Shape Identification in Temporal Data Sets

Machon B. Gregory*
Dept. of Computer Science
University of Maryland

Ben Shneiderman†
Dept. of Computer Science
and Human-Computer
Interaction Lab
University of Maryland

Abstract

Shapes are a concise way to describe temporal variable behaviors. Some commonly used shapes are spikes, sinks, rises, and drops. A spike describes a set of variable values that rapidly increase, then immediately rapidly decrease. The variable may be the value of a stock or a person's blood sugar levels. Shapes are abstract. Details such as the height of spike or its rate increase, are lost in the abstraction. These hidden details make it difficult to define shapes and compare one to another. For example, what attributes of a spike determine its "spikiness"? The ability to define and compare shapes is important because it allows shapes to be identified and ranked, according to an attribute of interest. Work has been done in the area of shape identification through pattern matching and other data mining techniques, but ideas combining the identification and comparison of shapes have received less attention. This paper fills the gap by presenting a set of shapes and the attributes by which they can identified, compared, and ranked. Neither the set of shapes, nor their attributes presented in this paper are exhaustive, but it provides an example of how a shape's attributes can be used for identification and comparison. The intention of this paper is not to replace any particular mathematical method of identifying a particular behavior, but to provide a toolset for knowledge discovery and an intuitive method of data mining for novices. Spikes, sinks, rises, drops, lines, plateaus, valleys, and gaps are the shapes presented in this paper. Several attributes for each shape are defined. These attributes will be the basis for constructing definitions that allow the shapes to be identified and ranked. The second contribution is an information visualization tool, TimeSearcher: Shape Search Edition (SSE), which allows users to explore data sets using the identification and ranking ideas in this paper.

Index Terms: Information Visualization, time series, shape identificaiton, temporal data, graphical user interface, lines, spikes, sinks, rises, drops, plateaus, valleys, gaps

1 Introduction

Shapes are a succinct way of describing the behavior of a temporal variable. For instance, a spike describes a sharp increase followed by a shape decrease. A shape describes a behavior abstractly. Therefore, the rate a spike increases or the height of the peak, as well as other details about the variable's behavior are lost. The absence of these details makes it difficult to compare one shape to another. For example, given a spike, how can it be described or compared to another spike? A lot of work has been done identifying a particular shape in a specific data set, but little work has been done to examine individual shapes and generalize their use.

Shapes such as spikes, drops and increasing lines are used by professionals in many different fields to describe the behavior of temporal variables. Stock market analysts use shapes to describe changes in stock prices. Published research results offer concrete evidence of the usefulness of shape identification. For example, spikes were used by Balog et al. to understand the mood of bloggers in relation to world events[2] and by Dettki and Erisson to analyze the seasonal migration patterns of moose[4]. These shapes are obvious in a visual representation to the informed observer, but they are often hard to describe precisely and compare to other shapes of the same type. The ability to identify and rank shapes of interest in a visualization of temporal data sets can be helpful to novice analyst and in knowledge discovery.

This paper examines eight simple shapes: lines, spikes, sinks, rises, drops, plateaus, valleys, and gaps. A spike is defined as a significant increase in value followed by a significant decrease in value in a set of sequential points. A sink is a significant decrease in value followed by a significant increase in value in a set of sequential points. A line is a set of sequential points with the same general behavior.A rise is a sustained increase in value in a set of sequential points. A drop is a sustained decrease in value in a set of sequential points. A plateau is a temporary increase in value in a set of sequential points. A valley is a temporary decrease in value in a set of sequential points. A gap is a specific type of valley where the values temporarily decrease to zero.

Each shape will be assessed by a set of measurable attributes. For example, a line shape's primary attributes are its endpoints and slope. An attribute, such as the "spikiness" of a spike, may be manifested as one or more measurements of the shape's attributes. Each measurement or set of measurements represents a different behavior. The attributes are used to define a shape's behavior and compare and rank the shapes. A shape definition consists of one or more constrained attributes. For instance, a line with the slope constrained to be positive defines an increasing line. A shape can have many definitions that identify different behaviors of interest. A ranking metric is one or more attributes by which a shape is compared to other shapes of the same definition and ranked. A ranking metric results from one or more calculations performed over values associated with a particular variable. The shapes that will be discussed are not an exhaustive set of shapes, nor are the attributes. This paper presents the idea of identifying behaviors of interest through shape identification, then ranking the shapes according a set of attributes.

The shapes and attributes that will be discussed are simple, as are the measurements of the attributes. This work is not a replacement for pattern mining techniques, used to identify a unique behavior in a data set. But the work presents a way of thinking about an identified behavior of interest and how it is defined and can be compared to other behaviors.

A subset of the shapes with multiple definitions for each are incorporated into TimeSearcher: Shape Search Edition (SSE), an information visualization tool. SSE is built upon TimeSearcher 1[11], and allows for the exploration of temporal data sets by identifying shapes of interest and ranking them according to a ranking metric. SSE visualizes shapes and provides a numerical ranking metric, which allows the shapes to be compared. SSE can identify shapes like increasing, decreasing, and volatile lines, as well as spike, sinks, rises, and drops. SSE has several definitions for each

*e-mail: machon.gregory@gmail.com
†e-mail:ben@cs.umd.edu

of the shapes to identify different types of behaviors.

2 BACKGROUND

A lot of research has been done to understand how to define shapes. Some of the research, such as Agrawal et al.'s shape definition language (SDL)[1] and Hochheiser, et al.'s timeboxes[10], focuses on allowing users to define a shape of interest and then identify them in a data set. Research in the area of pattern discovery has focused less on the definition of the pattern and more on the value of the identified pattern. Many of the papers on pattern discovery start to answer the question "How significant or interesting is the identified pattern?" Much of the work in this area takes an automated approach, examining sets of values in a data set and determining their value based on some function. The idea that patterns can be evaluated to estimate their value to the user is one of the ideas presented in this paper.

2.1 Shape Definition

Providing an expressive language for identifying and comparing shapes is one goal of this paper. Agrawal et al. and Hochheiser et al. present two distinct methods of defining shapes. Both are expressive, but for different reasons. Agrawal et al.'s SDL provides a language consisting of an alphabet and a set of operators to define a shape; Hochheiser et al.'s research has focused on visual widgets as the method of defining shapes.

SDL provides a simple alphabet to describe point to point transitions in time series data. For example, the user defined symbol "Up" may indicate a significant increase in a stock price from one time point to the next. Using the symbols and the operators users can define an alphabet to describe any shape. The symbols define the amount of variation from point to point and the operator describe the relationship between the symbols.

TimeSearcher 1, an information visualization tool for exploring time series data, provides different techniques for defining shapes. The TimeSearcher 1 uses timeboxes and several other types of queries to allow users to visually define shapes. Timeboxes facilitate shape definition by allowing users to visually specify a range of values for the x and y coordinates of the data points within a shape. In addition to the timeboxes, TimeSearcher 1, includes an angular query widget. The angular query widget allows users to define a range of slope values. The timeboxes are a fairly course-grained approach to defining shapes, the angular queries provide a much more granular approach.

QuerySketch[16] allows the user to define shapes of interest using freehand sketches. Similarly, QueryLines provides a structured method of creating shapes using a series of line segments. QueryLines[14] combines the point-to-point expressiveness of SDL and the dynamic visual query language of TimeSearcher 1. QueryLines is an information visualization tool that incorporates visual shape definition and user defined rankings to identify shapes of interest in temporal and ordered data sets. QueryLines, like TimeSearcher, can be bound by the x or y values, or both. QueryLines also allows the user to specify a set of contigous line segments that define a shape; identified shapes are ranked according to their similarity to the user defined shape.

SDL, TimeSearcher 1, and QueryLines enable users to define shapes of interest and locate their occurrences within a data set. SDL is an expressive solution that can be tailored to the needs of its users, but it could be hard to use effectively by common users. On the other hand, TimeSearcher 1, is less expressive, but provide the users with the ability to define shapes in terms they understand (what they can see visually see). Keogh et al. extended timeboxes create variable time timeboxes (VTT) to increase their expressiveness[11]. VTT allows user to define a shape and then locate it over a range of values. Other research offers expressive ways of defining shapes over categorical data, such as temporal logic[12] and regular expressions[7], but the techniques do not easily transfer to temporal data sets. QueryLines has the expressiveness of SDL in visual query tool, but it is unable to express higher level behaviors, like anomolous spikes.

2.2 Shape Evaluation

In SDL and TimeSearcher 1, the significance of a shape is based strictly on whether the shape conforms to the definition or not. Although, the values used by the angular query widget could be used to define the significance of the identified shape, it is not an inherent capability in the tool. Because all shapes have the same significance they can not be compared to one another. However, research in the area of pattern discovery[9, 8] focuses on evaluating the significance of identified shapes. The ability to evaluate the significance of a shape implies that the identified shapes are comparable by some measurable attribute. For example, Dubinko et al.'s research in visualizing the evolution of social network tags, defines "interestingness" as the likelihood of a tag occurring during a particular period of time[5]. The definition of "interestingness provides a measurable attribute, frequency of tags occurrence during a particular period of time, by which tags can be compared. Similarly, clustering techniques are used to identify patterns of interest. In this technique, similar patterns are grouped together into a cluster[6, 3]. Patterns identified using this technique can be compared based on the size of the cluster, the larger the cluster the more interesting the the pattern. Yang et al.'s STAMP algorithm uses statistics to measure the importance of identified patterns[17]. Each of these techniques provides a metric by which an identified pattern can be compared to another pattern. Unfortunately, these techniques are primarily associated with pattern discovery techniques and offer users little control over what patterns are identified.

Garofalakis et al. recognized the "lack of user controlled focus in the pattern mining process" and introduce a set of algorithms deemed SPIRIT, Sequential Pattern Mining with Regular Expression Constraints[7]. This research combines the ability to identify significance by using some measurable attributes, frequency, and an expressive definition language, regular expressions. The regular expressions provide users with the ability to constrain the results returned by the pattern mining algorithm to just the patterns of interest to the users. This paper goals are to provide capabilities similar to the SPIRIT algorithms, shape identification and ranking techniques using a user defined shape definition. Going beyond the SPIRIT algorithms this paper presents techniques that allow users to define what is"interesting" in terms that are familar to them.

In addition to assisting users in defining shapes this paper presents attributes by which shapes can be ranked. There are many novel techniques for identifying similar patterns, but few offer users the ability to direct the ranking of the results. The idea of ranking data according to user-specific feature is not new, Seo and Shneiderman's created the rank-by-feature framework to assist users in selecting a feature that may interest them[15].

3 SHAPE DEFINITIONS

There are an infinite number of shapes; many of them are too complex to describe succinctly or create mathematical definitions to identify them. But there are a set of simple shapes that are commonly used to describe a particular behavior. In the following sections several shapes will be described, as well as their attributes. These attributes will be used to provide examples of shape definitions and ranking metrics. Additionally, examples explaining how the shapes, their definitions and ranking metric may be used to answer different types of queries will be given. Line, spike, sink, rise, drop, plateau, valley and gap shapes will be discussed.

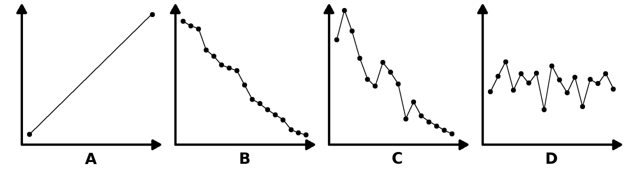


Figure 1: Graphs A through D show examples of line shapes. A shows a 2-point increasing line and B a multi-point constantly decreasing line. C is an example of a multi-point decreasing line that could be identified by a linear regression calculated using the values that compose it. The last graph, D, is an example of a volatile line, where volatility is a measure of the standard deviation of the values in the line.

3.1 Line Shapes

The simplest shape, a line, is defined as one or more line segments created by a set of contiguous time points. In a 2D Cartesian plane, a geometric line can be defined using the equation, $y = mx + b$, where m is the slope, b is the y-intercept, and x is an independent variable. A line segment is a portion of a line defined by its endpoints. Line shapes are interesting because they can be used to describe any other shape, but they are most useful in describing consistent behaviors, such as generally increasing, decreasing, stable, or volatile periods. For instance, a stock that consistently rises over a period of time can be described by an increasing line shape. Depending on how its attributes are constrained, a line shape can be used to generalize the behavior of a set of time points or identify a specific behavior, that is characterized by limited range of value changes between time points. For example, a linear regression identifies a relationship between a set of variables, that generalizes their behavior, but calculating the slope of each individual line segment can identify a specific behavior.

The attributes associated with line shapes are the length, slope, and volatility. The length attribute is the number of time points in the shape. The slope attribute is a measure of the rate at which the line shape is changing. The slope definition varies depending on whether the goal is to identify a particular behavior in the time series or to generalize the behavior of a set of time points. To identify a specific behavior, slope can be defined as the change in value between two time points. This definition is identical to the definition of slope for a geometric line. Using this definition of slope any constraint applied to the slope must be consistent across every line segment in the line shape. For example, if one line segment is increasing, all line segments in the line shape must be increasing. On the other hand, if the goal is to generalize the behavior of a set of time points, the slope definition should consider all of the points together. For example, the slope of a line shape may be defined as:

- amount of change between two time points that may or may not be contiguous
- the sum of the change of between all contiguous time points in the line shape
- the geometric slope of a linear regression computed over the time points in the shape.

These are examples of ways of calculating slopes. Figure 1C shows a line that could be identified using a linear regression, the set of values in the line have a decreasing trend. Each of these definitions describes a different behavior that may be of interest. Using different definitions for slope will result in different slope calculations for line shapes, therefore identifying different behaviors.

The term volatility can refer to the relative rate at which a stock increases and decreases. The same definition will be used to describe the volatility attribute of a line shape. The standard deviation of the values within a line shape can be used as a measure of a line's volatility. Figure 1D is an example of a volatile line. Other calculations may be more appropriate for measuring the volatility of line shape depending on the behavior of interest.

The slope, length, and volatility are attributes by which line shapes can be defined and ranked. Constraining the slope of a line shape to be a positive or negative value creates two definitions of line shapes, increasing and decreasing, respectively. According to the slope definition, an increasing line shape will characterize different behaviors. Constraining each individual line segment in a line shape to be negative creates a monotonically decreasing line, like the line shape in Figure 1B. In addition to constraining the slope of the line, the number of time points can also be constrained. Figure 1A is an example of a 2-point line shape; Figures 1B, 1C and 1D are examples of multiple point line shapes.

3.2 Spike and Sink Shapes

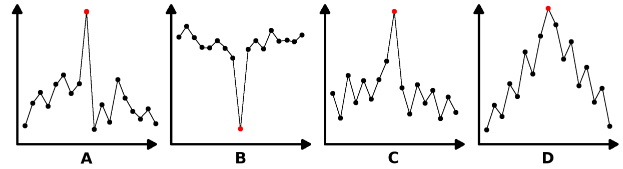


Figure 2: These graphs are examples of spike and sink shapes. The red dots are the peak points. Graph A, B, and C are graphs that may be ranked high based on its relative or angular height. The relative height is a measure of the difference between the peak point and average value of the remainder of the points. The angular height is the measure of the angle created by the two edges that meet at the peak point. An edge may consist of one or more points. Graph D is a spike that could be identified using a linear regression calculated over the points in the edges to the right and left of the peak point.

Spikes and sinks describe a temporal behavior in which a variable has a significant change over a period of time in one direction and then a significant change in the opposite direction. The point at which this change in direction occurs is the peak point. A spike, specifically, is a significant increase followed by a significant decrease. A sink is just the opposite, a decrease followed by an increase. Spikes and sinks are used by stock market analyst to describe the behavior of stock prices. Similarly, a doctor may say when blood pressure spikes there is a rapid rise then fall in pressure. Although the general behavior of spikes and sinks are understood more information is need to identify and compare particular instances of the shapes.

The attributes associated with spike and sink shapes are the significance of the increase or decrease and their duration. The significant can be manifested in one or more attributes. The significance of the change can be measured by the absolute, relative, or angular height of the peak point. The absolute height is the absolute value of the peak point. The angular height is defined by the angle created at the peak point. The relative height is defined as the height of the peak point relative to all the other points in the time series. This definition will identify spikes and sinks whose behavior is significantly different then the rest of the the points in the time series. For example, the equation, $|(max - mean)|/\sigma$ could be used to define the relative height of a spike or sink.

The relative height attribute of a spike or sink shape is affected by the behavior of all the time points in the time series. The absolute and angular height definitions have the ability to identify spikes and sinks in a volatile time series. Volatile time series are characterized by large changes in opposite directions between a set of consecutive time points. The duration attribute is given by the sum of time points contained in both edges plus the peak point. Constraining these attributes can identify a specific spike or sink shape within a time series.

The absolute, angular, and relative height attributes, as well as the duration and edge slope attributes can be constrained to define different spike and sink shape behaviors and they can be used as a ranking metric to compare and rank the shapes. The duration attribute can be constrained to identify sink and spike shapes that occur over a specific period of time. For instance, a three point and multiple point definition could be defined. The three point shape contains exactly three time points, a peak point and a single point on each side. Three points is the smallest number of points that a spike or sink shape can contain. The multiple point shape contains more than three time points.

The peak height can be constrained to create a definition that will identify shapes which are greater or less than a particular height. The slope of the leading or trailing period of change can also be used to define behaviors of interest for spike and sink shapes. By using these attributes to create shape definitions and rank shapes, particular behaviors of interest can be identified in temporal data sets.

3.3 Rise and Drop Shapes

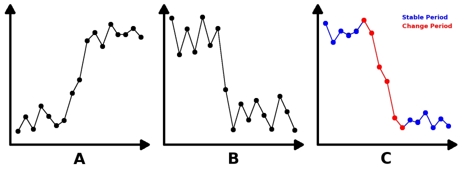


Figure 3: The graphs above are examples of rise and drop shape. Graph A is a rise. B and C are drops. Graph C shows the three periods of drop and rise shapes: the leading stable period, the change period, and the trailing stable period.

Rise and drop shapes are used to describe a sustained change in the average value. These shapes can be divided into three distinct periods: a period of change that is preceded and followed by a periods of stability, Figure 3C. The stable periods are drawn in blue and the period of change in red. A rise shape has a change period that increases in value, while a drop shape decreases in value, as seen in Figures 3A and 3B, respectively. Each period must consist of one or more time points; there is a single transition point between each period; and the time points in the shape must be contiguous. Drop and rise shapes contain a minimum of five points. The periods of stability separate these shapes from spikes, sinks and lines.

Stable time points have very low volatility, which could be measured by the standard deviation of the points or some other definition. In drops and rises, if a set of time points is not stable, it is changing. A rise and drop shape describes a person's heart rate at the start and conclusion of a aerobic workout, respectively. At the start of a workout a healthy person's heart rate will transition from a resting rate of approximately 65 beats per minute (bpm) to 140 bpm. During the period prior to and after the transition the active and resting heart rate will be stable until the conclusion of the workout. This is the type of behavior a rise or drop shape could identify.

The length of the periods, the change significance, and the average value of the stable periods are some of the attributes associated with rise and drop shapes. The length of a period is defined by the number of time points contained within that period. The change significance, like the previous shapes, can be defined by the slope of that period, and the slope can be defined in several different ways based on the behavior of interest. The average value of the stable period is the mean of the points in the period.

Period length is the most intuitive attribute to constrain when creating shape definitions for rise and drop shapes. A definition that limits the length of the change period to just two points is useful in identifying rapid change.

3.4 Plateaus, Valleys and Gaps

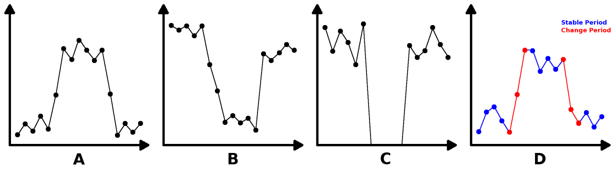


Figure 4: Graphs A, B and C show a plateau, valley and gap shape, respectively. Graph D shows the periods associated with plateau, valley and gap shapes.

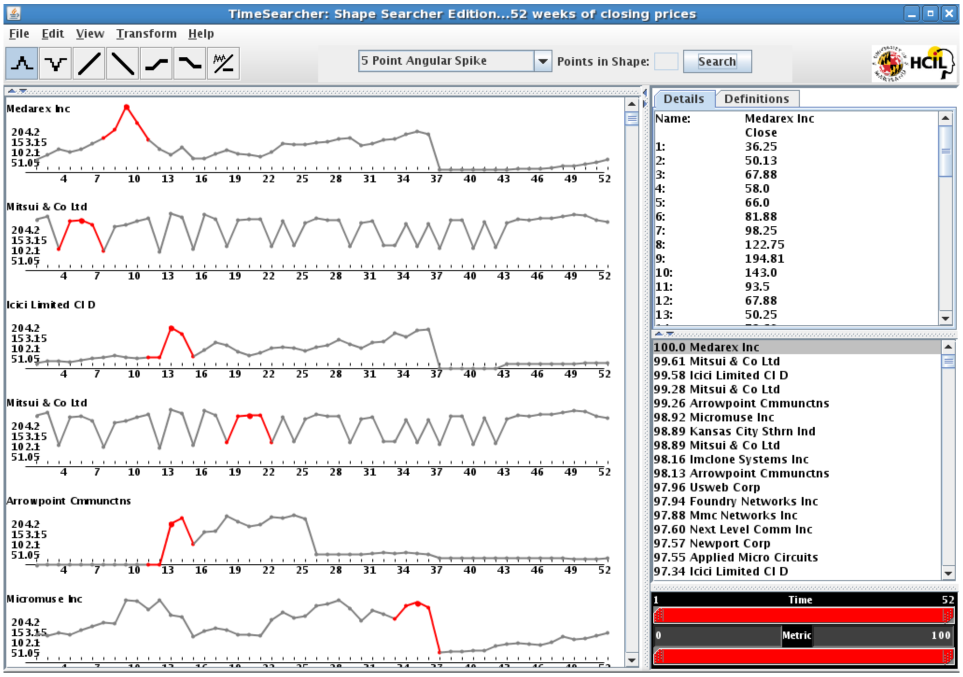
Plateaus, valleys, and gaps are used to describe temporary changes in variable. They differ from spikes and sinks because the temporary value is sustained for a measurable period of time. These shapes consist of leading, intermediate, and trailing stable periods, as well as departing and returning change periods, as shown in Figure 4D. A plateau has an intermediate stable period, whose average value is greater than the leading and trailing stable periods (Figure 4A), while a valley has an intermediate period, whose average value is less than the average value of other two stable periods (Figure 4B). A gap is a specific type of valley where the intermediate period's values are zero (Figure 4C). Using the workout example, a plateau describes a person's heart rate during his or her entire workout. Prior to the beginning and after the end of the workout, the heart rate is stable at a resting rate of 65 bpm. At the start of the workout, to the heart rate leaves the resting rate and rises to approximately 140 bpm. This heart rate is maintained throughout the workout. At the conclusion of the workout, the heart rate returns to the resting heart rate and remains there. Plateaus, valleys, and gaps are very similar to drops and rises, with one important difference. Drops and rises do not define the behavior that occurs after the trailing stable period. Therefore, several ranking metrics, such as the length of the intermediate stable period (the trailing stable period in the drop and rise shape), have a different meaning in plateaus, valleys and gaps than in drop and rise shapes.

The ranking metrics are similar to the ranking metrics for drops and rises, but they are calculated over the additional portions of the plateaus, valleys, and gaps. Although the calculations are same, the meanings are different. For example, using the workout example, the difference between the mean of leading and trailing stable periods in plateau shapes may signify a strengthening of the heart. On the other hand, the difference between the leading and trailing periods in a rise shape signifies a more strenuous workout.

Definitions that constrain the length of the stable periods are useful when examining plateau, valley and gap shapes. By limiting the length of a particular period, shapes with a specific duration can be identified. Definitions that measure the difference between the leading and trailing stable periods can also be useful.

4 TimeSearcher: Shape Search Edition

TimeSearcher: Shape Search Edition (SSE) is an information visualization tool that allows users to identify shapes and rank them according to one or more attributes. TimeSearcher SSE is an extension of TimeSearcher 1. TimeSearcher SSE can identify several definitions for each of the following shapes, lines, spikes, sinks, rises and drops and they can each ranked according different attributes. The definitions and ranking attributes are primarly static, but some of the definitions require user input.



Name:	Medarex Inc
	Close
1:	36.25
2:	50.13
3:	67.88
4:	58.0
5:	66.0
6:	81.88
7:	98.25
8:	122.75
9:	194.81
10:	143.0
11:	93.5
12:	67.88
13:	50.25

100.0 Medarex Inc
99.61 Mitsui & Co Ltd
99.58 Icici Limited Cl D
99.28 Mitsui & Co Ltd
99.26 Arrowpoint Cmmunctns
98.92 Micromuse Inc
98.89 Kansas City Sthrn Ind
98.89 Mitsui & Co Ltd
98.16 Imclone Systems Inc
98.13 Arrowpoint Cmmunctns
97.96 Usweb Corp
97.94 Foundry Networks Inc
97.88 Mmc Networks Inc
97.60 Next Level Comm Inc
97.57 Newport Corp
97.55 Applied Micro Circuits
97.34 Icici Limited Cl D

Figure 5: This is a screenshot of TimeSearcher Shape Searcher Edition (SSE). The upper panel shows the seven buttons labeled with the shapes that TimeSearcher SSE can identify and rank. Each shape has several definitions that can be selected from the drop down box to the right of the shape buttons. Some of the shape definitions require user defined input, such as the number of time points in the shape. The left side contains the shapes window, which displays the currently identified shapes for the loaded data set. The window in the upper right contains the details and definitions tab. The details tab displays the time points and values of a particular time series. The definition tab displays an explanation of the selected shape definition. The window in the left center is the rankings window. Once a shape and definition have been chosen from the upper panel the ranking metric value and label for each shape will be shown in this window. The lower right corner contains the dynamic query bars. These bars allow the shapes to be filtered based on the ranking metric and the endpoints associated with a shape.

4.1 Interface

TimeSearcher SSE consists of four primary windows. The shapes window on the left side contains time series graphs displaying each of the identified shapes. The tabbed window on the upper right side shows a details view, the time points and associated data values, of a time series in the details tab and the current shape definition in the definitions tab. The rankings window is on the right side in the center. This window displays the ranking metric for an individual shape and the label for the time series in which it is located. The shapes, details and rankings windows are tightly connected. Scrolling in the shapes window causes the rankings window to scroll, so that the first item in the rankings window is the same as the first graph in the shapes window. Selecting an item in the ranking window will cause the details for that time series to be shown in the details window and graph containing the shape to be the first one shown in the shapes window. Similarly, mousing over a graph in the shapes window will cause the details of the graph to be shown. The window on the lower right hand side contains range sliders which filter the identified shapes based on its endpoints and the value of the ranking metric.

The graphs in the shapes window are a visual representation of time series. These graphs make it easy to identify the shapes created by plotting the values in a time series. The graph's y-axis is labeled with the range of values that the variable takes on throughout the entire data set. The x-axis is labeled with the time points. The axes are drawn in black, while the time series is plotted in gray. Each time point is represented by a small gray dot and each consecutive dot is connected by a gray line. Each shape is shown in its own graph; if a time series has more than one unique occurrence of a shape, then the graph of the time series will appear more than once. Each shape is labeled in the graph with red lines instead of gray; points of interest in the shape are marked by large red dots. A significant point may be the peak point in a spike or sink shape or the change period in a rise or drop shape.

4.2 Spike and Sink Shape Identification

TimeSearcher SSE has three definitions for both spike and sink shapes. The definitions define a three, five, and seven point spikes and sinks and each of these shape definitions can be ranked according to its relative and angular height. Each of the shape definitions and ranking metrics are described below:

- **3-Point Spike/ Sink** – a spike or sink shape containing exactly three time points with a single time point on both sides of the peak point.
- **5-Point Spike/ Sink** – a spike or sink shape containing exactly five time points with two time points on both sides of the peak point.
- **7-Point Spike/ Sink** – a spike or sink shape containing exactly seven time points with three time points on both sides of the peak point.

- *Angular Height* – the measure of the angle created at the point where the edges meet. Figure 6A shows the component's angular height calculation. Using the trigonometric function, $cos(\alpha+\beta) = cos(\alpha)*cos(\beta) - sin(\alpha)*sin(\beta)$, the angle created by the edges of the spike is equal to $cos(\alpha+\beta) = (dy1*dy2-1)/\sqrt{(1+dy1^2)*(1+dy2^2)}$ where $dy1 = |y1-y2|$ and $dy2 = |y2-y3|$. A linear regression calculated over the points to the right and left of the peak point, defines the increasing and decreasing edges for the 5 and 7-point spikes.
- *Relative Height* – the height of the peak point from the mean of the time series measured in standard deviations. The relative height is given by the equation $|max - mean|/\sigma$. Figure 6B shows the values of the relative height calculation.

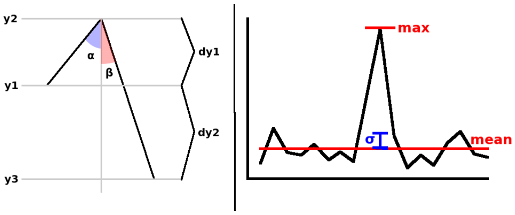


Figure 6: The diagrams above show how the angular and relative height attributes are calculated. The first image shows the components of the angular height equation, $cos(\alpha+\beta) = (dy1*dy2-1)/\sqrt{(1+dy1^2)*(1+dy2^2)}$. The angular height a measure of the angle created where the two edge of spikes and sinks meet. The second image shows the components of the relative height equation, $|max - mean|/\sigma$. The relative height is the height of a spike or sink relative to the rest of the shape.

All of the definitions and ranking metrics are static, and require no input from users. Each shape is computed when the data is loaded. Values such as the mean and standard deviation are only calculated once and stored within the internal representation of a time series, an Entity object. The function that identifies the spikes and sinks takes a parameter that defines how many points will be in a spike or sink. These shapes are identified simultaneously. The class attempts to identify shapes as efficiently as possible, by only passing through the data once. Figures 7 – 9 show examples of spikes and sinks identified by TimeSearcher SSE.

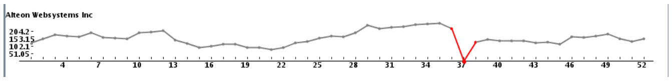


Figure 7: Example of a three point sink ranked according to its angular height. This sink identifies a missing value in this stock market data.

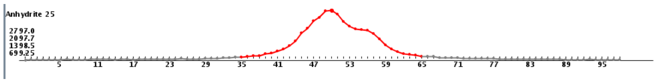


Figure 8: Example of a 31 point spike identified in X-ray diffraction data ranked according to its angular height.

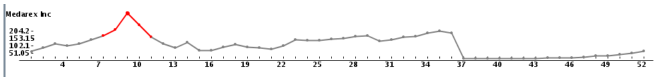


Figure 9: Example of a five point spike in a stock price that is highly ranked according to its angular height.

4.3 Line Shape Identification

TimeSearcher SSE contains four definitions for both increasing and decreasing line shapes and a single definition for volatile lines. The first three shape definitions for increasing and decreasing lines are two point, multiple point, and monotonic slope line shapes. The fourth definition is a monotonic slope line shape with a constraint placed on the minimum length. The two point and multiple point definitions are ranked according to their slope, while the monotonic slope definition is ranked according to its length and slope. Volatile lines are defined and ranked according to their standard deviation. The definitions for each shape and ranking metrics are listed below:

- **2-Point Line** – a line shape that contains only two time points. An increasing line has a positive slope, while a decreasing line's slope is negative.

- **Multiple Point Line** – a line shape that contains multiple time points. An increasing line has a positive geometric slope, while a decreasing line's slope is negative. There are several ways to measure the slope which are discussed below.

- **Monotonic Slope Line** – a line shape where each line segment's geometric slope has the same sign, positive or negative.

- *Slope* – the geometric slope is given by the equation, $(y2-y1)/(x2-x1)$. The slope of a two point line shape or a line segment can be calculated using the geometric slope equation. A multiple point line's slope is a measure of the geometric slope of the linear regression calculated over the points in the line shape. The slope of a monotonic slope line is calculated in the same fashion.

- *Length* – the number of time points contained in the line shape.

This class contains functions to identify multiple point lines and lines with monotonic slopes. Both functions are passed parameters by the user to identify multiple point lines of a particular length and monotonic slope line greater than a particular length. This allows the user to specify a minimum length for the monotonic slope lines, eliminating the two point lines, which are always monotonic. Figures 10 – 12 shows examples increasing and decreasing lines identified by TimeSearcher SSE.

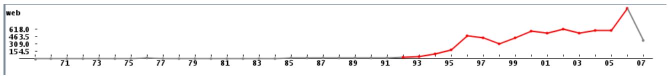


Figure 10: Example of a fifteen point increasing line ranked according to slope. This line shows the web the term “web” increasing over a fifteen year period.

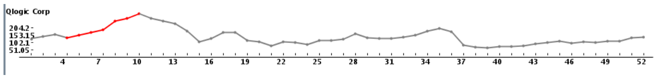


Figure 11: Example of a monotonically increasing line in stock market data ranked highly due to its slope.

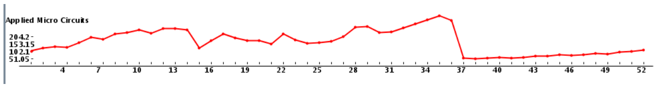


Figure 12: Example of a volatile line shape ranked highly according to its standard deviation.

4.4 Rise and Drop Shape Identification

TimeSearcher SSE contains three definitions for both rise and drop shapes. These definitions are ranked according to their slope and the length of their stable periods. The definitions defined by TimeSearcher SSE are general definitions described in Section 3.3 and the same definition except the length attribute of the stable periods is constrained to be a minimum length. Listed below are the definitions:

- **Rise or Drop** – a sustained change in values. These shapes consist of three distinct time periods: a stable period, followed by a period of change, concluding with another stable period.

- **Drop or Rise with Multiple Point Stable Period** – a rise or drop shape that contains multiple points in each of its stable periods.

- *Slope* – the geometric slope of the period of change. The slope of the period of change and a line shape are calculated the same way.

- *Length of the Stable Periods* – the lowest number of time points between the two stable periods.

A point is stable if it lies within one standard deviation of the mean of the other points within the stable period. If a point is not stable then it is changing. Figures 13 – 14 are examples of rise and drop shapes.

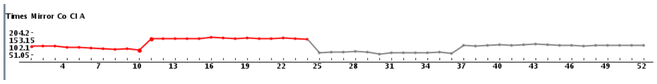


Figure 13: Example of a rise shape in stock market data. The shape is highly ranked according to the length of its stable periods.

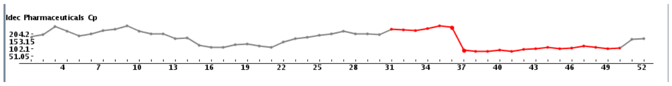


Figure 14: Example of a drop shape in stock market data. This drop was identified using the “stable period greater than x” definition which is ranked according to the slope of the change period.

5 TimeSearcher SSE Case Study

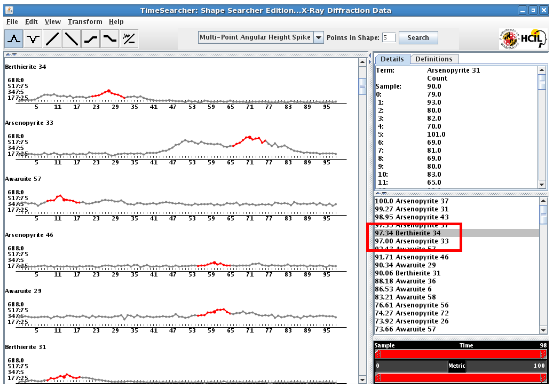
TimeSearcher SSE was given to an user to evaluate. The user participated in 4 one hour sessions. He examined two different data sets, network traffic and X-ray diffraction data. The user has been developing information visualization tools to examine network traffic data for the last 2 years. In his previous position he spent 5 years as a research physicist using tomography and angular and energy dispersive X-ray diffraction to idenitify unknown materials.

The network traffic data set consisted of a the number of server connections per hour made by a particular set of internet protocol (IP) addresses over a year. The other three data sets consider of angular dispersive X-ray diffraction (ADXRD) and energy dispersive X-ray diffraction (EDXRD) data. X-ray diffraction is used to observe properties of materials, such as their chemical composition or a specific physical property, by shining an X-ray beam on a material across a range of angles and measuring the scattered intensity of the X-ray as a function of the incident and scattered angle, polarization, and wavelength. The intensity readings produced by this process can be used as a fingerprint for a material. The fingerprint consists of intensity readings at various angles. For example, the element copper (Cu) may produce high intensity readings at angles 19.65, 23.0, and 33.5. Similarly, materials containing copper, such as covellite (CuS), would produce high intensity readings at similar angles. In ADXRD, peaks are very sharp, as opposed to EDXRD, where peaks are much broader. Two of the data sets contained ADXRD data and the other EDXRD data. The X-ray diffraction data was gather from an online database[13]. The goal of the X-ray diffraction case studies was to simulate the identification process of a unknown substance, by examining a set of materials with one or more common elements.

The ability to identify the common element would replicate using a set known elements to identify the elements that make up an unknown material. The network traffic data was sparse and while exploring the user noticed small spikes when particular IP addresses connected to the server, but due to the sparseness of the data set, he was unable to make any significant discoveries.

The X-ray diffraction data had more interesting results. The users was able to identify common spikes in each of the data sets. Iron (Fe), silicon (Si), and sulfur (S) were the common elements in each of the sessions. In each of the session the user was able to identify the common element using spike identification and rankings. Each of the data sets yield different results.

In the second session using the silicon ADXRD data some of the shortcomings of TimeSearcher SSE were shown. It was limited in its ability to display extremely large data sets. Each sample contained approximately 8500 intensity readings over a large range of angles. TimeSearcher SSE was not designed to handle such a large number of time points in a single time series, so each sample was divided into 85 separate samples with 100 time points in each sample. This caused shapes to be split over multiple graph limiting TimeSearcher ability to identify certain shapes. TimeSearcher SSE was also limited in its ability to dynamic define shape definitions. But once these shortcoming were identified the user was able to use different a different shape to find the behavior he was interested in. He began to look for increasing lines at the end of the time series and decreasing lines at the beginning of the following time series.



Term: Arsenopyrite 31
Count
Sample: 90.0
0: 79.0
1: 93.0
2: 80.0
3: 82.0
4: 70.0
5: 101.0
6: 69.0
7: 81.0
8: 69.0
9: 80.0
10: 83.0
11: 65.0

100.0 Arsenopyrite 37
99.27 Arsenopyrite 31
98.95 Arsenopyrite 43
97.34 Berthierite 34
97.00 Arsenopyrite 33
91.71 Arsenopyrite 46
90.34 Awaruite 29
90.06 Berthierite 31
88.18 Awaruite 36
86.53 Awaruite 6
83.21 Awaruite 58
76.61 Arsenopyrite 56
74.27 Arsenopyrite 72
73.92 Arsenopyrite 26
73.66 Awaruite 57

Figure 15: The user in the case study was able to identify spikes in Beritherite and Arsenopyrite at similar angular positions with TimeSearcher SSE. The top two graph show the match spikes. The element name and range of angular positions are in the upper left corner. TimeSearcher SSE was loaded with raw powder X-ray diffraction data for Beritherite, Awarite, and Arsenopyrite, materials all containing Fe. A ten point spike ranked according to its angular height was used to identify these spikes. The angular height value is squared in red in the ranking window. This discovery indicates that spikes at this position may be caused by the presence of Fe in the materials.

In the third session the user was able to identify some spikes with matching intensities and angular positions in the iron ADXRD data set. The materials in the data set were Arsenopyrite (FeAsS), Berthierite ($FeSb_2S_4$), and Awaruite (Ni_3Fe), which all contain Fe. By experimenting with spikes containing varying number of points, the user was able to see that Arsenopyrite and Berthierite both had spikes at angular position 33 and 34, respectively. A ten point spike definition was used and the results were ranked according to their angular height. The Arsenopyrite and Berthierite spikes were ranked consecutively, with normalized values of 97.00 and 97.34, respectively, as shown in Figure 15. The third material, Awarite, did not a have a spike ranked at the 33^{rd} or 34^{th} angular positions. But using an eight point spike definition ranked according to the angular height, Awarite and Arsenopyrite appear in the ranking window close together. The angular height of Awarite's spike at position 33 has a normalized value of 98.3, and the Arsenopyrite a value of 96.3, as shown in Figures 16. Although a definition and ranking metric that ranked the spikes in similar position for all three materials together was not found, a correlation could be drawn from the results.

The best results were given by the EDXRD data, which was generated by using infrared diffraction. In the fourth session IR diffraction data was collected for Anhydrite ($CaSO_4$) and Baryte ($BaSO_4$). Spikes at similar wavelength with similar intensities were identified in each material, as shown in Figure 17. The spikes had normalized values of 99.97 in both the Anhydrite and Baryte samples. The spike definitions were able to describe and rank the spike in the X-ray diffraction data set.

In addition to the discoveries made in the user case study the authors have used TimeSearcher SSE to analyze several data sets. These data sets included weekly closing prices for several hundred stocks, word frequency for the book, *The Making of Americans* by Gertrude Stein, and 125 Human and Computer Interaction (HCI) keywords over 37 years in a database 40,000 abstracts. In each of these data sets the authors were able find interesting patterns, unexpected anomalies, and errors in the data sets via shape identification and ranking.

6 Conclusion

Shapes are used to describe variable behaviors. Most people are familiar with the behaviors that are described by shapes. By providing users attributes to describe and rank shapes particular behaviors can be more easily identified and knowledge discovery becomes a more intuitive process.

References

[1] R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zaot. Querying shapes of histories. In *Proc. 21st International Conference on Very Large Databases*, pages 502–514. Morgan Kaufmann Publishers, Inc, 1995.

[2] K. Balog, G. Mishne, and M. Rijke. Rijke. why are they excited? identifying and explaining spikes in blog mood levels. In *Proc. 11th Meeting of the European Chapter of the Association for Computational Linguistics*, 2006.

[3] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 16–22, 1998.

[4] H. Dettki and G. Ericsson. Screening radiolocation datasets for movement strategies with time series segmentation. *Journal of Wildlife Management*, 72:535–542, 2008.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. of the 15th International WWW Conference*, 2006.

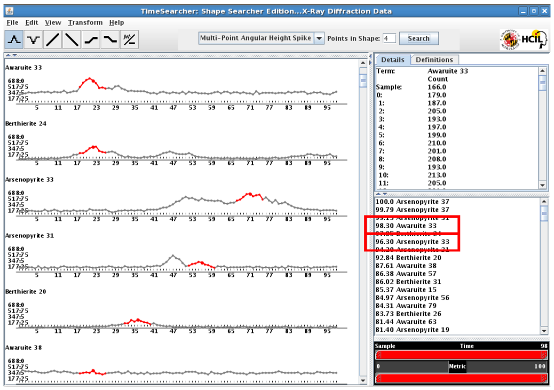
[6] T. C. Fu, F. L. Chung, V. Ng, and R. Luk. Pattern discovery for stock time series using self-organizing maps. In *in Workshop on Temporal Data Mining, 7th International Conference on Knowledge Discovery and Data Mining*, pages 27–37. ACM Press, 2001.

[7] M. N. Garofalakis, R. Rastogi, and Shim. Spirit: Sequential pattern mining with regular expression constraints. In *Proc. of the 25th International Conference on Very Large Databases*, pages 223–234, 1999.

[8] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 33–42, 1999.

[9] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 214–218. AAAI Press, 1998.

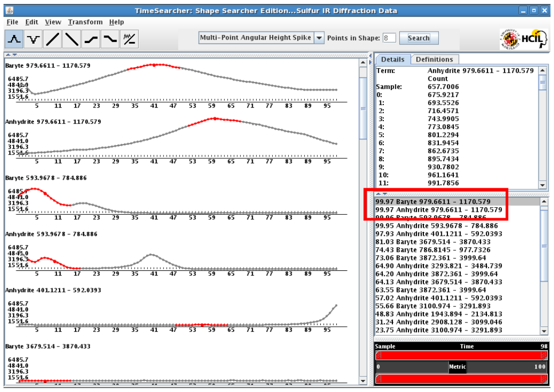
[10] H. Hochheiser. *Visual Queries for finding patterns in time series data*. PhD thesis, University of Maryland Computer Science Dept, 2002.



```
Term:       Awaruite 33
            Count
Sample:     166.0
0:          179.0
 1:         187.0
 2:         205.0
 3:         193.0
 4:         197.0
 5:         199.0
 6:         210.0
 7:         201.0
 8:         208.0
 9:         193.0
 10:        213.0
 11:        205.0
```

```
100.0 Arsenopyrite 37
99.79 Arsenopyrite 37
99.15 Arsenopyrite 31
98.30 Awaruite 33
97.85 Berthierite 24
96.30 Arsenopyrite 33
94.20 Arsenopyrite 31
92.84 Berthierite 20
87.61 Awaruite 38
86.38 Awaruite 57
86.02 Berthierite 31
85.37 Awaruite 15
84.97 Arsenopyrite 56
84.31 Awaruite 79
83.73 Berthierite 26
81.44 Awaruite 63
81.40 Arsenopyrite 19
```

Figure 16: The user in the case study was able to identify spikes in Awarite and Arsenopyrite at similar angular positions with TimeSearcher SSE. The first and third graphs show the similar spikes. The material name and range of angular positions are located in the upper left corner of the graph. TimeSearcher SSE was loaded with raw powder X-ray diffraction data for Beritherite, Awarite, and Arsenopyrite, material all containing Fe. A eight point spike ranked according to its angular height was used to identify these spikes. The angular height value is squared in red in the ranking window. This discovery indicates that spikes at this position may be caused by the presence of Fe in the materials.


TimeSearcher: Shape Searcher Edition...Sulfur IR Diffraction Data
File Edit View Transform Help
Multi-Point Angular Height Spike
Points in Shape: 8
Search
HCIL
Details
Definitions
Baryte 979.6611 - 1170.579
Anhydrite 979.6611 - 1170.579
Baryte 593.9678 - 784.886
Anhydrite 593.9678 - 784.886
Anhydrite 401.1211 - 592.0393
Baryte 3679.514 - 3870.433
6485.7
4840.0
3196.3
1551.6
5 11 17 23 29 35 41 47 53 59 65 71 77 83 89 95

```
Term:       Anhydrite 979.6611 - 1170.579
            Count
Sample:     657.7006
0:          675.9217
 1:         693.5526
 2:         716.4571
 3:         743.9905
 4:         773.0845
 5:         801.2294
 6:         831.9454
 7:         862.6735
 8:         895.7434
 9:         930.7802
 10:        961.1641
 11:        991.7856
```

```
99.97 Baryte 979.6611 - 1170.579
99.97 Anhydrite 979.6611 - 1170.579
99.96 Baryte 593.9678 - 784.886
99.95 Anhydrite 593.9678 - 784.886
97.93 Anhydrite 401.1211 - 592.0393
81.03 Baryte 3679.514 - 3870.433
74.43 Baryte 786.8145 - 977.7326
73.06 Baryte 3872.361 - 3999.64
64.90 Anhydrite 3293.821 - 3484.739
64.20 Anhydrite 3872.361 - 3999.64
64.13 Anhydrite 3679.514 - 3870.433
63.55 Baryte 3872.361 - 3999.64
57.02 Anhydrite 401.1211 - 592.0393
55.66 Baryte 3100.974 - 3291.893
48.83 Anhydrite 1943.894 - 2134.813
31.24 Anhydrite 2908.128 - 3099.046
23.75 Anhydrite 3100.974 - 3291.893
```

Figure 17: TimeSearcher SSE was loaded with infrared spectroscopy data for Anhydrite ($CaSO_4$) and Baryte ($BaSO_4$). A user was attempting to identify spikes with similar intensity at the same wavelength. The user was able to do this with a sixteen point spike ranked according to its angular height. The values of the ranking metrics are shown in the ranking window inside the red square.

[11] E. Keogh, H. Hochheiser, and B. Shneiderman. An augmented visual query mechanism for finding patterns in time series data. In *Proc. of the 5th International Conference on Flexible Query Answering Systems*, pages 240–250. Springer, LNAI, 2002.

[12] B. Padmanabhan and A. Tuzhilin. Pattern discovery in temporal databases: A temporal logic approach. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

[13] Rruff project database. http://rruff.info/.

[14] K. Ryall, N. Lesh, H. Miyashita, S. Makino, T. Lanning, T. Lanning, D. Leigh, and D. Leigh. Querylines: approximate query for visual browsing. In *in Extended Abstracts of the Conf. on Human Factors in Computing Systems*, pages 1765–1768. ACM Press, 2005.

[15] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proc. of the IEEE Symposium on Information Visualization*, pages 65–72. IEEE Press, 2004.

[16] M. Wattenberg. Sketching a graph to query a time series database. In *Proc. of the 2001 Conference Human Factors in Computing Systems, Extended Abstracts*, pages 381–382. ACM Press, 2001.

[17] J. Yang, W. Wang, and P. S. Yu. Stamp: Discovery of statistically important pattern repeats in a long sequence. In *Proc. of the 3rd SIAM International Conference on Data Mining*, pages 224–238. SIAM, 2003.