**TITLE**

**A thesis submitted in partial fulfillment of the requirements for the degree**

**MASTER OF SCHENCE**

**in**

**DATA SCIENCE AND ANALYTICS**

**by**

**THOMAS CANNON**

**AUGUST 2020**

**at**

**THE GRADUATE SCHOOL OF THE UNIVERSITY OF CHARLESTON, SOUTH CAROLINA AT THE COLLEGE OF CHARLESTON**

**Approved by:**

**ABSTRACT**

**TITLE**

**A thesis submitted in partial fulfillment of the requirements for the degree**

**MASTER OF SCHENCE**

**in**

**DATA SCIENCE AND ANALYTICS**

**by**

**THOMAS CANNON**

**AUGUST 2020**

**at**

**THE GRADUATE SCHOOL OF THE UNIVERSITY OF CHARLESTON, SOUTH CAROLINA AT THE COLLEGE OF CHARLESTON**

We present a new method for categorizing Gamma-Ray Burst (GRB’s) emission episodes with similar light curves from the Burst and Transient Source Experiment (BATSE), onboard NASA’s Compton Gamma-Ray Observatory (CGRO). We compare normalized time-series data from any two respective GRBs’ 64ms light curves using several statistical tests. The comparisons are used in the construction of similarity matrices as input in a hierarchical clustering algorithm. With the new application of this data mining tool, we begin to see similar GRB light curves cluster together by emission properties that exist independent of their amplitude and time scales, leading to a unique understanding of GRB physics.

ACKNOWLEDGEMENTS

Test

**CONTENTS**

[LIST OF FIGURES 4](#_Toc45571593)

[TITLE 5](#_Toc45571594)

[Introduction 5](#_Toc45571595)

[Data 5](#_Toc45571596)

[Statistical Tests Description 5](#_Toc45571597)

[Agglomerative Hierarchical Clustering Description 5](#_Toc45571598)

[Methods 5](#_Toc45571599)

[Results 5](#_Toc45571600)

[Conclusion 5](#_Toc45571601)

[Future Work 5](#_Toc45571602)

[REFERENCES 6](#_Toc45571603)

[FIGURES 7](#_Toc45571604)

[APPENDICES 8](#_Toc45571605)

LIST OF FIGURES

Test

TITLE

## **Introduction**

GRB emissions, while their complexities range, have a defined, non-random structure. However, the cause of GRB prompt emission is still highly debated. GRBs are made of pulsed radiation, and in recent years, have been studied thoroughly to give a greater understanding of the physics behind a GRB event. The properties of these pulses have been measured \cite[]{gol83,nor96,lai96,ram00,nor02} and can be fitted by a four-parameter empirical model \cite[]{nor05}. The Norris model allows for extracting the shape of a single pulse to several overlapping pulses in a GRB lightcurve \cite[]{nor05,hak08,hak09}. Unfortunately, when a burst is made up of a large number of overlapping pulses, it becomes difficult to differentiate between said pulses and accurately understand the processes of the GRB event \cite[]{hak09}.

Many GRBs are classified into two different categories: Long and Short \cite[]{kou93,muk98,hak03}. These categories were selected mainly on the basis of duration. However, recent work has shown that similar correlative pulse properties not only exist in both Long and Short bursts \cite[]{hak09}, but that the Long and Short bursts share common trends of these different property correlations \cite[]{hak11}. This suggests that the Long and Short bursts originate from similar physical processes.

Because there seems to be no discernible difference in the way that GRB light curves are formed, we develop a method to compare every GRB against every other GRB. This is done by a two-sampled $\chi^2$ test \cite[]{pre92} and what is known as an Agglomerative Hierarchical Clustering analysis \cite[]{jai99}.

## **Data**

The data used for this investigation is archived GRB time-series data that was collected by BATSE on board the CRGO\footnote{See http://www.batse.msfc.nasa.gov/batse/grb/catalog/current/}. The time resolution is of both 256ms and 64ms per burst. The BATSE experiment collected data at 256ms until a specified trigger criteria was met. Once met, BATSE would then record the remainder of the burst in 64ms time resolution. Because of this, some GRB samples in our working dataset will have multiple time resolutions along their lightcurves.

## **Statistical Tests Description**

Tests

## **Agglomerative Hierarchical Clustering Description**

An Agglomerative Hierarchical Clustering algorithm works from the ``bottom up"; meaning it begins with every element in its own cluster. As described by Jain, et. al. \cite[]{jai99}, and modified to represent GRB light curves, the algorithm proceeds as follows:

\item With a matrix of comparative 'similarity distances' between every GRB to every other GRB, each distance is initially treated as its own cluster.

\item The two clusters considered the most similar out of the entire matrix are then merged into one cluster. The matrix is then updated to reflect the merger by considering new position values that represent the combined cluster.

\item If every GRB is within the same cluster, then stop. Otherwise, return to step $2$.

Fortunately, there are several programming languages that will take matrix inputs and return clustering results. In our case, we used an IDL routine called CLUSTER\\_TREE.\footnote{ For the IDL clustering routine used see \url{http://www.exelisvis.com/docs/cluster\\_tree.html}}\footnote{For the analogous MATLAB routine see \url{http://www.mathworks.com/help/stats/linkage.html}}\footnote{For the analogous Python Scipy routine see \url{http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html}}

## **Methods**

Before the 64ms light curve data could be evaluated with the two-sampled $\chi^2$ test, it needed to appropriately prepared. Under every BATSE GRB detection there is considerable background of gamma-ray flux. This background is often changing due to the CGRO's orbit in and out of areas of higher and lower gamma-ray radiation around earth. This sometimes gives the background a slope, which needed to be subtracted out. In order to subtract out the background, we begin by defining the area of the background as everywhere outside the middle $90\%$ of the GRBs' fluence. The times representing this bounded fluence are called t90 times, and were selected for convenience from a table of published values\footnote{see \url{http://www.batse.msfc.nasa.gov/batse/grb/catalog/current/tables/duration\\_table.txt}}. An additional $10\%$ of the t90 times were added to either end to ensure that the total fluence of the GRB was accounted for. This area, which represents the burst, was then removed from the GRB light curve to leave only recorded background data of the times surrounding the GRB event. This background was then linearly fit, and the slope of the fit was then subtracted from the light curve, leaving a flattened background.

The burst event regions were then also selected by their t90 times. This time, there was no additional time added because we operated under the assumption that similar bursts would have similar t90 times and therefore still evaluate well under the $\chi^2$ test. These regions were then scaled to a common fiducial time scale.

It is from here that we proceed with gathering 'similarity distances' in the form of $\chi^2$ values. With these values in a single document, we run them through the clustering algorithm to produce the relationships of every GRB to every other GRB.

## **Results**

With the clustering completed, we move the data into a dendrogram for easy visualization. With our large dataset, the dendrogram becomes too large to visualize on a single page. Therefore, as seen in Fig. 1, we sample a part of the dendrogram for preliminary results discussion. The dendrogram visualizes different GRB light curves as leaves in branches. Each branch represents a cluster of GRBs that have similar characteristics as seen in Fig. 2. The vertical length of a branch is a measure of how similar a GRB or cluster of GRBs is to its connecting GRB or cluster of GRBs. Sometimes, the tree will display several adjacent GRBs in a tight cluster that show evidence for an evolving continuum of GRB properties as seen in Fig. 3. Also, There are some GRBs that are almost completely unique. These are displayed as branches on the tree whose nodes, common to the rest of the tree, break off very high up. In the algorithm, these GRBs would have been selected last as a comparable relative to any other GRB or cluster of GRBs. Curiously, these GRBs also have an incredibly high signal to noise.

## **Conclusion**

We are still only beginning to discover what this data mining technique is teaching us about GRB light curves. Currently, this technique allows us to compare GRB light curves in a new and interesting way. By using a two-sampled $\chi^2$ test results as `similarity distances' in a Agglomerative Hierarchical Clustering Algorithm, we are able to achieve our comparisons. This technique seems to be good at finding GRBs that evolve similarly and possibly have similar progenitors. However, with a large amount of new data to investigate, we are still determining the meaning of the results of the analysis. We are also working on ways to improve and speed up the computationally expensive process.

## **Future Work**

While the method is sound the data preparation could be improved upon. A better definition of the burst start and end times would work nicely. Also, a more rigorous analysis needs to be done with the results. We have developed a new tool with which to explore GRB physics. However, in the larger scheme of things, this method does not have to be used simply for GRB light curves. It is easily adaptable to any transient or two dimensional data.

REFERENCES

Test

FIGURES

Test

APPENDICES

Test