Integrating time-series visualizations within parallel coordinates for exploratory analysis of incident databases

Michael Butkiewicz[a], Thomas Butkiewicz[b], William Ribarsky[b], and Remco Chang[b]
[a]University of Pittsburgh
[b]University of North Carolina at Charlotte

ABSTRACT

Aviation disaster prevention has always been of interest to homeland security, especially after the recent use of aircrafts as weapons by terrorists. With better understanding of the deficiency of different types of aircraft and their corresponding effects on the craft's safety, better maintenance and response plans can be devised to prevent disasters from occurring. In this paper, we present a visual analytical technique to examine the Federal Aviation Agency's Accident/Incident Database, which contains more than 90,000 incidents across 53 dimensions over the last 30 years, for identifying trends of relationships between dimensions over time. Our technique is based on the integration of the ThemeRiver technique directly within a parallel coordinates framework, and simultaneously presents both a "forward flow view" and a "backward flow view" between each dimension. The forward flow view shows the trends over time of each of the elements-of-interest in the first dimension, while the backward flow view illustrates how the elements in the second dimension contribute to the overall trends seen in the first dimension. Through the use of our technique, we were able to identify characteristics of aircrafts and suggest plausible explanations to their common failures.

Keywords: Aviation Disaster Prevention, FAA, Parallel coordinates, time-series, ThemeRiver, incident reports.

1. INTRODUCTION

Understanding the causal relationships between aircraft types, types of failure, and time is important for many reasons. From the perspective of homeland security, a better understanding of what types of aircrafts are prone to certain types of failures can lead to the creation of more accurate and precise prevention and response plans. From a commercial perspective, manufacturers and operators can devise better maintenance plans to address the weaknesses of the aircrafts and possibly improve upon the existing designs. Fortunately, the Federal Aviation Agency (FAA) has been collecting accident/incident reports for the past 30 years. The Accident/Incident Data System (AIDS) contains more than 90,000 records, with each record detailing 53 aspects (dimensions) of the conditions and causes of each incident. Through examination of these records, we propose that we can better understand why and how certain aircrafts fail.

The core of our method for examining the FAA accident/incident database is a visual analytical technique that we have developed. This technique is based on a combination of two well-known visualization methods: parallel coordinates and ThemeRiver. The parallel coordinates method is a technique that depicts relationships between the dimensions in a dataset, while the ThemeRiver technique is designed to reveal the temporal patterns in single dimension. By merging the two methods together, our new technique exceeds the capabilities of both predecessors and is capable of examining relationships between multiple dimensions over time.

Inherent to the traditional parallel coordinates plot are several issues that limit its ability of display easily identifiable patterns and trends. Each plot consists of parallel axes with each representing one dimension of the data. The limitations of this structure arise from the point-to-point method used to map data items across the axes. Each data item in the plot is represented by a single line which connects each adjacent dimensional axis at a point of value. This simplified method ensures that the connection of two axes can at most only represent a two dimensional relationship. In addition, the layering of these simple lines due to the use of large datasets also introduces the issue of occlusion.

Many times it is desirable to view two dimensions with respect to time, a third dimension. This allows the detection of trends of relationships over time which is applicable to many areas. With the current infrastructure this would require a dimensional time axis to be placed directly between the two axes of interest. This insertion not only accentuates the problem of display space limitations for high dimensional data, but also due to the relationship having been stretched out over three adjacent axes this method serves to greatly obscure the intuitive pattern recognition granted by the typically simple connection of two points.

We find interest in the use of ThemeRivers precisely because they do not suffer from the problems and limitations mentioned above. A ThemeRiver provides a macro view of the thematic changes in a dataset, set over a serial dimension. For our use, we assign time to this serial dimension so that we are able to view the trends of relationships over time in a straightforward manner. The strength of representation of these themes is shown by the widening and shrinking of a band of color associated with a particular theme. The task of effectively scaling each theme is made rudimentary since all of the data for a particular theme is encapsulated within one specific band. Furthermore since the ThemeRiver is comprised entirely of these naturally scalable bands, the entire visualization is able to scale to any size regardless of the actual number of data items represented. This inherent scalability makes it an ideal candidate for visualizing trends over time in our 90,000+ record dataset.

We present a method to create a hybrid of these two types by replacing the point-to-point connections found in-between the axes of a traditional parallel coordinates plot with collections of space-filling ThemeRivers. In this we seek to solve the inherent limitations placed upon the parallel coordinates structure by these simple connections. This hybrid technique provides a more natural theme and pattern presentation along with improved occlusion and clutter management. We show our hybrid technique's strength by testing it on a large, highly dimensional dataset in which the successful identification of trends of relationships over time between dimensions is crucial to its understanding.

2. PREVIOUS WORK

The well-known parallel coordinates visualization technique has been used for decades in both information and data visualization fields, and has been thoroughly researched. Seminal papers dealing with the development, refinement, and use of the parallel coordinates technique include Inselberg[4], Wegman[7], and Fua[1]. Parallel coordinates' inherent ability to handle datasets with high dimensionality and the ease of customizing the ordering of dimensions for comparison against each other are the two primary reasons we choose to use it as the basic framework for our visualization.

ThemeRiver is a relatively newer technique that was introduced by Havre[2][3] as a method for illuminating trends, patterns, and relationships (all in regards to keyword occurrence) over time in large collections of documents. ThemeRiver was found to be effective, in that users found it intuitive and understandable, easier to follow due to its connectedness, and useful for identifying major trends. However, a noted shortcoming was difficulty in identifying minor trends, as they became buried under major ones, and thus deemphasized. We sought to harness ThemeRiver's great ability to depict trends in one-dimensional temporal data and integrate it directly within the multi-dimensionality supporting framework of parallel coordinates.

Singular ThemeRiver plots have been added as coordinated visualization to various visual analytics systems to provide a supplemental time-series visualization of sub-sets of data selected in other coordinated frames. A good example of this is shown in Wang's[6] interactive visual analytics system for exploring a large (>60,000 records, ~120 dimensions) geo-referenced event database. Here, a ThemeRiver is placed alongside the rest of the system, in which other tools are provided for the user to filter the records by categorical information. The ThemeRiver presents the patterns of all events that have not been filtered out over the databases time-range. Our system differs from these coordinated visualization implementations of ThemeRiver in that we do not provide a single separate ThemeRiver as a supplemental visualization in support of a system of other visualizations, but instead we provide multiple ThemeRivers as needed, directly integrated within a parallel coordinates structure. As such, filtering is no longer an external event controlling the scope of the ThemeRiver, but an iterative exploratory process supported with intermediate ThemeRivers for immediate feedback. Furthermore, the multiple instances of ThemeRivers in our system each present unique views of the complex temporal relationships between multiple dimensions in the underlying data.

Attempts to improve the ability of a parallel coordinates plot to better handle time-series data include moving from parallel axes aligned linearly in a single dimension across the image, to a layout in which the time axis resides in the center, while the other parallel axes surround the central time axis in a circular pattern. In this way, there is essentially a time axis "between" each pair of parallel axes, but only one actual time axis drawn. Tominski[5] presents this "TimeWheel" arrangement as well as a 3D modification. Parallel coordinates plots are already known to be fairly non-intuitive, and we argue that these complicated arrangements are even more so. They also require changes in viewing angles and thus navigation to fully observe their overlapping (in 2D) structures. We are not aware of any prior work that integrates ThemeRiver plots directly within the framework of a traditional parallel coordinates arrangement as we present here.

3. DATA SET

The FAA Accident/Incident Data System (AIDS) database contains over 90,000 incident reports with as many as 53 dimensions, for all categories of civil aviation from the year 1978 onward. These reports are generated as a result of events associated with the operation of an aircraft that affects or potentially affects the safety of operations but fails to meet the thresholds of personal injury or aircraft damage as set by the National Transportation Safety Board's definition of an accident. This definition of an accident is based on the idea of damage or failure which adversely affects the structural strength or performance of the aircraft, requiring either major repair or replacement of the affected part. Excluded from this definition however are things such as engine failure, punctured holes in the skin, and damage to landing gear, flaps, wheels, brakes and other such accessory components.

While these incidents may be considered minor in perceptual isolation, it is their reoccurrence over a span of time that causes problems for the aeronautic industry. In order to understand why these incidents occur in specific circumstances and more importantly what can be done to prevent their frequent occurrence it is first necessary to identify the patterns behind their occurrences. It becomes clear then, that to find unifying themes of occurrence the incidents must be viewed both in their entirety and in respect to a time span.

Viewing such a large number of data items at once typically overloads the viewing space of a parallel coordinates plot which will not adequately be able to scale given any reasonable amount of screen space and as such the issue of debilitating occlusion arises. With our hybrid technique's use of ThemeRivers and their encapsulating and scalable nature this data load is once again manageable. As we have stated before, traditional parallel coordinates plots struggle to maintain their intelligibility when a temporal relation is extended via dimensional time axes as a unifying element present in the relationships between all other dimensional axes. With the use of the ThemeRiver visualization, which has been specifically designed to allow easy identification of thematic information over a time span, our hybrid technique provides a significant improvement in this area while still maintaining the ridged structure of the underlining parallel coordinates plot to guide the user though the dataset's dimensions.

4. IMPLMENTATION

It was important in the design of our hybrid system that the basic functionality of both the original parallel coordinates plot and the ThemeRiver be preserved. The overlying structure of the parallel coordinates plot allows the user to focus their attention on the relationship between fixed sets of dimensions. While our approach introduces the idea of filtering at each axis to allow these relationships to affect each other in a chaining method, we still provide the basic isolation of each relationship through the division of axes.

4.1 Framework

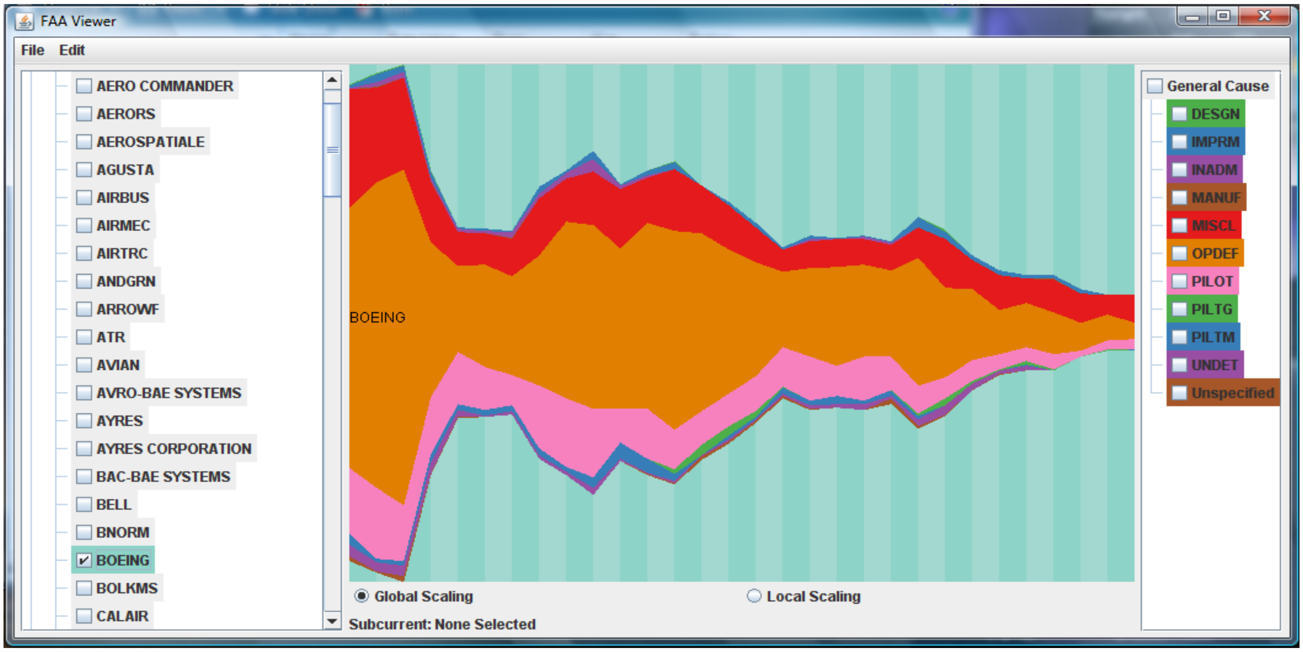
Our hybrid technique's over-arching layout stays in line with the axis driven design of a traditional parallel coordinates plot. Each axis still can represent at most represent one dimension in the data and these axes are still placed parallel to each other with relationships depicted in-between. The major difference in use of these axes comes from their effect on the relational depictions.

Each category in the first dimensional axis traditionally materializes as a separate starting point from which data items branch out of. If handled naively our hybrid technique would in turn require the generation of a separate ThemeRiver for each of these categories. In order to limit this potentially unwanted clutter we allow the user the option of selecting categories-of-interest from a given axis, which in turn serves to limit the number of ThemeRivers needed. We also use the user's choice of categories to enable each axis to act as a filter for the data flowing through it. We can see this in Figure 1 where only Boeing has been selected on the airplane-make dimensional axis and thus any axes beyond will only contain data items involving Boeing aircrafts.

Our motivation behind such filtering methods concerns our end goal of improving the clarity of trends of relationships through an exploratory analysis. These datasets are typically large in nature so as to bolster the strength of correlations at a statistically reliable level. Once general areas of interest are identified, which we will discuss more in Section 4.2, some method is need to isolate these areas for further exploration. The ability to refine the dataset through the selection of categories-of-interest combined with the filtering effect this has on the chaining together of axes provides the user with a powerful method of narrowing the amount of data items to a more meaningful range.

4.2 Data presentation

One common use for a ThemeRiver is to display themes based on variations in the frequency of keywords in a corpus occurring over time. To integrate the ThemeRiver into our parallel coordinates structure we modify this basic idea. Since each ThemeRiver corresponds to a selected category in the first dimension, our corpus becomes the subset of the axis' dataset that contains this category. Our keywords in this case are all of the possible categories from the second dimension. Following this logic we employ a two step rendering process to relate the two axes. The first step is a "forward flow view" that shows the trend over time for the selected category-of-interest in the first dimension. This is then complimented recursively by the "backward flow view", which illustrates how the categories in the second dimension contribute to the overall trend seen from the first dimension. The resulting combination of the two views generated from this process can be seen in Figure 1.



● Global Scaling ○ Local Scaling

Subcurrent: None Selected

Fig. 1. The 'forward flow' consists of the entire ThemeRiver, showing the number of incidents over time for Boeing (the sole selected category in the 1st dimension.) The 'backward flow' consists of the colored currents within the main river, which show the proportions and trends of categories in the 2nd dimension in the overall river of the selected category in the 1st dimension. Shading lines in the background indicate units of time, in this case years.

The connections found between axes in a traditional parallel coordinates plot provide a direct connection for the viewer to follow. With the breaking of these ties in the presentation of the more abstracted ThemeRiver view we are left with the task of making the associations between categories and their relative locations within the ThemeRiver easily perceivable to the viewer. Our solution to this is a two palette coloring system in which the color palette of the axis to the left of a ThemeRiver will provide its background colors and the palette of the axis to the right will determine the foreground colors. This ensures the user can easily distinguish from which dimensions the ThemeRivers and the currents they contain originate from. See Figure 2 for an example of the color palette swapping process.

One aspect of the ThemeRivers' presentation which turns out to be extremely helpful in aiding the identification of areas of interest with in the dataset is the manner in which the collective ThemeRivers in-between a set of axes are scaled in relation to each other. Our findings show that there are two scaling methods which provide distinct and invaluable perceptual information. The first approach is to scale each ThemeRiver with an equal height in what we call the local scaling method. The second is to scale each ThemeRiver's height in direct proportion to the maximum amount of data items expressed in that ThemeRiver at any one. To give an example of this second method, which we will call the global scaling method, if we compared the aeronautic giant Boeing with a company we will call X, which only produces several aircraft every year, we would expect that the number of incident reports filed for Boeing aircrafts would dwarf the number of the lesser known company based solely on number of planes in operation. Since the global scaling method takes this discrepancy into account, the ThemeRiver for Boeing would likewise dwarf company X's ThemeRiver in display space. This global scaling method is useful because it intuitively presents the differences in volume/frequency between multiple categories, but can lead to trends in minor categories being suppressed. The local scaling option can be used to overcome this shortcoming by dividing the available screen space uniformly. In Sections 4.2.1 and 4.2.2 we further discuss the strengths of each of these types of scaling

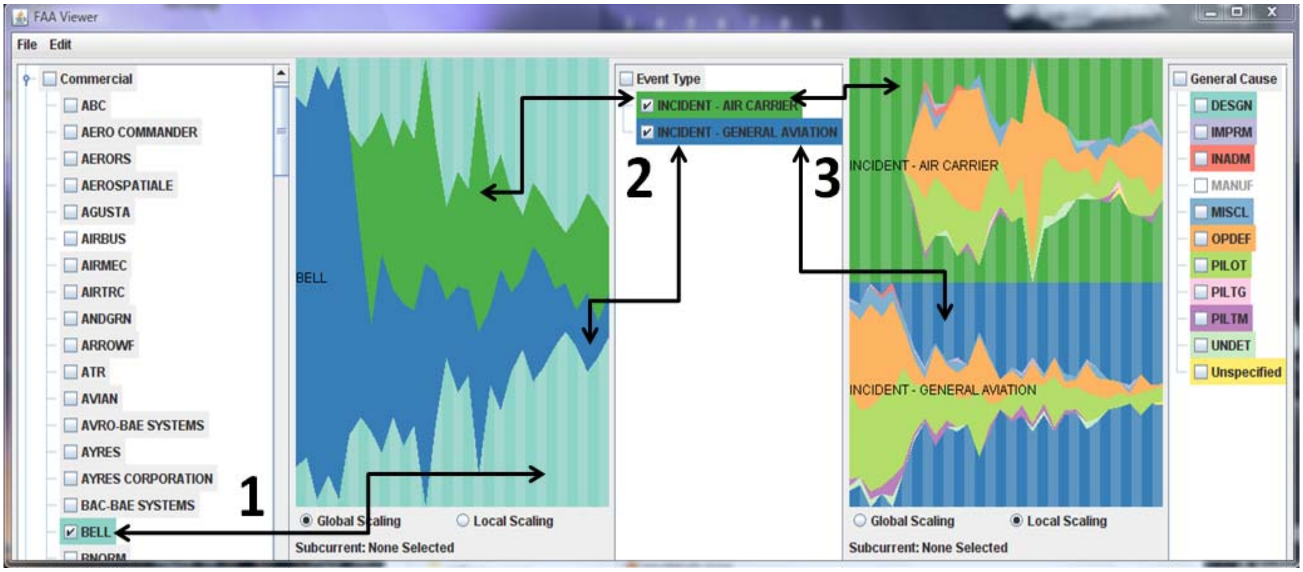


Fig. 2. Step 1) The color of the selected category in the 1st dimension is used as the background for its ThemeRiver. Step 2) The sub-currents found inside this ThemeRiver are then colored according to the categories in the 2nd dimension to which they correspond. Step 3) Note that the green "Air Carrier" sub-current found in the 1st display now has its own ThemeRiver in the 2nd display with the same green as its background. Thus the two palettes have swapped, foreground to background and vice versa.

4.2.1 Forward Flow View

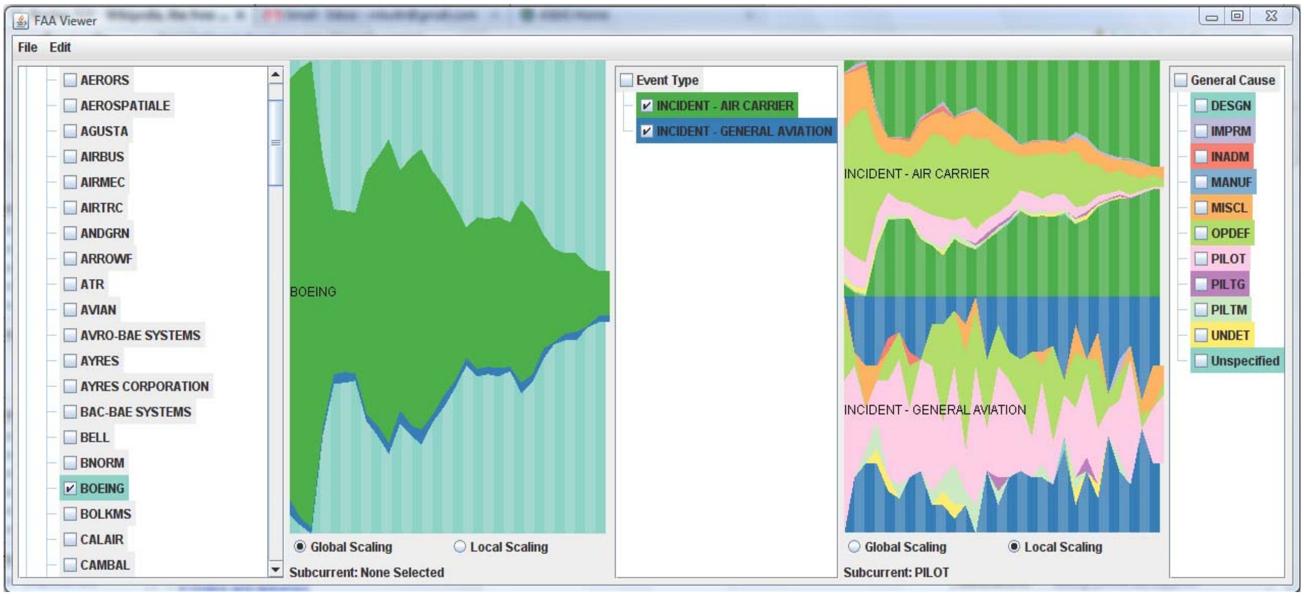
This view is best thought of as the two dimensional relationship between the selected category in the first axis and time. In this sense it displays much of the same information as if we had placed the first axis next to the time dimensional axis in a traditional parallel coordinal plot. To leave it at this, however, would to suggest little improvement with our new system. One advantage to the use of a ThemeRiver type display is the interpolation between the basic time units which provides addition relational information for the events of one year to the next. When we combine the forward flow view with the local scaling option we see its true power for the identification of areas of interest in the first dimensional axis. We will extend our previous example of Boeing and company X to include the ThemeRivers of tens or hundreds of selected categories. Our hybrid technique will allocate screen space based on category strength and as such the strongest (most influence in dataset) and weakest (least influence in dataset) categories become readily apparent.

4.2.2 Backward Flow View

The encapsulation of the data items as they pass from one axis to the next granted by this view is one of the most critical steps in our stride towards ensuring clarity of trends in these relationships. With this clear separation, the distinct features of each theme become readily apparent, and identifying them is facilitated. When used in conjunction with the local scaling option, this view provides an easy way to compare the compositions of multiple ThemeRivers rather than the specific quantity of their data items. An excellent example of this type of usage is our FAA AIDS dataset example seen in Figure 3. Here we are able to readily identify an interesting relationship in the data from simple observation of the composition of each ThemeRiver, without having to go in and inspect the specific number of data items in each.

5. DISCUSSION

In Figure 3 we illustrate a simple example scenario showing how results can be found using the AIDS database. Looking at the first ThemeRiver (left) in Figure 3 we see that Boeing aircrafts have almost always been used for commercial air carrier flight types as opposed to private general aviation flight types. Moving on to the second set of ThemeRivers (right) we find an interesting ratio in the data between incidents attributed to pilot error and those attributed to other persons involved in the operation (known as OPDEF). In the air carrier ThemeRiver we can observe that the ratio is heavily sided with the OPDEF (green) as a general cause while in the general aviation ThemeRiver we see the ratio alarming turn in its majority to the pilot (pink) sub-current. One conjecture to make from this is that the pilots flying Boeing aircraft for personal uses (which is assumed to fall under general aviation) simply do not have comparable training to a pilot flying it for its intended use as an air carrier flight type (assumed to be more commercially based.) We are not domain experts and further investigation is of course necessary, but this may turn out to be an area for stricter regulations by the FAA for increased training hours and programs for these types of pilots.



☑ BOEING

● Global Scaling ○ Local Scaling

Subcurrent: None Selected

☑ INCIDENT - AIR CARRIER

☑ INCIDENT - GENERAL AVIATION

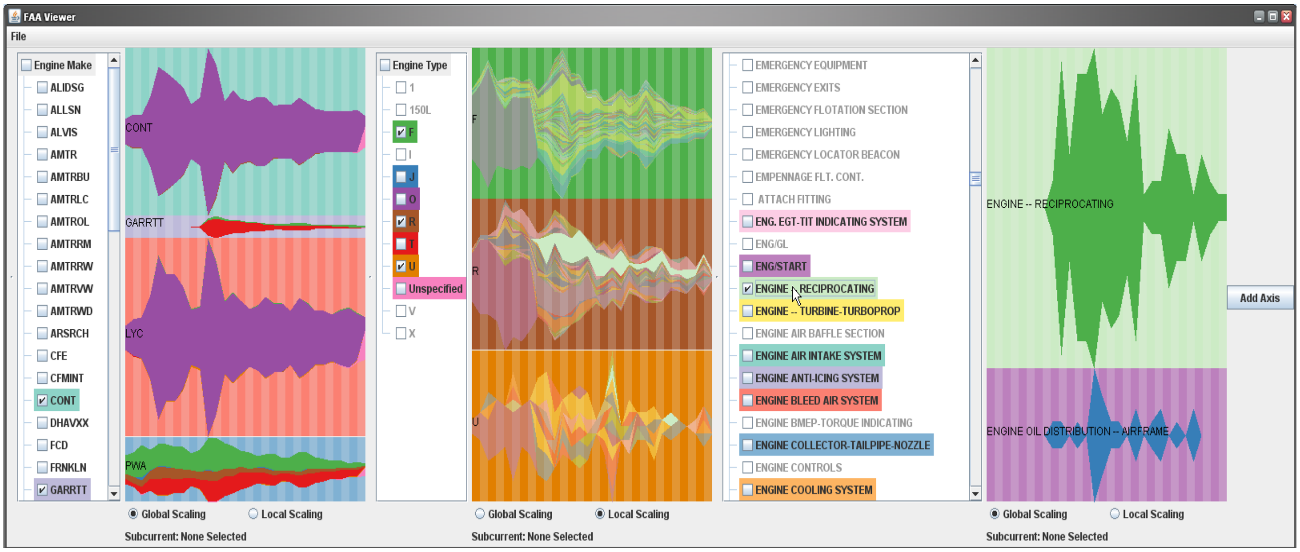
○ Global Scaling ● Local Scaling

Subcurrent: PILOT

Fig. 3. Note the difference in ratio of OPDEF (green) error to pilot error (pink) between the second set of ThemeRivers. Also note that the radio button for scaling has been turned to local scaling. This is to better allow the comparison of the percentages of the sub-currents. If the scaling had been turned to global, the air carrier ThemeRiver would take up almost all of the screen space, dwarfing the general aviation ThemeRiver.

While ThemeRivers have solved the problem of occlusion in parallel coordinates, this hybrid technique may still suffer with the effects of crowding. This problem is often seen in the use of local scaling when the number of ThemeRivers or sub-currents within exceed a certain limit and force the scaling to reduce them to an unusable range. While with global scaling this often serves to eliminate uninteresting portions of data from the user's attention there is still the case where a user may not be able to find the data he is looking for. To aid in this detection we provide the highlighting of sub-currents upon mouse over of the name of the category in its axis (see Figure 5) as well as name presentation upon selection of a sub-current directly inside of a ThemeRiver.

Because our hybrid technique is primarily focused on theme identification in datasets where many of the dimensions may have an effect on one another, we see a wide range of application areas. It is clear that our approach, which works well on the AIDS database, could easily be extended to other types of mass transit as they face very similar issues of maintenance and repair. Branching out further, we see many other scenarios for use in areas of security and defense. A look into various trends of crime could be explored through a database of police reports. Comparing dimensions such as re-education efforts, location of occurrence, number of repeat offences, and type of crime may allow the user to explore the effects of re-education on various populations of criminals. Using this as a guide one may then be able to determine which types of criminals respond best to these efforts, and which efforts are most effective overall.



Engine Make: ☑ CONT, ☑ GARRTT

Engine Type: ☑ F, ☑ R, ☑ U

☑ ENGINE -- RECIPROCATING

● Global Scaling ○ Local Scaling

Subcurrent: None Selected

○ Global Scaling ● Local Scaling

Subcurrent: None Selected

● Global Scaling ○ Local Scaling

Subcurrent: None Selected

Fig. 5. Here we see a more complicated use of the backwards flow view in the middle set of ThemeRivers. The highlighting method has been used here by holding the mouse over the “Engine – Reciprocating” category in the primary cause (third) axis. This highlighting allows us to look through the crowding caused by the abundant sub-currents to see the disparity between the percentage (remember we are using local scaling, so we are looking at percentages instead of raw quantity) of incidents cause by “Engine – Reciprocating” in the three different engine types. As shown here, this is a major primary cause for failure in engine type R, where as types F and U show little problem.

6. FUTURE WORK

Due to the encapsulation of the data items within each separate ThemeRiver, our current presentation looses the ability to step back and view in one take the entire span of the relationship from one axis to the next originally given by the all inclusive nature of a traditional parallel coordinates plot. To provide this wide, over-reaching view of the general relationship between the two axes we plan to allow the joining of several categories together to make one composite ThemeRiver. This would then be used to either create groups of categories, or if the user desires to see the total relationship, may be used to group all categories into one all encompassing ThemeRiver. This option not only provides the user with yet another tool to display the data as they see best, but also allows them to mitigate the clutter resulting from many separate and possibly redundant ThemeRivers.

Because the theme recognition aspect of a ThemeRiver is driven by keyword occurrences, the problem of how to display continuous data and free text must be addressed. In order to limit the categories for display with continuous data, the user, who is assumed to have domain knowledge, will be able to set relevant intervals for these dimensions to be used as categories. In the AIDS dataset a free-text narrative dimension is given by those involved and describes the incident along with possible causes. This dimension of the data is therefore extremely helpful in not only in identification of the cause of specific occurrences, but could also serve as further validation of results obtained from other dimensions. To take advantage of the possible validation granted by this use, we intend to allow the user to coordinate the dataflow with the free-text analysis program INSPIRE. In this way the user could export their refined data into the program and take advantage of INSPIRE’s keyword analysis and data item clustering into order to explore the full narratives in their themes-of-interest.

7. CONCLUSIONS

Our method of integrating the ThemeRiver into a parallel coordinates system attempts to solve some long standing problems the visualization community has typically struggled with. The task of providing supplemental temporal information to each two dimensional relational depiction in a traditional parallel coordinates system requires means which invariably complicate the results to the point of unintelligibility. Further confounding the relational depiction is the oclusion caused by an inability to properly scale to large datasets.

With their easy to understand metaphor of one main stream comprised of several distinct currents, the ThemeRiver was a natural choice for lending clarity in the matter of relational trend depiction. Its forward flowing property incorporates the aspect of time which is critical in the ability to track changes in a given relation. The encapsulation provided by its depiction of sub-currents eliminates the previous issues of occlusion caused by the overlapping data flow. This encapsulation however can lead to the crowding of ThemeRivers and currents within. While the former is only partially mitigated by the use of scaling and axis filtering, the latter is effectively solved with current highlighting and direct current selection for name retrieval.

In our work towards creating an easy and intuitive tool for exploratory analysis we provide two distinct scaling options to allow the user to identify areas-of-interest in the dataset. As areas-of-interest are discovered, additional axes can be used as a powerful method for narrowing down the dataset to a more meaningful range. The clear depictions of themes and patterns provide the user with the information necessary to draw conclusions, which can be supported by relationships drawn directly from the underlying dataset.

REFERENCES

[1] Fua, Y.-H., Ward, M. O., and Rundensteiner, E. A., “Hierarchical parallel coordinates for exploration of large datasets,” in Proceedings of IEEE Visualization ‘99, pp 43-50, 1999.

[2] Havre, S., Hetzler, B., and Nowell, L., “ThemeRiver: visualizing theme changes over time,” in Proceedings of IEEE Symposium on Information Visualization 2000 (InfoVis), pp. 115-123, 2000.

[3] Havre, S., Hetzler, E., Whitney, P., and Nowell, L., “ThemeRiver: visualizing thematic changes in large document collections,” IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 9-20, Jan/Mar 2002.

[4] Inselberg, A., “The Plane with Parallel Coordinates,” The Visual Computer, vol. 1, no. 1, pp. 69-91, August 1985

[5] Tominski, C., Abello, J., and Schumann, H., “Axes-based visualizations with radial layouts,” SAC ’04: Proceedings of the 2004 ACM Symposium on Applied Computing, ACM, pp. 1242-1247, 2004.

[6] Wang, X., Miller, E., Smarick, K., Ribarsky, W., and Chang, R., “Investigative Visual Analysis of Global Terrorism,” Computer Graphics Forum, vol. 27, no. 3, pp. 919-926, 2008.

[7] Wegman, Edward J., “Hyperdimensional Data Analysis Using Parallel Coordinates,” Journal of American Statistical Association, vol. 85, pp. 664-675, 1990.