Data set report (second exercise)

Justinas Lekavičius

March 23, 2022

Abstract

This is the second report of the chosen multidimensional data set "Bitcoin Heist Ransomware Address Dataset" for the Multidimensional Data Visualization second exercise. The report contains usage of direct visualisation methods, providing comments, conclusions and insights on the analyzed data.

Contents

---

1 The chosen data set and visualisation methods

This section contains the brief introduction (recap) of the selected data set, and the direct visualisation methods chosen for the exercise.

1.1 Data set

The chosen data set is called "Bitcoin Heist Ransomware Address Dataset", accessed via UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets/BitcoinHeistRansomwareAddressDataset`) and used for the first exercise. The data set features multivariate and time-series characteristics with Bitcoin transaction data. It contains Bitcoin network addresses and their features, that are used for ransomware payment identification. The selected data set can be used to detect specific transaction patterns and potential ransomware payments, and direct visualisation methods may help bring more insights and conclusions.

1.2 Direct visualisation methods

For multidimensional data set visualisation, Orange program was used. The methods scatter plot, linear projection and Radviz were selected for data set visualisation.

2 Multidimensional data set visualisation

This section covers the procedure for getting the data set ready for visualisation, as well as the images produced with various visualisation methods. Observations and insights are also provided.

2.1 Preparation of data set

The data set is imported into Orange using the CSV File Import module, as the data set is a .csv file. However, the CSV file contains 2916697 rows, i.e., objects. Due to this, the performance of the program is degraded severely, and the decision was made to use a selected array of data for visualisation. Firstly, the Select Rows module was used to pick only the objects that are not "white" (not associated with ransomware) and leave only ransomware-related objects, bringing the number of instances down to 41413. It could be argued that the number of objects is still large, however, the variety of ransomware classes was also taken into account. Then, Data Sampler was used to randomly pick 50 percent of data (fixed proportion of data), resulting in 20707 instances. Finally, Preprocess was used to normalize the data to [0, 1] interval.

2.2 Visualisation using scatter plot

Firstly, scatter plot was used to visualise the data set with different x and y projections.

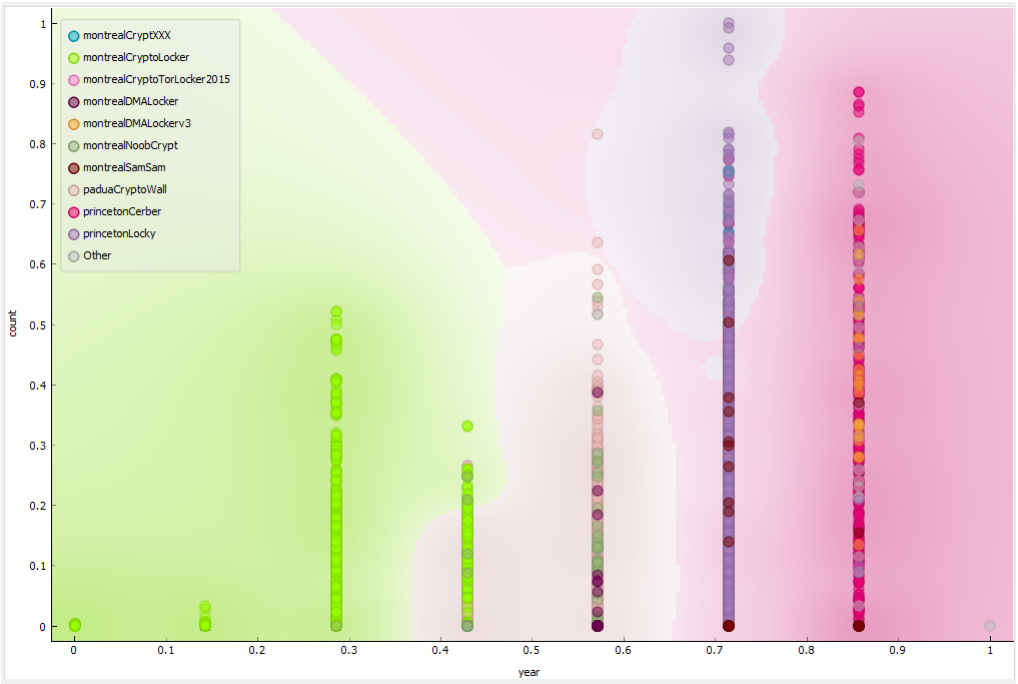

Figure 1: Using the scatter plot visualisation method to display the variety of ransomware attacks.

For example, using the x "year" and y "count" axes, we can visibly see the variety of ransomware attacks throughout the years (spanning from 2009 to 2018). CryptoLocker ransomware attacks were popular in the earlier point of the timeline, while more varying attacks such as CryptoWall, Locky, Cerber and DMALockerv3 appear later in the timeline, likely coinciding with the increasing popularity of cryptocurrency transactions. Furthermore, it is visible by the y axis (count) that the amount of starter transactions connected to address through a chain tends to increase as well. The different types of ransomware are indicated by different color dots, with the legend in upper left corner. Color regions also indicate the dominance of certain ransomware classes.

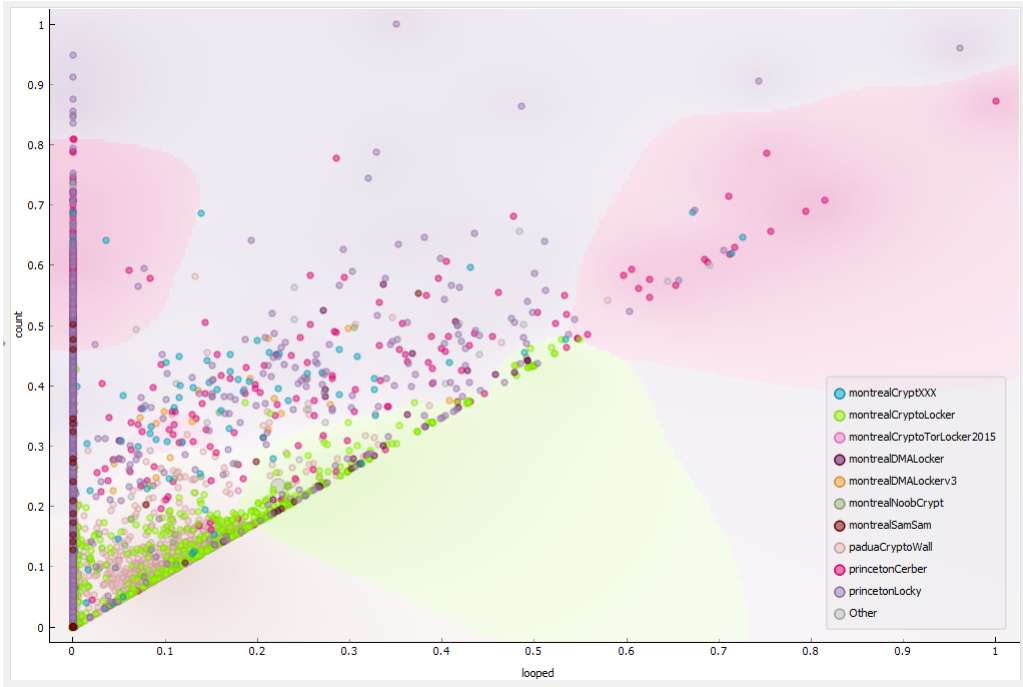

Figure 2: Using the scatter plot visualisation method to display the correlation between "looped" and "count" attributes.

Moreover, using the x "looped" and y "count" axes, we can see positive correlation between the "looped" and "count" attributes. The positive correlation is indicated by increase of "looped" values, along with "count" values (increase from left to right). In other words, the amount of starter transactions that are connected to address through a chain increases with the amount of transactions that are connected to address in more than one directed path.

2.3 Visualisation using linear projection

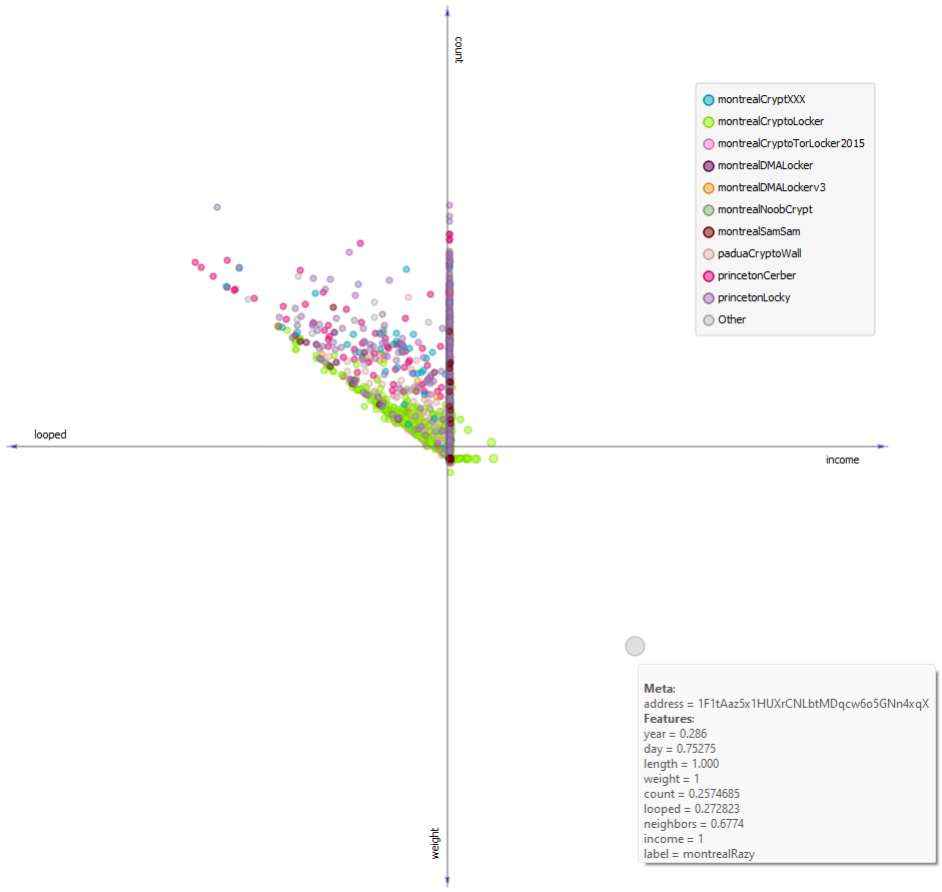

Figure 3: Using the linear projection method to display several features.

The linear projection visualisation method was used to display the objects' several features, and the visualisation method can be considered a better alternative to scatter plot. In Figure 3 for example, not only the positive correlation between the amount of starter transactions connected to address through a chain and the amount of transactions connected to address in more than one directed path is visible, but also outliers can be displayed more prominently. As an example, an outlier is visible in the lower right side of the figure, indicating an object of "montrealRazy" ransomware type and extremely large weight (fraction of Satoshis which originate from a starter transaction) and income (amount of Satoshis output to address). Both have maximum values of 1 (because of normalization to range [0,1]).

2.4 Visualisation using Radviz

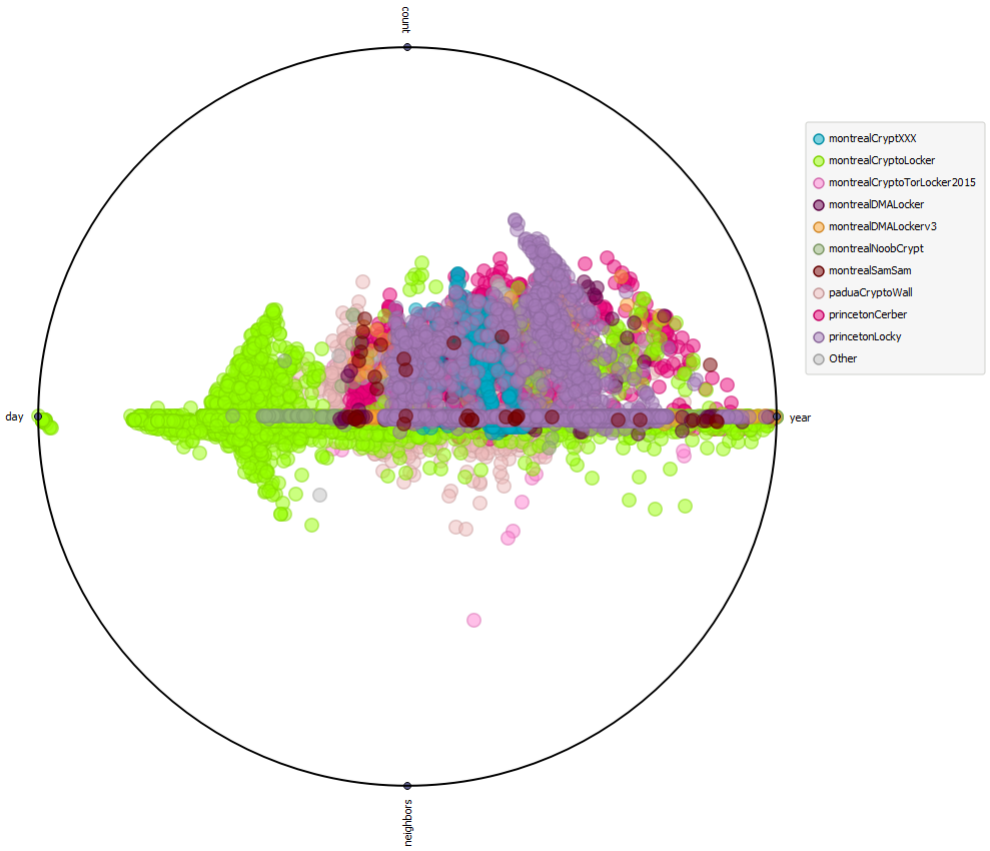

Figure 4: Using the linear projection method to display several features.

Finally, The Radviz visualisation method was utilised to display the data set object characteristics. "year", "count", "day" and "neighbors" features were used to reveal more information not seen using other visualisation methods. For instance, one of the more prominent ransomware classes "montrealCryptoLocker" was more rampant in the earlier point of the timeline (2009 - 2018), and displayed the characteristic of having a large number of transactions which have the specific address as one of its output addresses. However, the "count" attribute tended to have lower values, indicating the lower chance of the address having connections to starter transactions through a chain. In contrast, the "princetonLocky" ransomware class transactions (which are more dominant in the second half of the timeline, displayed in bright violet dots) tended to have lower "neighbors" values, indicating low number of transactions which have the

address as an output address. Nevertheless, the "count" values are generally higher. This method of visualisation is suitable for displaying several features at the same time and works generally well, however, was not perfectly suitable for this data set, due to inability to show categorical variables with more than two values.

3 Conclusions

To sum up, various visualisation methods may be used to get different insights and conclusions using the provided data set. Scatter plot can be used to easily detect correlation (and its type) between two variables, however is limited by only two axes. The Linear Projection and Radviz visualisation methods have the advantage of being able to visualise more than two features of the same data set, enabling the ability to detect more insights. Therefore, the scatter plot visualisation method is more suitable to detect correlations and relations between at the most two variables, while Linear Projection or Radviz can be used to display relations between several variables, and also detect outliers easily. However, it should be noted that the features should be selected sensibly, otherwise data that is not useful may be displayed. Furthermore, the best visualisation methods also depends on the data set itself, for example, Radviz may not show categorical variables with more than two values, as was the case with the data set selected for this exercise. Thus, several direct visualisation methods should be tested, and then the most suitable one may be selected.