Visual Analytics for Open Source Intelligence

Rodrigo Alves Carvalho

Federal Police Departament - Brazil and
University of Oxford - England

⤳✳⤲

ABSTRACT

Enhanced data visualisation technologies can contribute decisively to companies that rely on making sense of a great amount of information as their primary business objective. Graphical techniques often employed to present processed information and findings to the final customer, such as charts and scatter plots, could be improved and customized to aid analysts during the investigation phase. After examining the internal analysis tasks and data storage strategy from a security consultancy company, this work suggests an approach to enhance knowledge extraction, insight generation and hypothesis testing by applying novel techniques of interactive data visualisation and mining.

KEYWORDS: Visual Analytics, Data Mining, Open source tools.

1. Introduction

The Internet has certainly expanded information-sharing capabilities between individuals, companies and governments. Besides highly improving some services that would otherwise require manual interaction or live presence (e.g. transferring funds between accounts, shopping for groceries and more), its indexing technologies also enabled anyone to browse the Web to gain knowledge about virtually any topic in a timely manner. Even though its reliability must be assessed, the fact is that information is available, possibly from multiple sources.

In addition, the development of more intuitive content posting tools such as blogs and social media networks made possible for users from all backgrounds to disclose information on the Web, thus democratizing online content generation. Without regard to the potential benefits and harms

this possibility might bring, it contributed decisively to the great amount of data stored on the Internet today. Such diversity in available data is a great opportunity for any entity looking to extract knowledge from it.

1.1 VISUAL ANALYTICS

Visual Analytics is rapidly becoming a crucial technology to many fields of knowledge. The Institute of Electrical and Electronics Engineers (IEEE) recognises its interdisciplinary nature by holding a conference that "...includes both fundamental research contributions within visual analytics as well as applications of visual analytics in science, engineering, business and commerce, medicine and healthcare, media and social interaction, arts and humanities, public safety, logistics, and other disciplines." (IEEE, 2014).

Its importance resides in the fact that, in contrast to traditional text-based search methods like Structured Query Language (SQL) queries and spreadsheets filters, and also to statical, low dimensional visualisation techniques (e.g. line graphs), "Visual representations and interaction technologies give users a gateway into their data, letting them see and understand large volumes of information at once. To facilitate analytical reasoning, visual analytics builds on the human minds ability to understand complex information visually." (THOMAS; COOK, 2006).

1.2. OVERVIEW OF THE COMPANY

This work will analyse the working processes performed by a security consultancy company which claims to conduct its investigations in an ethical manner, avoiding grey areas such as covert operations, active monitoring and privacy settings bypassing. Therefore, this company is highly dependable on the knowledge obtained from open sources on the Internet and, similarly to other firms, the amount of available working information is surpassing its analysis capability.

A first inquiry on how the employees carry out their tasks reveals that intelligence information is being collected from the web just as it would be from newspapers: searching through many sources, selecting the most relevant ones, and filling in a spreadsheet with their content, metadata and

internal classification attributes. Moreover, client final reports and internal intelligence documents are being produced as text files, with location data being visually inserted into an attached map.

1.2.1. Internal Working Processes

There are two main groups considered in this study:

The collection team cares about the direction of the investigation, or what is the subject being searched for, and the collection of data, harvesting as much relevant information as possible. Animal Rights is the generic topic been constantly monitored from open source sites in the Internet, without prejudice to more focused searches, depending on some client or strategy requisites.

Relevant information include scheduled events (e.g. date, location, type of campaign, targets), headlines related to big laboratories and court decisions on related entities, whether companies or animal rights defendants. Its sources range from online newspapers to social media (e.g. Twitter, Facebook) and blogs.

The analysis team cares about processing, finding links between previously collected information, and dissemination, effectively producing the intelligence reports. They also hold regular meetings with the clients, in order to better understand their requirements and thus inform the collection team which direction to take.

1.2.2. Information Flow and Storage

The first task performed every day is searching the Internet for relevant information and producing the Daily Update report, in the format of a text email to be sent to the clients early in the morning. Hereupon, the Collection team fills the Source Table with the metadata (e.g. URL, date and time) and the content of the information gathered.

At this stage, the Analysis team already has enough information to work with. Consulting both the Daily Update report and the Source Table,

they fill up their own Report Table, essential to produce more thorough documents such as Weekly Reports and Client-tailored statements. This table describes events, comprising attributes like "Incident Type", "Target company" and "Group involved".

Finally, the Statistics Table merges rows from the Source and Report Tables, and serves both as a consolidated data repository and an information source for plotting graphs normally included in Quarterly Reports and Risk Assessment documents. Figure 1 illustrates in a simplified diagram the work process between both groups, and also the data sources they interact with.
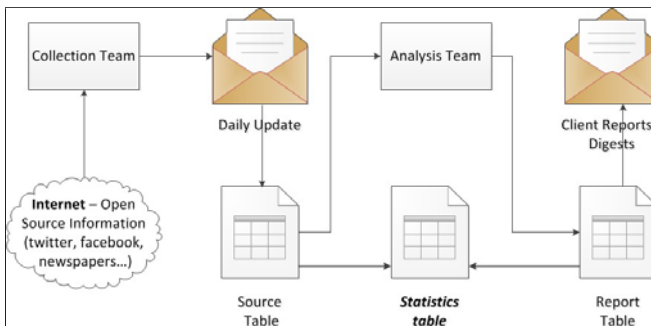

Figure 1. Simplified company workflow.

1.3. OBJECTIVE

The ever increasing amount of information analysts have to cope with extends the necessary time to produce an intelligence report, and might also cause relevant information to be missed. In addition, having different data sources (i.e., distinct spreadsheets storing redundant information) increases maintenance complexity.

Even when applying techniques such as spreadsheet filtering and indexed searching, there may still be a large amount of returned results, possibly containing multiple attributes. Finding relationships among this raw, text-based information is an inefficient duty. Furthermore, spreadsheets filters hamper insight generation and causality discovery by not providing an intuitive interaction capability based on previously returned results.

Based on the initial assessment of the internal working processes carried out by the target company, the goal of this work is to research novel Visual Analytics technologies and suggest some that could make data exploration in the Statistics Table more manageable and intuitive, enhancing its visualisation and consequently insight generation.

2. Technologies Survey

How to make an informed decision? Intuitively, more available data leads to better judgments. However, too much information might also hamper the analysis. So, what information is really necessary, and how should it be made available?

Software that reduce the bulk of data to be analysed without losing the relevant knowledge within are as crucial as novel ways of displaying and interacting with information that merge human cognition and computer processing capabilities. This section suggests some of the available technologies.

2.1. MULTIVARIATE DATA ANALYSIS

Because of the intrinsic complexity of the data stored in the Internet, multivariate data analysis (statistics techniques used to examine data represented by more than one variable), are necessary to convert it into knowledge (HAIR JR. *et. al.*, 2013), aiding relationship finding within high dimensional datasets.

There are two main ways to represent information. Metric data refers to numeric values among which arithmetic operations can be performed. However, some previous analysis and transformation might be necessary: after all, its dimensions might implement different ranges (e.g. a variable that stores values ranging from 200 to 500 might have a greater impact on the analysis results than one whose values vary from 0 to 5), scales (e.g. comparing values among non-proportional temperature scales like Celsius and Fahrenheit might lead to false conclusions) and also contain measurement errors.

Differently, non-metric data refers to categorical values, among which arithmetic operations make no sense. There are two main types of attributes: nominal, which only provide enough information to distinguish one object from another (e.g. its colour), and ordinal, allowing the objects to be ordered (e.g. street numbers).(TAN *et al.*, 2013).

2.1.1. Clustering

Among the multivariate techniques, Cluster Analysis is commonly applied when the variables describing the objects to be examined are not related (in other words, one cannot be predicted or explained by another). So, all variables are analysed simultaneously, in an effort to find an internal structure among the objects that could simplify their analysis and thus unveil previously unidentified relationships.

In order to do that, Cluster Analysis groups instances of a dataset into clusters, aiming to improve the similarity between the ones inside the same group at the same time as maximizing the difference between instances in different clusters. Such analysis might be arguably described as a mere classification task, in which humans perform really well. Although this might be true in the case of few, low-dimensional instances (e.g., classifying 10 car instances based only in 3 dimensions: fuel consumption, type and price range), computer-aided cluster analysis is indispensable when the number of dimensions and objects increases significantly.

There are four main stages in a Cluster Analysis process, and the decisions taken in any of them can significantly impact the final result:

1) **Defining Analysis Purpose Definition**. Taxonomy Description serves both to classify objects and compare the results with theoretically-defined assumptions. Data Simplification aims to reduce the number of discrete entities to consider by grouping the most similar instances into clusters to be further analysed. Finally, Relationship Identification takes advantage of the classification and simplification tasks to unveil previously unidentified relationships.

2) **Establishing Research Design**. Decisions regarding dataset sample size and representativeness, identified outliers removal, need for standardizing numeric variables and similarity measure to apply should be taken at this stage. Previous analysis experience might be helpful in answering some of these questions.

3) **Deriving Clusters and Assessing Overall Fit**. Deciding whether to apply hierarchical or non-hierarchical procedure, and also which algorithm to choose will depend on the sample size and variable values. At the end, the results (e.g. cluster with disparate sizes) might indicate the need for running the analysis again, with different settings.

4) **Interpreting and Validating the Cluster Solution**. The final stage involves analysing the identified relationships and defining whether they apply to the general population.

Finally, it is important to assess if relevant information was not lost during the clustering process, as that could lead to inaccurate results.

2.1.2. K-means

K-means is a well-known and simple clustering algorithm that divides instances from a dataset into non-overlapping clusters. For that, it creates a one-level partitioning in which each object is more similar to the prototype that defines the cluster (i.e. the centroid, which is usually the mean of a group of points) than to the prototype of any other cluster (TAN *et al.*, 2013). The basic K-means algorithm is defined as:

1	Select K points as initial centroids.
2	**repeat**
3	Form K clusters by assigning each point to its closest centroid.
4	Recompute the centroid of each cluster.
5	**until** Centroids do not change.

The first task is to define the K number of clusters to be achieved. This is not a definite decision, and further algorithm executions can be performed with increasing values of K until the quality of the final cluster solution stabilises.

Next, the initial centroid for each cluster must be chosen, with strategies ranging from random selection to applying a hierarchical procedure beforehand. The former usually results in poor cluster solutions and the latter, although more effective, does not scale as well. Again, the approach will vary depending on the considered dataset.

During the iteration phase, instances are grouped according to their similarity towards the previously defined centroid, which is usually expressed in terms of the mean of all objects inside the cluster. Each object is assigned, one at a time, to the cluster with the closest mean to its value. The mean is then updated, including the recently joined object, and the process iterates with the next object. The algorithm terminates when the centroids do not change any more (i.e. no objects are shifting clusters).

2.1.3. Knime

KNIME, or Konstanz Information Miner, is a data analytics platform that serves many purposes such as data pre-processing, analysis and reporting. It integrates functionality from other open source platforms (e.g. statistics functions from R[1] and visualisations functions from Weka[2]). It embeds all these functions in specific building blocks, called nodes, which are divided in main categories such as Data View, Mining, Time Series and more. Chaining these blocks together allow the researcher to prepare and run a complete data analysis workflow in a very intuitive way.

Some of the main benefits of KNIME, besides offering multiple useful tools through a user-friendly interface which manages the data flow seamlessly, are the extensions developed by its community: being a widely recognized open source platform, it encourages developers from distinct areas of knowledge, such as biology and chemistry, to implement customized nodes that fit their own analysis needs, and by that help other professionals from the same field.

---

1   "The r project for statistical computing." [Online]. Available: http://www.r-project.org/

2   "Weka 3 - data mining with open source machine learning software in java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

Figure 2 illustrates a workflow that could be used to measure similarity between instances in a dataset. It involves nodes from the Data IO, Data Manipulation and Mining categories respectively. It is the researcher's job to configure the settings and establish the connection between the nodes.
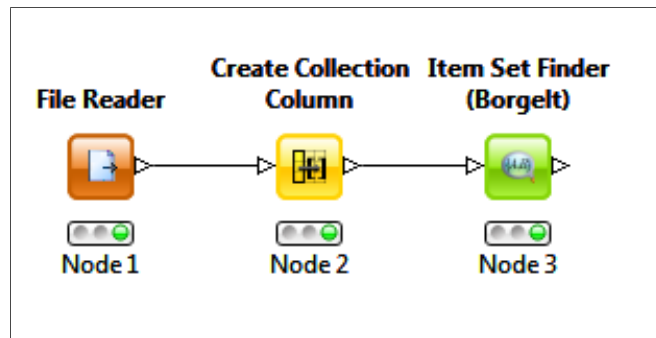

Figure 2. Knime flow example.

2.2. VISUALISATION TECHNIQUES AND TOOLS

"A picture is worth a thousand words" is probably the cliché that can better describe this section. Especially when considering the ever-increasing amount of data humans need to make sense of, Visualisation Techniques and Tools are a crucial link between previously computer-mined data and human cognition. Some examples will be discussed in this section.

2.2.1. PARALLEL COORDINATES

Parallel coordinates consist on a relatively recent visualisation technique, first mentioned in 1985. It has been increasingly researched and applied to multivariate data and high-dimensional geometry since 1991 (HEINRICH; WEISKOPF, 2013). Due to its resemblance with traditional Cartesian coordinates, as both techniques domains are represented in the xy-plane, but with the advantage of representing more than two dimensions at a time (by placing lateral axes alongside either the X or Y axes), Parallel Coordinates are considered a powerful yet familiar visualisation option. The first layer of the graph displays

the parallel axes, or dimensions, which together represent one variable each. Then, the data-points layer maps every instance of the sample to one polyline (a series of lines which connect to one another in the axes). The exact connection point depends on the assumed variable value for the current dimension.

Furthermore, Parallel Coordinates can work well with other multivariate techniques. In the case of Clustering, for instance, it can be used to identify sets of data exhibiting similar characteristics or, if the sample is too big, to visualize the pre-computed clusters and provide the necessary interaction to enhance data exploration and, consequently, insight generation and hypothesis testing.

Different artifices can further enhance the represented information. For instance, when visualisation is deteriorated by many overlapping polylines, a Parallel Coordinates graph plotted using density functions can render an image similar to Figure 3 (HEINRICH; WEISKOPF, 2013). In this case, colours were applied to denote differences in the concentration of discrete polylines.
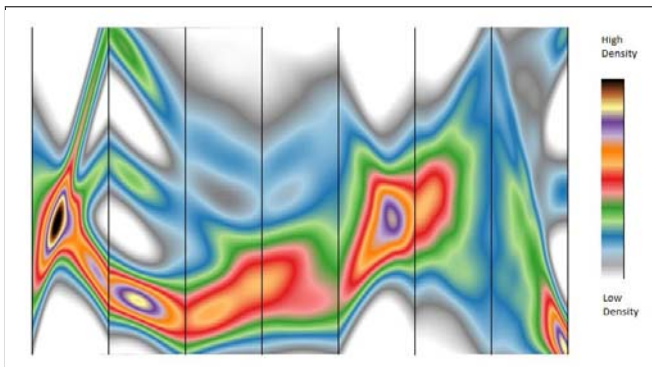

Figure 3. Density functions applied to Parallel Coordinates graph (HEINRICH; WEISKOPF, 2013).

In addition, Parallel Coordinates might be implemented in a way that "...enables the user of a software to change parameters interactively and get immediate feedback from the system (HEINRICH; WEISKOPF, 2013)." Brushing is a very common technique that allows the user to filter sub-samples of the whole population instantly by selecting a specific

range in one or more axes. This way, only the filtered polylines will be high- lighted through all the dimensions, reflecting the selected instances variable values. After some analysis, the user can then decide which action to take against this sub-sample: label, replace, delete and more, depending on the final objective.

2.2.2. Geospatial Data

Despite all the benefits in visualising information with Parallel Coordinates, Geospatial Data is still better represented in a map. In addition, displaying it in conjunction with date information can add a better representation of objects changing over time.

Geospatial Data comprises point, which define a single location based in X and Y coordinates, typically latitude and longitude; line, "...one-dimensional defined object having a length and direction and connecting at least two points"; and polygon, "A two-dimensional figure with three or more sides intersecting at a like number of points." (PADMANABHAN *et al.*, 1992).

There are different Geospatial Data formats that can be used to store map information. The ArcGis - Geographic Information System, developed by Esri, was one of the first widely-adopted map solutions, as its capabilities range from manipulating geospatial and nonspatial data to plotting maps and interacting with the rendered information. For that, it implements shapefiles, "*...an Esri vector data storage format for storing the location, shape, and attributes of geographic features.*"[3] The mandatory file extensions are .shp, designating the file which stores the actual maps geometry data (points, lines and polygons); .shx, an index file to quickly retrieve information from the .shp file, and the .dbf, a database file format used to store the attributes information (e.g. addresses).

More recent data formats, which are better integrated with web technologies are KML - Keyhole Markup Language, and GeoJSON - JavaScript Object Notation. The former is an extension of XML - Extensible Markup

---

3   "ShapefilesHelp | ArcGIS." [Online]. Available: http://doc.arcgis.com/en/arcgis- online/reference/shapefiles.htm

Language, including tags describing spatial and nonspatial information, and is the format used by Google Maps and Google Earth. The latter, although working seamlessly with Javascript web applications, is not restricted to them. Like KML, it stores spatial and nonspatial data in the same file.

Depending on the source of the geospatial data, it might come in any format, not necessarily the one the researcher might be willing to work with. Therefore, software like GDAL - Geospatial Data Abstraction Library are very useful. It consists of a library for geospatial data formats that comes with a variety of command line utilities for data translation and processing.[4] Two relevant tools are "*gdalinfo*", which displays information about a specific file, helping to retrieve its format and specifications (e.g. the coordinate system used), and "*ogr2ogr2*", that converts data between different formats. The way to display the data into a map will be discussed in Subsection 2.2.4.

2.2.3. Glyphs

Whether used to express instances in maps or charts, "*... a glyph is a visual representation of a piece of data or information where a graphical entity and its attributes are controlled by one or more data attributes.*" (WARD, 2010). Some examples of graphical attributes that can be customized are:

- Size: in terms of length, area or volume;
- Position: whether in one, two or three dimensions;
- Shape: including colour, line style and filling pattern;
- Dynamics: speed and direction of motion, rate of flashing;
- Material: texture, opacity, saturation.

In relation to the mapping between graphical and data attributes, there are three possible options:

- One-to-one mapping: each data instance attribute is mapped to one glyph attribute, always aiming to achieve intuitive pairings (e.g. the more calories in a product, the bigger the glyph size);

---

4   "GDAL: GDAL - geospatial data abstraction library." [Online]. Available: http://www.gdal.org/

- One-to-many mapping: one data attribute is represented by two or more glyph attributes (e.g. representing amount of calories with both glyph size and opacity, the latter aiming to reinforce the difference between "light" and "heavy" food);

- Many-to-one mapping: two or more data attributes mapped to the same glyph attribute. This is normally done to facilitate comparison between intra-instance values, considering that they are represented on the same scale (e.g. disclosing the amount of calories in a product, both in its pure form and mixed with other product, using different heights of a vertical bar).

The designer must be cautious when defining which data attributes are being represented, and which mapping to implement, so the final glyph is not difficult to interpret. Furthermore, the cognition of the end users of the visualisation must be considered, in order to make graphical representations more familiar (e.g. setting up possible glyph colours to patterns they are used to).

It is also important to consider biases: for instance, humans can judge line length more accurately than colour; also, closer glyphs are easier to compare than distant ones, so graphical attributes with a low level of distinctiveness should not be applied to a visualisation where glyphs might be distant one from another.

Finally, the layout in which glyphs are displayed on the screen can follow three strategies. A uniform fashion makes both glyphs scale and position homogeneous on the screen. In a data-driven strategy, their position depends on a specific instance attribute value; and if an implicit or explicit data structure is considered to arrange the glyphs on the screen (e.g. a previously identified hierarchy among the data objects resulting in a decision-tree-like graphic design), this is called a structure-driven strategy.

2.2.4. D3.js

Among all visualisation frameworks available today, D3.js is one of the most versatile and promising. In fact, it is highly compatible with all the Visualisation Tools and Techniques discussed so far in this work. In addition, "*D3's emphasis on web standards gives you the full capabilities*

of modern browsers without tying yourself to a proprietary framework, combining powerful visualisation components and a data-driven approach to DOM manipulation."[5]

The mentioned web standards are HTML - Hyper Text Markup Language, that describes all the elements in a Web page; SVG - Scalable Vector Graphics, defining vector-based graphics to be drawn in a Web page; CSS - Cascading Style Sheets, which can define styles for any element described by either HTML or SVG; and DOM - Document Object Model, the "*interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents.*"[6]

Being an open-source Javascript library that provides seamless integration with such native web technologies, D3.js inspired programmers all over the world to develop new extension packages[7], turning it into a very powerful yet relatively simple tool.

3. PROPOSED SOLUTION

The Statistics Table referred in Subsection 1.2.2. stores information about events from 2003 to current days, and is considered a repository for all the company's structured data. It has approximately 50.000 rows and 32 distinct column titles. As a single event might hold multiple values for a specific column title, some columns are replicated in order to accommodate them all (e.g. identified individuals participating on an event), comprising total of 189 columns. Their categories, titles, data types and quantities are described in Table 1. For the sake of clarity, some possible values for data validated columns are shown in Table 2. Finally, this dataset only comprises locations in the UK. Although the actual data was not provided by the company, but only the Statistics Table columns headers, datatypes and validation values, the following analysis considers a rough estimate number of rows, observed during one of the company meetings. Nevertheless, dummy data was generated to aid the understanding of some of the techniques described.

---

5   "D3.js - data-driven documents." [Online]. Available: http://d3js.org/

6   "World wide web consortium (W3C)." [Online]. Available: http://www.w3.org/

7   Data Visualization Libraries Based on D3.JS - Mike McDearmon'. [Online]. Available: http://mikemcdearmon.com/portfolio/techposts/charting-libraries-using-d3.

Table 1. Statistics table columns titles and types

Statistics Table

Category	Column Title	Data Type	Qty
Date / Time	Date of Incident (from)	Date	1
	Date of Incident (to)	Date	1
	Time of Incident (from)	Time	1
	Time of Incident (to)	Time	1
Event	Reliability	Validation	1
	Incident Type	Validation	3
	Type of Campaign	Validation	5
	Specific Campaign	Free Text	1
	Sub Campaign	Free Text	1
	Type of Direct Action	Validation	2
	Demonstration Announcement	Validation	1
Targets	Company Indirect Target	Free Text	2
	Company Direct Target	Free Text	3
	Individual Target	Free Text	2
	Company of Indiv. Target	Free Text	4
Participants	Group Involved	Free Text	4
	Individuals Involved	Free Text	14
	No. of Attendees	Free Text	1
Location	Town	Free Text	1
	Region	Free Text	1
	Country	Free Text	1
Source	Date of Posting	Date	20
	Source	Free text	20
	Author	Free text	20
	URL	Int. number	20
	Report Title	Free text	20
	Report In Full	Free text	20
	Document Reference	Int. number	20

Table 2. Examples of data validation values for the dimensions considered

Data Validation Values		
Incident Type	**Type of Direct Action-**	**Campaign type**
March	Arson	Anti-Vivisection
Sit-In	Bomb	Anti-Fur
Lock-On	Incendiary Device	Anti-Hunting
Intrusion	Hoax Bomb	Anti-Meat Industry

Both the size and the high dimensionality of the Statistics Table indicate that multivariate data analysis techniques could be applied to extract knowledge from it. However, an overall goal for the analysis must be established beforehand. Considering the great amount of collected Individuals Involved names, this work will suggest a way to separate dangerous demonstrators from peaceful protesters.

Performing a first analysis on the dataset, it is noticed that there is no dependence relationship between the attributes. In addition, the great majority of them are non-metric, and there is also geospatial and datetime information that could be relevant. Taking into account both the current dataset and the available mining, visualisation and interaction technologies which could boost insight generation and hypothesis testing, the overall goal will be split in three related sub-objectives:

1) To group previously identified individuals, so newcomers can be automatically profiled;

2) To test motivation hypothesis for more dangerous actions and also suggest possible future targets;

3) To unveil supposed third-party financing and professional commitment.

Once the specific objectives are defined, a feature subset selection should be performed to remove redundant or irrelevant attributes to them, which could reduce classification accuracy and the quality of the clusters. (TAN *et al.*, 2013).

In the interest of simplicity, a manual selection based on empiric knowledge, rather than a systematic approach, will be performed. As a result, the dimensions to be considered in this work are highlighted in Table 1.

3.1. PROFILING NEW AND EXISTING INDIVIDUALS

In order to develop a Taxonomy Classification regarding the activism profile of the Individuals Involved, and also to perform Data Simplification against this big dataset without losing relevant information, Cluster Analysis was chosen as the multivariate technique to be implemented. The Knime platform can be very useful in running the necessary data transformations and also the clustering algorithm.

In order to assign individuals into activism-profiled groups, the other clustering variables considered for similarity measuring are Campaign Type and Incident Type of the events they attended.

However, as there are multiple columns describing them (3 for Incident Types, 5 for Campaign Types and 14 for Individuals Involved), their values should be aggregated in collections, so only 3 dimensions are considered, instead of 22.

Following the steps in Subsection 2.1.1, the next thing to assess is the sample size. Considering there are long lasting activist groups, all events since 2003 should be considered, so maximum representativeness is achieved. Although outliers might be detected, the decision to remove them can only be taken when analysing the actual dataset. Also, as the clustering variables are nominal, there is no need to standardise the data.

3.1.1. Similarity measure

After that, a similarity criteria between pair of objects must be established, to allow later grouping. For instance, considering the 5 different possible values in the Campaign Types for every single event, coefficients ranging from 0 to 1 could be assigned to pairs of events (in which 0 indicates that, for that dimension, none of their values match, and 1 states that all values match, no matter their order within the collection).

Decimal coefficients might be considered (e.g. similarity coefficient of 0.4 when 2 out of 5 values match between the pair or events). However, it is necessary to analyse the actual dataset in order to suggest optimized inter- mediate coefficients, by testing them and verifying the outcomes. After all, different coefficients might be chosen if the great majority of events hold only 3 values for Campaign Types, or if the amount of blank values is highly disparate among them. Although the same principle could be applied to the other clustering variables, Individuals Involved should require further analysis, as it might describe individuals from more than one group per event.

3.1.2. Applying the K-means algorithm

Then, it is necessary to define the clustering process to be applied. K-means is a good candidate both because it is simple to implement and cost-effective in terms of computer processing. The first step is to define the K amount of desired clusters and their initial centroids. As stated on the previous subsection, if the final result is not satisfactory, both parameters can be changed for a new algorithm execution.

Initially, 2 clusters will be sought. Their initial centroids should be as far apart to each other as possible, in order to optimize the algorithm execution and achieve better results. As a first attempt, they will be chosen according to empirical evidence related to Campaign Types: one will comprise the most frequent combination of values regarding animal killing (e.g. Anti-fur, Anti-vivisection, Anti-foie gras) and the other will relate to using animals for human joy (e.g. *Anti-Animal Circuses, Anti-Zoos, Anti-Horse Racing*).

After all the objects have been grouped, it is time to analyse the final cluster solution and decide whether to run the algorithm again using a different specification. For instance, by choosing initial centroids based on the most frequent combinations of attributes. When the solution is finally accepted, the researcher should be able to interpret the revealed relationships among different Campaign Types and Incident Types that form the groups' profile, and also identify which individuals are most sympathetic to them. Lastly, it should be assessed if the findings can be applied to the general population, and thus to future events.

3.2 DETECTING POSSIBLE MOTIVATION FOR MORE DANGEROUS ACTIONS, AND INFERRING NEW TARGETS

Once the activism-profiled clusters and the individuals within are defined, investigating the Type of Direct Action they are regularly involved in and the most frequent Target Companies might give some clues about their motivation and level of capability, and also hypothesize possible future targets.

As stated in subsection 2.2.4., relationship finding can be highly enhanced by plotting and interacting with information in a Parallel Coordinates graph. In this case, its axis layer would comprise, in addition to the previously identified Profiled Groups, the dimensions Individuals Involved, Type of Direct Actions, Target Company and Date of Incident (from). Figure 4 reflects such graph, in which dummy data in CSV - Comma Separated Values format was provided to a D3.js script, and colours distinguish individuals from different groups. Once plotted, patterns can be identified by applying the brushing technique to some specific dimensions values.

For instance, if a sub-sample containing more dangerous Type of Direct Actions (e.g. bomb or incendiary device) is selected through brushing, a greater convergence of polylines towards specific individuals would indicate they applied such harmful capabilities in multiple occasions, thus allowing to distinguish them from other individuals that, despite identified, might be attending the same events peacefully.
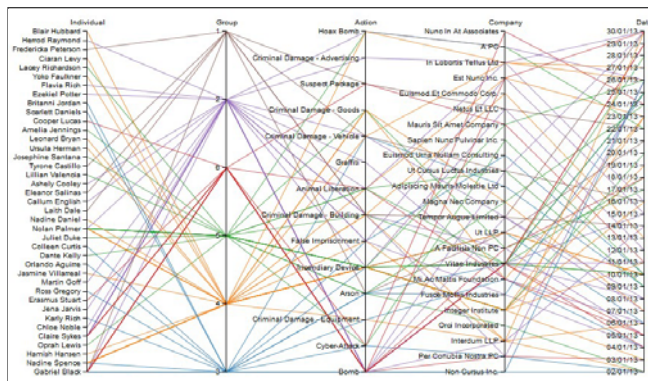

Figure 4. Dummy data plotted into a Parallel Coordinates graph.

Also, it would reveal whether these threatening people are more active regarding specific campaigns or against determined companies, by assessing polyline concentration on the Profiled Groups and Target Company dimensions respectively. If confirmed, this fact might demonstrate a possible relationship that deserves further investigation in other information sources and which could point out that a dissatisfied former employee is behind those critical attacks, for instance.
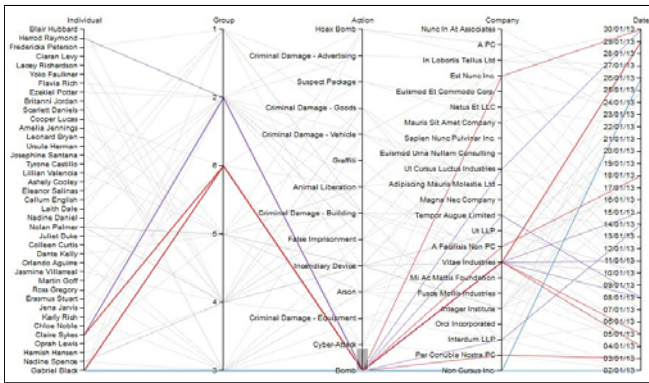
Figure 5. Identified patterns by brushing the value bomb from the action axis.

Figure 5 reflects a similar situation, as it clearly shows that 3 individuals are involved in approximately 15 occurrences of bomb: Gabriel Black, Claire Sykes and Herrod Raymond. The different colour intensity suggests the latter was involved in only one incident. On the other side, the formers are far more active, and participate in two groups each.

According to the blue polylines in Figure 4, Group 3 has multiple individuals and is involved in different actions. However, when brushing the action bomb, it is noticed that only Gabriel Black was involved in such attacks. Also, he targeted specifically Non Cursus Inc two times in the same month. Further brushing reveals that those were the only registered events for that company, what could indicate a deeper relationship between the attacker and the target.

Finally, there is still the possibility to gain some geospatial insight if the interesting events are inserted into a map tool that uses glyphs to describe the most relevant attributes of the plotted instances. Essentially, if a sub-sample containing the most dangerous Individuals Involved is selected from the Parallel Coordinates graph and plotted into a map together with the location and date of the events in which they behaved criminally, this might reveal a travel pattern that, if compared to the addresses of similar companies already represented in the map, could suggest the next possible target.

3.3. UNVEILING SUPPOSED THIRD-PARTY FINANCING TO ATTACKS

In addition, glyphs on a map could also disclose the date in which such capable Individuals Involved attended distinct events. If its noticed that they had covered great distances within short periods of time, this might indicate some kind of sponsorship or professional commitment. Such assumptions could become stronger if, analysing the Parallel Coordinates graph again, it is observed that they did not choose specific Type of Campaigns and targeted Companies indiscriminately.

To plot this map, it is necessary to search for the fundamental geospatial information to render the UK map into the visualisation tool using a D3.js script. NaturalEarth[8], an online database providing worldwide map datasets in the ESRI shapefile, is a good choice. Presumably, the available files come with general map information only, thus it is necessary to append the variables from Table 1 to be displayed in the final map.

Although the Statistics table does not contain Latitude - Longitude information, once converted to a text file, e.g. CSV, then it is possible to merge its rows to the objects within the downloaded shapefile using the GDAL tool ogr2ogr. In this case, the matching parameter would be the town name, available in both files. Nonetheless, further data transformations might be necessary before merging both files, such as counting total events and attendees in one specific town. Again, Knime can be very useful in performing these tasks.

Lastly, it is worth mentioning that some reflection is crucial to define the mapping between data attributes and glyphs properties. A systematic approach that accounts for both the "...*hierarchy of concept categorization and the ordering of discriminative capacity of visual channels*" (MAGUIRE *et al.*, 2012) should be considered in order to maximize insight gaining. Additionally, considering patterns already familiar to the end user might enhance their cognition capabilities during the analysis.

---

8   Natural Earth [online]. Available: http://www.naturalearthdata.com/

4. Conclusion

This work proposed a viable approach to enhance the analysis of a highly dimension dataset stored in a spreadsheet file and containing information about animal rights activism events. Specific objectives were outlined in order to demonstrate the capabilities data mining tools and interactive visualisation techniques bring to knowledge extraction and insight generation. Although only the column headers and their datatypes were provided as a basis for this work, dummy data was generated and tested against the suggested approach, indicating it would also fit the actual dataset.

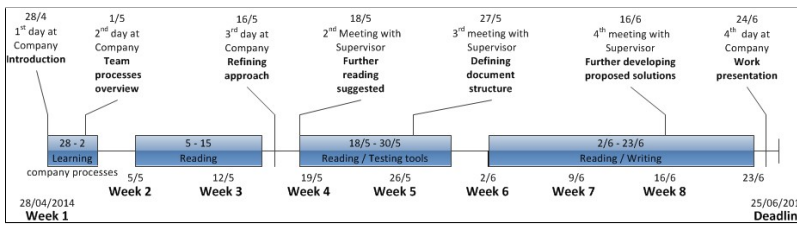Figure 6 reflects the activities performed during the length of the project.

Fig. 6: Activities performed during the project.

5. Acknowledgments

Rodrigo Alves Carvalho

Federal Forensic Expert at the Technical-Cientific Unit in the Rio Grande do Sul State and doctoral candidate at the Centre for Doctoral Training in Cyber Security, University of Oxford.

E-mail: rodrigo.carvalho@cs.ox.ac.uk

Resumo

Análise Visual para Inteligência de Fontes Abertas

Tecnologias de visualização de dados aprimoradas podem contribuir decisivamente para as empresas que dependem da interpretação de uma grande quantidade de informações como objetivo principal de seu negócio. Técnicas gráficas frequentemente empregadas para apresentar a informação processada e os resultados encontrados ao cliente final, como figuras e gráficos de dispersão, poderiam ser melhorados e personalizados para ajudar os analistas durante a fase de investigação. Após examinar as tarefas de análise interna e a estratégia de armazenamento de dados de uma empresa de consultoria de segurança, este trabalho sugere uma abordagem para melhorar a extração de conhecimento, a geração de cenários de interpretação e o teste de hipóteses, aplicando novas técnicas de visualização e garimpo de dados interativos.

Palavras-chave: Análise visual. Garimpo de dados. Ferramentas de fontes abertas.

6. References

HAIR JR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.. **Multivariate Data Analysis.** Harlow: Pearson New International Edition, 2013, 752 p.

HEINRICH, J.; WEISKOPF, D. State of the Art of Parallel Coordinates. **Eurographics 2013**, p. 95-116, t5+refs. [Online]. Available: http://classes.soe.ucsc.edu/cpms261/Fall13/papers/hcmarsh/ StateXofXArtXofXParallelXCoordinates.pdf

IEEE. **CALL FOR PARTICIPATION: VAST papers**. [Online]. Available: http://ieeevis.org/year/2014/info/call- participation/ vast-papers.

J. THOMAS, J.; COOK, K. A Visual Analytics Agenda. **Computer Graphics and Applications**, IEEE, vol. 26, n. 1, pp. 10–13, Jan. 2006.

MAGUIRE, E.; ROCCA-SERRA, P.; SANSONE, S. A.; DAVIES, J.; and CHEN, M..Taxonomy-based glyph design with a case study on visualizing workflows of biological experiments, **Visualization and Computer Graphics**, IEEE Transactions on, vol. 18, n. 12, p. 2603-2612, 2012.

PADMANABHAN, G.; YOON, J.; LEIPNIK, M. **A Glossary of GIS Terminolog**y. NCGIA (National Center for Geographic

Information and Analysis), 1992. [Online]. Available: http://faridesm.ir/Glossary.pdf.

TAN, P. N.; STEINBACH, M.; KUMAR, V.. **Introduction to Data Mining**. Harlow: Pearson New International Edition, 2013. 736 p.

WARD, M. **Interactive data visualization: foundations, techniques, and applications**. Natick, Mass: A. K. Peters, 2010.

❖❖❖