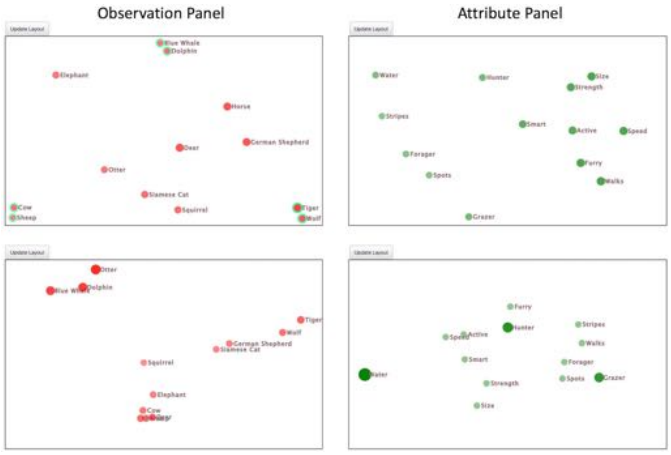

Fig. 5. Given the initial projection shown in Fig. 2, **(Top)** the analyst can move animals to express their desired similarities or differences to begin investigating their three questions about this animal dataset. **(Bottom)** After clicking "Update Layout," the data is reprojected with new attribute weights and observation weights. The analyst can now use node position, size, and opacity to determine the answers to all three questions without performing any further interactions.

5.1.2 Analysis on a Subset of the Dataset

Using the same subset of the animal dataset as in Fig. 2 and Fig. 3 for clarity, consider an analyst who wants to gain insights based on the three related questions mentioned in Section 1 using this dataset:

1. What attributes separate the *Tiger* and *Wolf* from the *Blue Whale* and *Dolphin* as well as from the *Cow* and *Sheep*?
2. What other animals are similar to those three groups?
3. What other attributes are correlated with the attributes that separate these three groups?

To answer these questions, after the data is initially projected (as shown in Fig. 2), the analyst would begin by using PrI to move the nodes for the animals of interest into three groups, as depicted in the top row of Fig. 5. Clicking "Update Layout" results in the final projection shown in the bottom row of Fig. 5.

Following this reprojection, the analyst can answer Question 1 by observing that the attributes *Water*, *Hunter*, and *Grazer* are large and opaque, thus leading to the insight that they describe the differences between the three groups of animals. Additionally, this projection gives the analyst the insight that *Horse*, *Deer*, *Elephant*, and *Squirrel* are similar to the *Cow* and *Sheep*; the *Siamese Cat* and *German Shepherd* are most like the *Tiger* and *Wolf*; and the *Otter* is similar to the *Dolphin* and *Blue Whale*. Thus, Question 2 is also answered from this projection.

However, Question 1 and Question 2 can be answered by existing techniques such as Andromeda [43]. What makes SIRIUS unique is that Question 3 can also be determined at a glance via the relative node positions in the attribute panel; more similar (or correlated) items will appear closer together in the projections. Thus, analysts can easily gain the insight that *Furry*, *Active*, *Speed*, *Smart*, *Strength*, and *Size* are all more correlated with *Hunter* than with *Grazer* or *Water*. Similarly, *Stripes*, *Walks*, *Forager*, and *Spots* are most correlated with *Grazer*. The *Water* attribute, in comparison to *Hunter* and *Grazer*, is not correlated with any other attributes.

Connecting these answers for Question 3 back to the animals, this means that the animals that have a high value in *Hunter* are more likely to have higher values for *Furry*, *Active*, *Speed*, *Smart*, *Strength*, and *Size* than animals that have a high value in *Grazer* or *Water*. Inspection of the animals in the observation panel (or domain knowledge) provides the insight that animals that are high in *Hunter* are animals like *Tiger* and *Wolf*. Similarly, animals that have a high value in *Grazer* (like the *Cow* and *Sheep*) are more likely to have higher values for *Stripes*, *Walks*, *Forager*, and *Spots* than animals that have a high value in *Hunter*

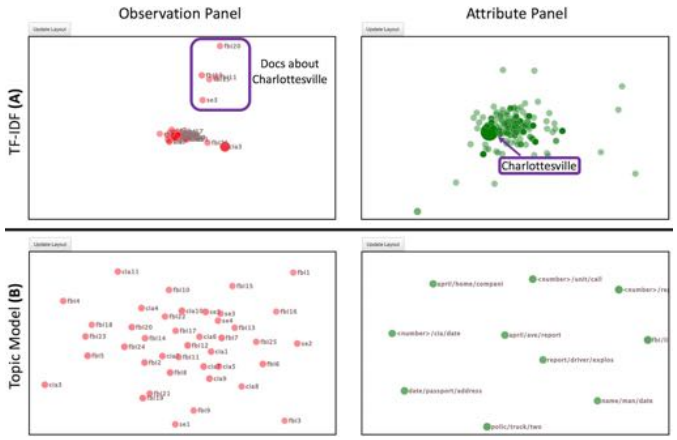

Fig. 6. The panels labeled **A** show an initial projection in SIRIUS with all extracted entities as attributes of a textual dataset, which immediately emphasizes *Charlottesville* as an important entity. The panels labeled **B** show an initial projection with topics learned through topic modeling as the attributes of the dataset. While this makes both the projection of the observations and the projections of the attributes clearer, the initial insight about *Charlottesville* is lost.

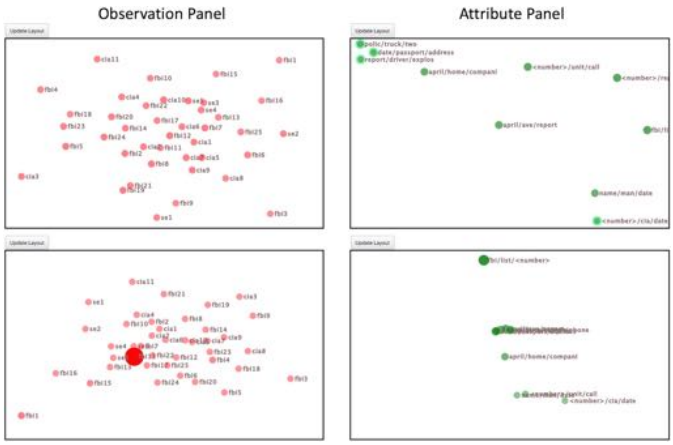

Fig. 7. From the initial projection of the topic modeled data shown in Fig. 6-B, nefarious activity can be uncovered by **(top)** using PrI on the attributes to separate topics of interest from generic or uninteresting topics. Clicking "Update Layout" produces **(bottom)** a visualization which reveals other topics that are very closely correlated with topics of interest. Additionally, the combination of emphasized attributes results in *fbi11* in the observation panel being highly emphasized. This document reveals crucial information to one of the three main terrorist plots in this dataset.

or *Water*. Animals that have a high value in *Water* (like the *Dolphin* and *Blue Whale*) are not as likely as animals that have a high value in *Hunter* or *Grazer* to have higher values in any of the other attributes.

5.2 A Text-Based Dataset

As mentioned in Sect. 4.3, our notion of importance, defined by Equations 5–6, results in a high importance for observations that have high values for important attributes and vice versa. This definition of importance is well-suited for text-based datasets in which a higher value for an attribute (e.g., an extracted entity) denotes that the associated extracted entity appears more often in that document (observation).

5.2.1 TF-IDF Data vs Topic Modeled Data

To demonstrate how SIRIUS can be used to explore textual data, we extracted entities from a synthetic intelligence analysis dataset and created a TF-IDF matrix in data preprocessing steps. Using SIRIUS to visualize this data (depicted in Fig. 6-A), we can immediately see that *Charlottesville* is greatly emphasized over other attributes. Given this is the initial projection, the emphasis on *Charlottesville* indicates that this entity has the highest sum of TF-IDF values across the entire dataset, hinting that there may be something nefarious occurring there. Since *Charlottesville* is the entity that has the largest influence on the documents in the observation panel, documents that mention *Charlottesville* (the 5 documents towards the top of the observation panel) are separated from the ones that don't (in the middle of the observation panel). Inspecting these *Charlottesville* documents provides insight on a terrorist plot in Charlottesville involving several individuals.

However, Fig. 6-A has many of the observations and attributes overlapping with each other, making them harder to distinguish from each other. Although the attribute node labels have been removed to improve clarity in the projection, these labels are still accessible by hovering over a node.[3] However, this issue can be also alleviated using topic modeling to essentially group attributes together (and thus better separate the documents) during an additional preprocessing step. Visualizing the topics in place of the extracted entities results in the much clearer initial projections shown in Fig. 6-B. The tradeoff in doing so is that the previous initial insight that Charlottesville may be the center of some nefarious activity is lost.

[3]There are many methods for improving the display of labels in scatterplot-like visualizations [10, 17] However, this is not the main focus of this paper; we instead focus on new interactive projection techniques for displaying both observations and attributes of high-dimensional data and highlight the insights that can be gained.

5.2.2 Example Analysis

To show analysis on text data with SIRIUS, we use the topic modeled data to improve clarity. A reasonable starting point with a dataset like this is to pick out attributes (i.e., topics) that seem more indicative of nefarious activity than others. Examining the topics shown in the attribute panel of Fig. 6-B uncovers a number of such topics, including one that focuses on passport information (*dates/passport/address*), another on police activity (*polic/truck/two*), and a third on explosions (*report/driver/explos*). In contrast, topics like *<number>/cia/date* seem perhaps generic or less useful to the initial investigation. Performing PrI by dragging the three attributes of interest into one corner to and *<number>/cia/date* into the opposite corner to expresses their desired similarities/dissimilarities (depicted in the top row of Fig. 7) and clicking "Update Layout" above the attribute panel results in the visualization depicted in the bottom row of Fig. 7. This visualization gives the insight that the topics *<number>/report/phone* and *april/ave/report* are very closely correlated with the three topics of interest that were moved. These new topics, along with *fbi/list/<number>* (which is now the most emphasized topic), are all worthy of further investigation.

However, most notably, this interaction resulted in the document *fbi11* being highly emphasized in the observation panel, giving insight on its strong association with the emphasized topics in the attribute panel. Reading the contents of this document reveals information that happens to be central to one of the three main terrorist plots contained within the dataset, as further analysis can confirm.

5.3 A Breast Cancer Dataset

To demonstrate the ability of SIRIUS to enable exploration of larger datasets, Fig. 8 shows the "Breast Cancer Wisconsin (Original)" dataset from the UCI Machine Learning Repository [32]. Similar to Fig. 6, the observation labels have been removed to improve clarity. In this visualization, many observations representing benign tumors naturally group together in the lower left of the observation panel, separating themselves from those representing cancerous tumors.

Looking at node size and opacity in the attribute panel, we can see that the attributes that describe this separation are *Bland Chromatin*, *Single Cell Epithelial Size*, and, most notably, *Clump Thickness*. Thus, an analyst can immediately gain the insights that *Bland Chromatin* and *Single Cell Epithelial Size* do help distinguish between observations, but *Clump Thickness* describes the separation between the different observations better than the others. Additionally, since *Clump Thickness*

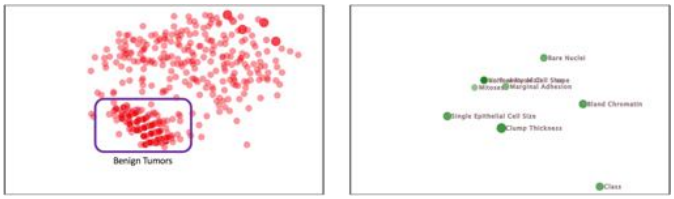

Fig. 8. An initial projection of the "Breast Cancer Wisconsin (Original)" dataset [32] in SIRIUS. Note that the dense group of nodes in the lower left of the observation panel correspond to benign tumors. The attribute projection reveals that the observation projection is best described by the *Clump Thickness* attribute. However, this attribute, along with *Single Cell Epithelial Size* and *Bland Chromatin* are the attributes that are most closely correlated with the *Class* attribute and therefore may be useful in diagnosing breast cancer in patients.

is the closest attribute node to *Class*, this means that *Clump Thickness* is the attribute that has the strongest correlation with *Class*. This correlation explains why observations seemed to be well-separated by class. Theses insights also mean that doctors may be able to use clump thickness, bland chromatin, and single cell epithelial size to help distinguish between cancerous and non-cancerous tumors. While these insights may be obvious to medical practitioners, the fact that SIRIUS immediately uncovers them demonstrates its ability to easily reveal critical information in high-dimensional datasets.

6 Discussion, Limitations, and Future Work

In our supplementary material, we thoroughly evaluate the capabilities of SIRIUS against 10 existing visual analytics projection techniques, emphasizing the different insights afforded. Here, we highlight that SIRIUS is enables exploratory data analysis on observations and attributes simultaneously and efficiently, as evidenced in Sect. 5. This includes insights such as attribute correlations while exploring observation similarities, which can be gained through few, simple interactions.

Using Andromeda [43] as a contrasting example, PaI and PrI are enabled on a single observation projection (accomplishing half of Goals 1 and 2). A separate Andromeda instance would need to be run simultaneously to display the attributes to enable the same interactions on an attribute projection (for the other half of Goal 1 and Goal 2). While this would provide the two projections and the same interactions within each, they would remain disconnected (falling short of Goal 3). Thus, the analyst would be forced to estimate or even guess how to manipulate one projection to reflect changes in the other. This process would be time-consuming and error-prone, easily resulting in incorrect conclusions. Therefore, we assert that SIRIUS provides a more powerful platform for performing data analytics tasks that incorporate both the observations and the attributes of a dataset, such as the example tasks in Sect. 5, than existing techniques.

However, using SIRIUS comes with the potential tradeoff of how one projection cannot be transposed on top of another (as exemplified by the bottom row in Fig. 5), which may be confusing for some analysts. Despite this, we highlight that every projection provided in SIRIUS is a valid projection that can provide rich, meaningful insights, as demonstrated in Sect. 5.

One potential limitation of our implementation of SIRIUS is in the usability and understandability of PrI caused by the fact that only a subset of nodes are used to calculate a new set of weights, which are then applied to all nodes. While Self et al. explore this limitation in [43], the user study described in [42] highlights the benefits that PrI can bring to the analysis process.

Another limitation of our implementation of SIRIUS is a "jumping" effect, which is most evident when performing PaI on each panel in sequence. This effect stems from our use of Equation 5 and Equation 6. For example, assume that PaI was just performed in the attribute panel. The weights for the changed attributes, W_A, are first fed into Equation 5. The observation weights, W_O, resulting from that equation is then fed into Equation 6 to determine a final set of attribute weights, W_A. Repetitions of this interaction follow the same flow; for small, subsequent changes to W_A, this series of steps results in similarly small changes to W_O. However, when an observation's importance value is then manipulated, the first step instead becomes that the new W_O is fed into Equation 6 to determine a new W_A. Since the W_A here is very different than the previous W_A, the projection of the attributes changes greatly. Therefore, although an analyst may expect both projections to change minimally, a large change is reflected in the attribute projection.

One cause of this "jumping" effect is that the cycle of Equation 5 and Equation 6 have a single convergence point, as suggested by the HITS algorithm [24]. However, since we want to explore alternative projections, we necessarily consider other pairs of weight vectors for which the cycle is not converged and therefore "jump" when the analyst switches their interaction back and forth between the projections.

A related cause of the "jumping" effect is the memorylessness of the interaction. When a new interaction is performed, our implementation of SIRIUS only uses the most recent interaction to update the visualization. One possible way to address this issue is to alter Equations 5–6 to incorporate previous interactions or projections. For instance, Leman et al. [30] suggest a weighted average of the weight vectors produced by the most recent interaction with vectors from previous interactions.

Lastly, a limitation to our implementation of SIRIUS is the size of data that is supported due to the the n^2 and n^2p^2 computational complexities for projection and interaction optimizations (respectively). However, recent performance improvements in the computation of WMDS projections [12] as well as PrI interactions (which others in the BaVA @ VT group are researching) can greatly increase the size of datasets that can be supported with interactive response times. However, this issue may also be alleviated for some datasets by implementing foraging features (e.g., searching and filtering for text data) in which only the most important observations and attributes are projected, such as StarSPIRE [4].

We intend to address these issues and limitations by continuing to develop our implementation of SIRIUS. This will also enable us to improve other aspects as well, such as the placement and size of the node labels in the projections or adding trails to highlight the impact the interaction had on the nodes' locations . With a more refined implementation in hand, we will be able to provide thorough evaluations on the time complexities of our refined algorithms as well as run user studies to evaluate interpretability and usability.

7 Conclusion

In this paper, we identified an opportunity for dual, symmetric, interactive projections of high-dimensional data to support the interconnectedness between observations and attributes in exploratory data analysis tasks. Given this need, we defined a generalized technique called SIRIUS, consisting of three principles: (1) dual projections of observations and attributes, (2) symmetric interactions on importances to explore projections, (3) symetrically relating importances of observations to importances of attributes. To concretize these principles, we described a specific implementation based on WMDS, weighted Euclidean distance, parametric and projection interactions, and dot-product importance calculations. A set of examples then demonstrated how SIRIUS provides insights across a range of diverse datasets, and we compared SIRIUS against a suite of existing techniques. SIRIUS offers new insight into both observations and attributes of high-dimensional data and their interrelationships, while maintaining a consistent symmetric mental model of each.

Acknowledgments

This research was partially supported by NSF grant IIS-1447416. We also thank the BaVA @ VT research group members for their contributions in developing SIRIUS as well as the reviewers for helping improve this paper.

References

[1] Z. Ahmed and C. Weaver. An adaptive parameter space-filling algorithm for highly interactive cluster exploration. In *2012 IEEE Conference on*

Visual Analytics Science and Technology (VAST), pp. 13–22, Oct 2012. doi: 10.1109/VAST.2012.6400493

[2] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo. Streamit: Dynamic visualization and interactive exploration of text streams. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, pp. 131–138. IEEE, 2011.

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pp. 111–117, Oct 2005. doi: 10.1109/INFVIS.2005.1532136

[4] L. Bradel, C. North, L. House, and S. Leman. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172, Oct 2014. doi: 10.1109/VAST.2014.7042492

[5] J. R. Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957. doi: 10.2307/1942268

[6] E. T. Brown. *Learning from Users Interactions with Visual Analytics Systems*. PhD thesis, Tufts University, 2015.

[7] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, Oct 2012. doi: 10.1109/VAST.2012.6400486

[8] S. Cheng and K. Mueller. The data context map: Fusing data and attributes into a unified display. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):121–130, 2016.

[9] J. Choo, H. Lee, J. Kihm, and H. Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 27–34, Oct 2010. doi: 10.1109/VAST.2010.5652443

[10] J. Christensen, J. Marks, and S. Shieber. An empirical study of algorithms for point-feature label placement. *ACM Trans. Graph.*, 14(3):203–232, July 1995. doi: 10.1145/212332.212334

[11] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 121–128, March 2010. doi: 10.1109/PACIFICVIS.2010.5429600

[12] S. Dash, A. Verma, C. North, and W. c. Feng. Portable parallel design of weighted multi-dimensional scaling for real-time data analysis. In *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 10–17, Dec 2017. doi: 10.1109/HPCC-SmartCity-DSS.2017.2

[13] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, Dec 2012. doi: 10.1109/TVCG.2012.260

[14] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 473–482. ACM, New York, NY, USA, 2012. doi: 10.1145/2207676.2207741

[15] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 121–130, Oct 2011. doi: 10.1109/VAST.2011.6102449

[16] D. W. Evans, P. T. Orr, S. M. Lazar, D. Breton, J. Gerard, D. H. Ledbetter, K. Janosco, J. Dotts, and H. Batchelder. Human preferences for symmetry: subjective experience, cognitive conflict and cortical brain activity. *PloS one*, 7(6):e38966, 2012.

[17] J.-D. Fekete and C. Plaisant. Excentric labeling: Dynamic neighborhood labeling for data visualization. In B. B. BEDERSON and B. SHNEIDERMAN, eds., *The Craft of Information Visualization*, Interactive Technologies, pp. 316 – 323. Morgan Kaufmann, San Francisco, 2003. doi: 10.1016/B978-155860915-0/50040-8

[18] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–871, 1971.

[19] P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations. In *In Proc of the NPIV 99*, pp. 9–16. ACM Press, 1999.

[20] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. ipca: An interactive system for pca-based visual analytics. *Computer Graphics Forum*, 28(3):767–774, 2009. doi: 10.1111/j.1467-8659.2009.01475.x

[21] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571, 2011.

[22] E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 107–116. ACM, New York, NY, USA, 2001. doi: 10.1145/502512.502530

[23] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):151–160, 2017.

[24] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[25] J. Krause, A. Dasgupta, J.-D. Fekete, and E. Bertini. Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *IEEE 6th Symposium on Large Data Analysis and Visualization*, 2016.

[26] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[27] J. B. Kruskal and M. Wish. *Multidimensional scaling*, vol. 11. Sage, 1978.

[28] B. C. Kwon, H. Kim, E. Wall, J. Choo, H. Park, and A. Endert. Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):221–230, 2017.

[29] C. H. Lampert, H. Nickisch, S. Harmeling, and J. Weidmann. Animals with attributes: A dataset for attribute based classification, 2009.

[30] S. C. Leman, L. House, D. Maiti, A. Endert, and C. North. Visual to parametric interaction (V2PI). *PloS one*, 8(3):e50474, 2013.

[31] Q. Li and C. North. Empirical comparison of dynamic query sliders and brushing histograms. In *IEEE Symposium on Information Visualization 2003*, pp. 147–153, Oct 2003. doi: 10.1109/INFVIS.2003.1249020

[32] M. Lichman. UCI machine learning repository, 2013.

[33] S. Liu, B. Wang, J. J. Thiagarajan, P.-T. Bremer, and V. Pascucci. Visual exploration of high-dimensional data: Subspace analysis through dynamic projections. *Technical Report UUSCI-2014-003*, 2014.

[34] Z. Liu, B. Jiang, and J. Heer. immens: Real-time visual querying of big data. *Computer Graphics Forum (Proc. EuroVis)*, 32, 2013.

[35] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: The vibe system. *Information Processing & Management*, 29(1):69 – 81, 1993. doi: 10.1016/0306-4573(93)90024-8

[36] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[37] D. Paranyushkin. Visualize any text as a network—textexture. retrieved november 18, 2012.

[38] K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[39] K. Pearson. Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, 6(2):566, 1901.

[40] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1759–1764. ACM, 2013.

[41] K. Schatz, C. Müller, M. Krone, J. Schneider, G. Reina, and T. Ertl. Interactive visual exploration of a trillion particles. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 56–64. IEEE, 2016.

[42] J. Z. Self, N. Self, L. House, J. R. Evia, S. Leman, and C. North. Bringing interactive visual analytics to the classroom for developing eda skills. 2015.

[43] J. Z. Self, R. Vinayagam, J. T. Fry, and C. North. Bridging the gap between user intention and model parameters for data analytics. In *SIGMOD 2016 Workshop on Human-In-the-Loop Data Analytics (HILDA 2016)*, p. 6, 06/2016 2016.

[44] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, 2002.

[45] J. Sharko, G. Grinstein, and K. A. Marx. Vectorized radviz and its application to multiple cluster datasets. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1444–1427, Nov 2008. doi: 10.1109/TVCG.2008.173

[46] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, Nov 1994. doi: 10.1109/52.329404

[47] Square Inc. Crossfilter: Fast multidimensional filtering for coordinated views, 2013.

[48] W. S. Torgerson. Theory and methods of scaling. 1958.

[49] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions-a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011.

[50] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, Dec 2012. doi: 10.1109/TVCG.2012.256

[51] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579 – 2605, 2008.

[52] J. Wenskovitch and C. North. Observation-level interaction with clustering and dimension reduction algorithms. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pp. 14:1–14:6. ACM, New York, NY, USA, 2017. doi: 10.1145/3077257.3077259

[53] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. doi: 10.1016/0169-7439(87)80084-9

[54] C. Xie, W. Zhong, and K. Mueller. A visual analytics approach for categorical joint distribution reconstruction from marginal projections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):51–60, 2017.

[55] L. Xu, Y. Xu, and T. W. Chow. Polsom: A new method for multidimensional data visualization. *Pattern Recognition*, 43(4):1668–1675, 2010. doi: 10.1016/j.patcog.2009.09.025

[56] Y. Xu, L. Xu, and T. W. Chow. Pposom: A new variant of polsom by using probabilistic assignment for multidimensional data visualization. *Neurocomputing*, 74(11):2018–2027, 2011. doi: 10.1016/j.neucom.2010.06.028

[57] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[58] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, 2013.

[59] G. G. Zanabria, L. G. Nonato, and E. Gomez-Nieto. istar (i*): An interactive star coordinates approach for high-dimensional data exploration. *Computers & Graphics*, 60:107–118, 2016. doi: 10.1016/j.cag.2016.08.007