A study on Data Outlier and Statistical Analysis in Digital Forensics

Mr.K.Vijay Babu[1]
CMR Engineering College,
Medchal.

Mr.V.NarsingRao[2]
Sphoorthy Engineering College,
Nadergul,R.R.District

Abstract

Statistics and statistical thinking have become increasingly important in a society that relies more and more on information and calls for evidence. Hence the need to develop statistical skills and thinking across all levels of education has grown and is of core importance in a century which will place even greater demands on society for statistical capabilities throughout industry, government and education.

A natural environment for learning statistical thinking is through experiencing the process of carrying out real statistical data investigations from first thoughts, through planning, collecting and exploring data, to reporting on its features. Statistical data investigations also provide ideal conditions for active learning, hands-on experience and problem-solving. No matter how it is described, the elements of the statistical data investigation process are accessible across all educational levels.

Real statistical data investigations involve a number of components: formulating a problem so that it can be tackled statistically; planning, collecting, organising and validating data; exploring and analysing data; and interpreting and presenting information from data in context. No matter how the statistical data investigative process is described, its elements provide a practical framework for demonstrating and learning statistical thinking, as well as experiential learning in which statistical concepts, techniques and tools can be gradually introduced, developed, applied and extended as students move through schooling.

1. Introduction

A natural environment for learning statistical thinking is through experiencing the process of carrying out real statistical data investigations from first thoughts, through planning, collecting and exploring data, to reporting on its features. There are four processes involved in a statistical investigation: Collection of data(information) Data for a statistical investigation can be collected from records from surveys.

Collection of data (information) Data for a statistical investigation can be collected from records, from surveys by direct observation or by measuring or counting. Unless the correct data is collected, valid conclusions cannot be made. Organisation and display of data Data can be organised into tables and displayed on a graph. This allows us to identify features of the data more easily. Calculation of descriptive statistics Some statistics used to describe a set of data are the centre and the spread of the data. These give us a picture of the sample or population under investigation. Interpretation of statistics

This process involves explaining the meaning of the table, graph or descriptive statistics in terms of the variable, or theory, being investigated.

The variable is the subject that we are investigating. The entire group of objects from which information is required is called the population. Gathering statistical information properly is vitally important. If gathered incorrectly then any resulting analysis of the data would almost certainly lead to incorrect conclusions about the population.

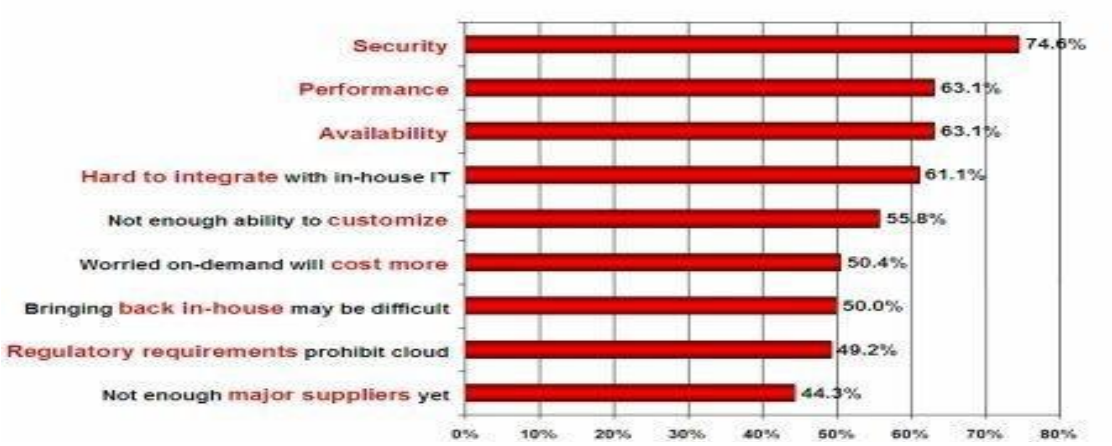
Data are individual observations of a variable. A variable is a quantity that can have a value recorded for it or to which we can assign an attribute or quality. Two types of variable that we commonly deal with are categorical variables and numerical variables.

The distribution of a set of data is the pattern or shape of its graph. For the example above, the graph has the general shape shown alongside: This distribution of the data is said to be negatively skewed because it is stretched to the left (the negative direction). Outliers that the horizontal is a number line with numbers in ascending order from left to right. are data values that are either much larger or much smaller than the general body of data. Outliers appear separated from the body of data on a frequency graph.

2. Literature Survey

A picture of a data set can be obtained if we have an indication of the centre of the data and the spread of the data. Three statistics that provide a measure of the centre of a set of data are: the mean the median and the mode. Many data sets have frequency distributions that are 'bell-shaped' and symmetrical about the mean. For example, the histogram alongside exhibits this typical 'bell-shape'. The data represents the heights of a group of adult women and has a mean of 165 and a standard deviation of 8.

Data carving is a very important topic in digital investigation and computer forensic. Researches are needed to focus on improving data carving techniques to enable digital investigators to retrieve important data and evidences from damaged or corrupted data resources.



.

Data accuracy: The definition of information accuracy depends on the particular application that the detector network is intended. as an example, in a very target localization drawback, the estimate of target location at the sink determines the information accuracy.

Latency: Latency is outlined because the delay concerned in knowledge transmission, routing and knowledge aggregation. It may be measured because the time delay between the knowledge the information the information packets received at the sink and therefore the data generated at the supply nodes. In this paper we tend to square measure discussing the various aspects of in-network information aggregation, totally different strategies bestowed already, and its options. the most aim of this paper is to spot however precisely the method of information aggregation works for digitial forensics, and what square measure the vital performance metrics those square measure settled with use of economical information aggregation methodology. Data carving is a very important topic in digital investigation and computer forensic. Researches are needed to focus on improving data carving techniques to enable digital investigators to retrieve important data and evidences from damaged or corrupted data resources.

1. **Model for Data Preparation**

Examiners begin by asking whether there is enough information to proceed. They make sure a clear request is in hand and that there is sufficient data to attempt to answer it. If anything is missing, they coordinate with the requester. Otherwise, they continue to set up the process.

The first step in any forensic process is the validation of all hardware and software, to ensure that they work properly. There is still a debate in the forensics community about how frequently the software and equipment should be tested. Most people agree that, at a minimum, organizations should validate every piece of software and hardware after they purchase it and before they use it. They should also retest after any update, patch, or reconfiguration.

When the examiner's forensic platform is ready, he or she duplicates the forensic data provided in the request and verifies its integrity. This process assumes law enforcement has already obtained the data through appropriate legal process and created a forensic image. A forensic image is a bit-for-bit copy of the data that exists on the original media, without any additions or deletions. It also assumes the forensic examiner has received a working copy of the seized data. If examiners get original evidence, they need to make a working copy and guard the original's chain of custody. The examiners make sure the copy in their

possession is intact and unaltered. They typically do this by verifying a hash, or digital fingerprint, of the evidence. If there are any problems, the examiners consult with the requester about how to proceed.

After examiners verify the integrity of the data to be analyzed, a plan is developed to extract data. They organize and refine the forensic request into questions they understand and can answer. The forensic tools that enable them to answer these questions are selected. Examiners generally have preliminary ideas of what to look for, based on the request. They add these to a "Search Lead List," which is a running list of requested items. For example, the request might provide the lead "search for child pornography." Examiners list leads explicitly to help focus the examination. As they develop new leads, they add them to the list, and as they exhaust leads, they mark them "processed" or "done."

For each search lead, examiners extract relevant data and mark that search lead as processed. They add anything extracted to a second list called an "Extracted Data List." Examiners pursue all the search leads, adding results to this second list. Then they move to the next phase of the methodology, identification.

2.Identification :

Examiners repeat the process of identification for each item on the Extracted Data List. First, they determine what type of item it is. If it is not relevant to the forensic request, they simply mark it as processed and move on. Just as in a physical search, if an examiner comes across an item that is incriminating, but outside the scope of the original search warrant, it is recommended that the examiner immediately stop all activity, notify the appropriate individuals, including the requester, and wait for further instructions. For example, law enforcement might seize a computer for evidence of tax fraud, but the examiner may find an image of child pornography. The most prudent approach, after finding evidence outside the scope of a warrant, is to stop the search and seek to expand the warrant's authority or to obtain a second warrant.

If an item is relevant to the forensic request, examiners document it on a third list, the Relevant Data List. This list is a collection of data relevant to answering the original forensic request. For example, in an identity theft case, relevant data might include social security numbers, images of false identification, or e-mails discussing identity theft, among other things. It is also possible for an item to generate yet another search lead. An email may reveal that a target was using another nickname. That would lead to a new keyword search for the new nickname. The examiners would go back and add that lead to the Search Lead List so that they would remember to investigate it completely.

An item can also point to a completely new potential source of data. For example, examiners might find a new e-mail account the target was using. After this discovery, law enforcement may want to subpoena the contents of the new e-mail account. Examiners might also find evidence indicating the target stored files on a removable universal serial bus (USB) drive—one that law enforcement did not find in the original search. Under these circumstances, law enforcement may consider getting a new search warrant to look for the USB drive. A forensic examination can point to many different types of new evidence. Some other examples

include firewall logs, building access logs, and building video security footage. Examiners document these on a fourth list, the New Source of Data list.

After processing the Extracted Data list, examiners go back to any new leads developed. For any new data search leads, examiners consider going back to the Extraction step to process them. Similarly, for any new source of data that might lead to new evidence, examiners consider going all the way back to the process of obtaining and imaging that new forensic data.

At this point in the process, it is advisable for examiners to inform the requester of their initial findings. It is also a good time for examiners and the requester to discuss what they believe the return on investment will be for pursuing new leads. Depending on the stage of a case, extracted and identified relevant data may give the requester enough information to move the case forward, and examiners may not need to do further work. For example, in a child pornography case, if an examiner recovers an overwhelming number of child pornography images organized in usercreated directories, a prosecutor may be able to secure a guilty plea without any further forensic analysis. If simple extracted and identified data is not sufficient, then examiners move to the next step, analysis.

3.Analysis:

In the analysis phase, examiners connect all the dots and paint a complete picture for the requester. For every item on the Relevant Data List, examiners answer questions like who, what, when, where, and how. They try to explain which user or application created, edited, received, or sent each item, and how it originally came into existence. Examiners also explain where they found it. Most importantly, they explain why all this information is significant and what it means to the case.

Finally, after examiners cycle through these steps enough times, they can respond to the forensic request. They move to the Forensic Reporting phase. This is the step where examiners document findings so that the requester can understand them and use them in the case. Forensic reporting is outside the scope of this article, but its importance can not be overemphasized.

Conclusion:

This paper is a helpful for data extraction , data carving and introduction to Data outlier detection and statistical analysis to study the various types of Investigation in the digital forensics methodology.

References

[1] K. Dasgupta, K. Kalpakis, and P. Namjoshi, “An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks”, IEEE 2003.

[2] . E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, “InNetwork Aggregation Techniques for Wireless Sensor Networks: A Survey”, IEEE Wireless communication 2007.

[3] M. Lee and V.W.S. Wong, "An Energy-aware Spanning Tree Algorithm for Data Aggregation in Wireless Sensor Networks," IEEE PacRrim 2005, Victoria, BC, Canada, Aug. 2005.

[4] Yong Yao and Johannes Gehrke, "Query processing in sensor networks," in Proceedings of the First Biennial Conference on Innovative Data Systems Research (CIDR), 2003.

[5] Dragan Petrovic, Rahul C. Shah, Kannan Ramchandran and Jan Rabaey, "Data Funneling: Routing with Aggregation and Compression for Wireless Sensor Networks," in Proceedings of the IEEE Sensor Network Protocols and Applications (SNPA), May 2003.N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson,J. Rexford, S. Shenker, and J. Turner. OpenFlow: Enabling Innovation in Campus Networks. ACM SIGCOMM Computer Communication Review (CCR), 38(2):69–74, 2008.

[6] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown, and S. Shenker. NOX: Towards an Operating System for Networks. ACM SIGCOMM Computer Communication Review, 38(3):105–110, 2008.

[7] R. Tagnipes. High Availability with Dynamic Load Balancers. GoGrid Blog, 4 Feb, 2013. http://blog.gogrid.com/2013/02/04/high-availabilitywith- dynamic-load-balancers/.

[8] S. Shin, P. Porras, V. Yegneswaran, M. Fong, G. Gu, and M. Tyson. FRESCO: Modular Composable Security Services for Software-Defined Networks. In ISOC Network and Distributed System Security Symposium (NDSS), 2013.

[9] V. Mann, A. Vishnoi, K. Kannan, and S. Kalyanaraman. CrossRoads: Seamless VM Mobility Across Data Centers through Software Defined Networking. In Network Operations and Management Symposium (NOMS), 2012 IEEE, pages 88–96, 2012.

[10] OpenFlow Network Research Center, Retrieved 14 June, 2013. http: //onrc.stanford.edu/.

[11] Syed Taha Ali, Member, IEEE, Vijay Sivaraman, Member, IEEE, Adam Radford, Member, IEEE,and Sanjay Jha, Senior Member, IEEE , A Survey of Securing Networks using Software Defined Networking .

[12] Takabi H, Joshi J B D, Ahn G. Security a nd privacy challenges in cloud computing environments. IEEE Security & Privacy;2010;8(6) :24–31.

[13] Esteves, R.M. and Chunming Rong, "Social Impact of Privacy in Cloud Computing" in 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), N ov. 30-Dec. 3 ,2010, pp. 593-596

[14] Ricardo vilaca, Rui oliveira 2009. Clouder: A Flexible Large Scale Decentralized Object Store. Architecture Overview. Proceeding of WDDDM '09