Digital Forensic Analysis of Internet History Using Principal Component Analysis

David W. Gresty, Diane Gan and George Loukas
C-SAFE Centre, Dept. of Computing & Information Systems, University of Greenwich, UK.
D.Gresty@Greenwich.ac.uk

Abstract--A modern Digital Forensic examination, even on a small-scale home computer typically involves searching large-size hard disk drive storage, a variety of host and web-based applications which may or may not be known to the investigator, and a proliferation of web-based Internet history artefacts that may be highly significant to showing the motivation of a suspect. Faster keyword searching and larger and more accurate sets of file hashes may point the investigator to relevant artefacts but when dealing with the new or the unknown, or there is a need to holistically profile the activity of the computer, the investigator is left with a manual and labour-intensive investigation. This paper proposes using an unsupervised statistical learning technique called Principal Component Analysis to provide a novel approach to the analysis of Digital Forensic Internet history. The approach groups and analyses artefacts to produce a high-level context view of the timeline data. The paper proposes a Principal Component Analysis approach and the selection of the appropriate number of Principal Components is described using the Scree test method. A case study of the approach is shown, first using a simulated set of data test comprising of 820 Mozilla Internet History artefacts and then using a set of 5900 Internet Explorer history artefacts from real-world browser data. The results of the analysis are presented in a tabular format that provides an accessible overall view of the activity within the timeline. They show a promising approach to effectively and simply represent large quantities of timeline data at a high-level where basic patterns of usage can be determined. Further work on enhancing the proposed approach to include low-level pattern rules is discussed.

Keywords--Digital Forensics; Internet History; Principal Component Analysis

I. INTRODUCTION

A. Timeline data

A digital forensic timeline is a time-ordered list constructed from point events recorded on a system that is under investigation. These are considered point events because the continuous use of the system is not typically recorded, rather what is normally examined within a forensic investigation is the end state of the system and its constituent files, and data that purports to show when the state of the system changed. Time-ordered lists of events can be constructed from artefacts at the file system, operating system and application level. The events broadly fall into the categories of creation, modification, access and in some cases destruction. Because time is a standard characteristic and the types of point events that are recorded on systems are broadly the same from source to source, it is possible to combine timelines from heterogeneous sources such as the file system, operating system or application logs. This ability to combine artefacts is the foundation of Super-Timeline Analysis [5]. A number of tools are available to produce and present timeline data, [14][15]. Although production is relatively straightforward, presentation that is more substantial than only showing artefacts per time period is not a trivial process. Existing work on timeline analysis tends to identify low-level events and possibly combine them together to form high-level events that are usable by an investigator, but this kind of method requires prior knowledge or known patterns of behaviour to search for [6].

Marrington's [9] doctoral thesis on 'computer profiling' identifies that traditional low-level models of computing behaviour are inappropriate and that "a framework for practically describing a computer system and its history at a level of abstraction suitable for a human investigator is still absent" and goes on to conclude that "digital forensics literature lacks a formal model which can be used in practical digital investigations to describe an entire computer system and its history".

Gladyshev and Patel [4] provide some interesting formalisation of the time boundaries to events and show that there is a transitive relationship between events. Abraham [1] discusses Event Chains, which are distinct actions relevant to an investigation that occur in a sequential order. Such an event chain can be represented $A \rightarrow B \rightarrow C$, but may, or may not have other optional events within the chain such as $A \rightarrow B \rightarrow D \rightarrow C$. This would therefore create an event chain rule AB*C with some kind of 'temporal sliding window' between the events. This fits rather nicely with Gladyshev and Patel [4], where the time of events A and B, $T^A < z < T^B$, where z is the 'sliding window' for the chain. Abraham's paper also discusses habitual and repetitive behaviour and the problem of an event chain ABAB being either a distinct pattern by itself or a repetition of AB, which by itself might not be significant.

B. Internet History

Internet history artefacts contain a time, which is quite often but not always UTC and may require processing to local dates and times [10]. Different browsers have different levels of resolution for the date time artefacts, with [10] showing that Microsoft Internet Explorer (IE) records artefacts at the 100 nanosecond level, up to Safari which records at the 1 second level. Boyd and Forster [2] describes in detail the structure of an IE Internet history record, which is implementation specific to that software but ultimately every record is a date/time and URL pair. The purpose, number and location of the records have significance to the software that uses them, but from a timeline point of view it is important to identify the presence of an artefact URL located at a reliable time such that it can be placed onto an event point timeline and where necessary de-duplicated.

C. Purpose of this Research

The research aim of this project is to profile systems based upon the timeline data, as this provides a heterogeneous characteristic that is present across a variety of artefacts. By profiling of the timeline we aim to show how the system is being used, any normal and abnormal behaviour of the system and potential identify the user of the system where there may be multiple possible users or usage characteristics. This paper focuses on web-based Internet history records for timeline analysis as they record activity that is highly interactive and involves a user.

Section II of the paper outlines a novel approach to the analysis of Internet history timelines using Principal Component Analysis. Section III demonstrates the Principal Component Analysis approach using two sets of data as case study. Section IV discusses enhancing the case study, compares the results of the two data sets and highlights some of the research issues that have been raised to date using this approach for timeline analysis. The paper concludes with an overview of the results, future work and references.

II. Method

A. Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised statistical learning technique for data reduction [7]. The use of unsupervised technique is desirable as it requires no prior knowledge of the data or any training phase before the analysis which might be considered a 'black box' process and is undesirable when considering codes of practice for electronic evidence such as the ACPO guidelines' [12] third principle which states that an independent third party should be able to replicated and verify the processes and results of a digital forensics investigation. As input, PCA uses a matrix of *cases* by *variables* and the output of the analysis is a reduced set of data, the Principal Components (PC), which have values showing how strongly the variables correlate to each Principal Component. For example, processing a sample dataset of children's height, weight and eye colour through PCA might reduce to two Principal Components: One component showing a high correlation between height and weight and very little correlation with eye colour, the second component showing little correlation with height and weight but a high correlation with eye colour.

Within the literature, there are different methods for the selection of the number of Principal Components to be used in the analysis, with the Eignevalue rule or Cattell's Scree test being popular methods. Within this paper we discuss using the Scree test [3], as we have found that this approach provides smaller number of Principal Components whilst capturing the highly variant data. To create a 'Scree plot' the Eigenvalues - values calculated from the square matrix of our variables - are plotted on a graph in descending order and where there is a significant drop off on the graph, where it is said to 'elbow', this would show an appropriate point to select the number of Principal Components. An example of a Scree plot can be seen in figure 1.

With respect to Internet history data we can see that the selection of a small number of Principal Components does not adequately capture the variety of the users' behaviour, and similarly too large a number of Principal Components does not adequately group related variables into a ‘behaviour’. An interesting area of further research that has been identified is that the amount of variance or repetition in a 'typical' Internet history is unknown. Empirically we have been selecting the numbers of Principal Components that capture 35-45% of the variance in the data. Ideally the Principal Component should contain a minimum of two, preferably three or more variables that highly correlate with the it.

B. Internet History Data

To analyse Internet history using PCA, events must be recorded on the timeline in a *case* by *variable* matrix where the cases are the time point events and the variables are occurrences of Internet artefacts. For example, if the Internet timeline showed five records with access to 'www.organisation.org' with a timestamp of 00:00:05 on a particular date then there would be a single case for the '00:00:05' event and the five occurrences would be recorded as a magnitude in a variable, which in our experiments would be called 'organisation.org'.

C. Time cases

There are a variety of levels of precisions when dealing with digital timestamps as noted in [10]. The second-level of precision is the common minimum level of time precision that can be seen across log files and meta-data that are suitable for constructing timelines for Internet history. We have performed tests data grouped using larger time windows than the 1-second level of precision, such as cases that contain all the events within a 5, 10 or 30 second window. There are advantages to grouping data in larger time windows, especially when the timeline has been constructed from more than one source and there is a concern that the artefacts are not synchronised, for example file system timestamps showing creation times before the web artefacts showing them appearing on the computer. The disadvantage of large time windows, especially very large time windows, is that data dependency can be introduced between the variables.

D. Number of variables

There is a need to sample a characteristic from the Internet history records to be the variable in the analysis, as although each individual record in the Internet history or point event on the timeline could be considered a distinct variable this would provide little to no grouping, and a larger number of variables compared to a small number of cases is undesirable for PCA. At this time, we have selected variable by the domain name contained within the Internet history record URI. From an investigative point of view there is a significant difference between artefacts 'mail.organisation.org', 'www.organisation.org' or even 'ftp.organisation.org' but by using the full 'authority' within the URI there would be three variables which would almost certainly be highly dependent upon each other. Consequently we have found it desirable to reduce further to the domain name part of the URI rather than the full authority, and such as 'organisation.org' would be used instead.

E. The Result of the Principal Component Analysis

After the PCA has been performed the result will be a matrix of the format *principal components* x *variables.* Each variable will have a Principal Component that it correlates with the most and consequently we then assign the variables with the maximum correlations to those Principal Components. After the variables have been assigned to the components it is possible to process the timeline replacing each of the URIs with the Principal Component number.

III. Case Study

A. The Test Data

To demonstrate the PCA analysis of Internet history we provide two case studies for comparison. Data set 1 comes from the Digital Corpora project ([13] [11]) and is a 'simulated' set of data, in that the data is from a real system and has a real user interacting with the system, but the parameters of that usage is a scenario. Data set 2 in comparison is from a 'real world' set of Internet history artefacts and shows a real user interacting with a system performing their day-to-day leisure, work and study activities.

Data set 1 is a forensic image that comes from the M57-Patents scenario referred to as 'jo-2009-11-20-oldComputer' and is an EnCase image of a 12.1GB NTFS formatted hard disk drive containing an installation of Windows XP SP3. The Internet history timeline was constructed from the 'Comprehensive Search' for Internet History using EnCase 6.19 and is based upon approximately 820 Mozilla artefacts dated between the 13th and the 20th of November 2009. The timeline was processed to produce 60 domain name variables and 580 distinct event cases from the original 820 event artefacts.

We constructed data set 2 from a real-world set of Internet history artefacts recorded from Microsoft Internet Explorer Version 10 on a Windows 8 PC over the period of a weekend. The artefacts were extracted and the timeline constructed directly from the WebCache database and contains 5900 artefacts between the 8th and the 10th of November 2013. The timeline was processed to produce 139 domain name variables and 1339 distinct event cases from the original 5900 event artefacts.

When combining timeline sources, such as from two different types of web browsers, care must be taken to ensure that a source which produces a large quantity of artefacts per time period, such as the Microsoft Internet Explorer web browser typically appears to record more artefacts than the Mozilla Firefox web browser. To ensure this does not skew the data, the correlation matrix is used during PCA rather than a covariance matrix. As we are assuming there is correlation between artefacts in our analysis, oblique rotation is preferable for the PCA and we have chosen the 'Direct Oblimin' algorithm using SPSS version 20.0.

B. Number of Principal Components

For data set 1 the Scree plot (Figure 1) shows an initial elbow at 3 values, with a steady decline in Eigenvalues until 15 components are reached at which point the graph flattens off until 45 components is reached, where a substantial fall-off is observed. For data set 2 (Figure 2) the Scree plot shows a much less obvious elbow and more of a concave shape. The data appears to slow its decline at 10 Principal Components rounding off around 20 components.

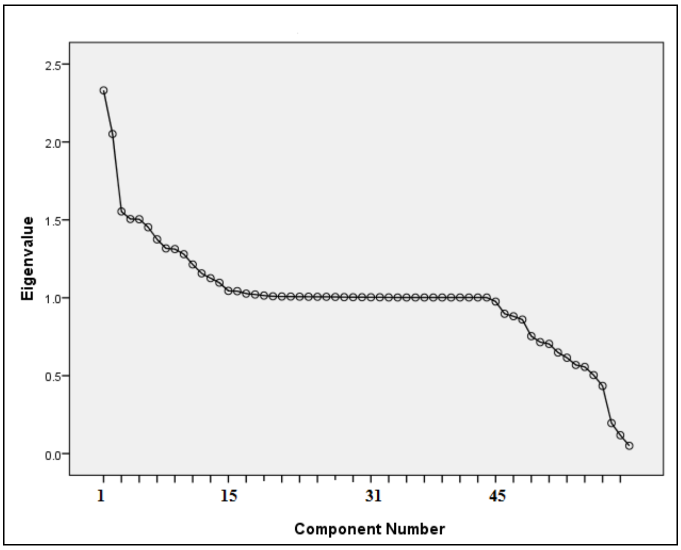


Figure 1 - Scree Plot of Principal Component Analysis on data set 1

Figure 1 shows that from the initial 60 domain name variables that 10% of variance can be accounted for within 3 variables, or more accurately 3 Principal Components. Increasing the number of components to 15 and will account for 35% of the total variance in the data set, but more components than that only increases variance by a small linear amount. Figure 2, although showing a larger data set of 139 variables is quite similar with the first 3 Principle Components accounting for approximately 9% of the variance. For this paper we have selected 20 Principal Components, partly based upon the shape that can be seen in Figure 2 at 20 components, but also that represent 35% of the data variance, which is the same variance level as we used for data set 1.

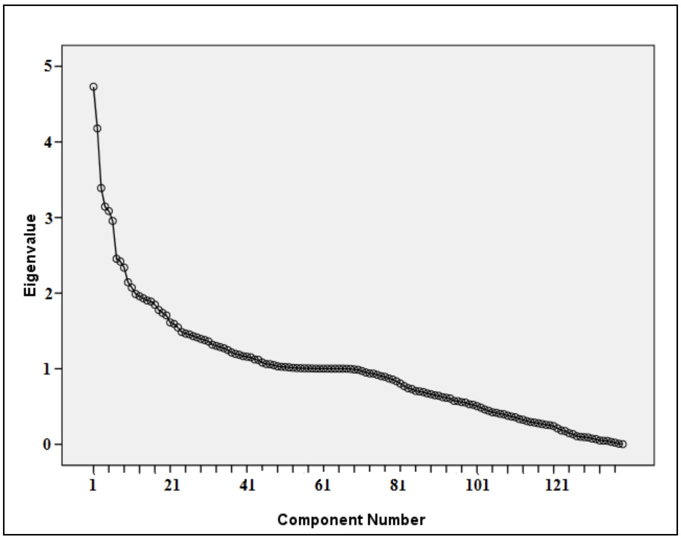


Figure 2 - Scree Plot of Principal Component Analysis on data set 2

The selection of the number of Principal Components to use and the variance in the data is an issue of substantial further research within this project.

C. *Processing the Sessions*

After each of the variables has been assigned to a Principal Component the timeline is processed to reassign the variables to the associate Principal Component number. At this stage artefacts are grouped into sessions of contiguous activity and the membership of the sessions are analysed. A period of contiguous access is classified as when there are temporally grouped point events on the timeline that are delimited by a time period of greater than threshold value X. For this case study 15 minutes has been chosen as the threshold X value.

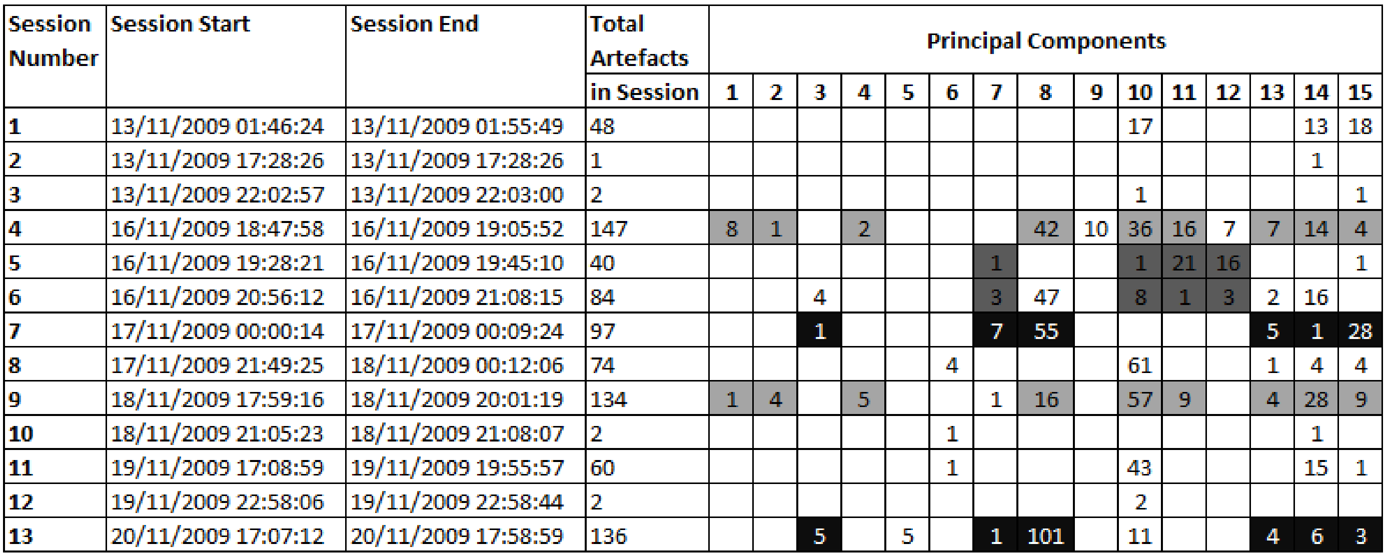
Using a 15 minute threshold with data set 1 we have 13 sessions. Data set 2, which was the larger, but more densely packed, set of data makes 7 sessions. The results of this session-level analysis can be seen in Tables 1 and 2. Both sets of data have two sessions that are quite close to the threshold value, 22 minutes and 25 minutes respectively, coincidentally occurring between sessions 4 and 5 in both data sets. There is an argument for extending X to cover this period but for this paper the sessions are kept separate. All the other session are delimited by substantial gaps.

D. *Analysis of the Sessions*

Table 1 and Table 2 show the results of grouping the Principal Components into the sessions of contiguous activity. For each sessions we see the start and end time of the session, the total number of artefacts that appeared in that session and the artefacts group per Principal Component per session. Shading of the Principal Components is used on the tables to indicate possible patterns of similarity.

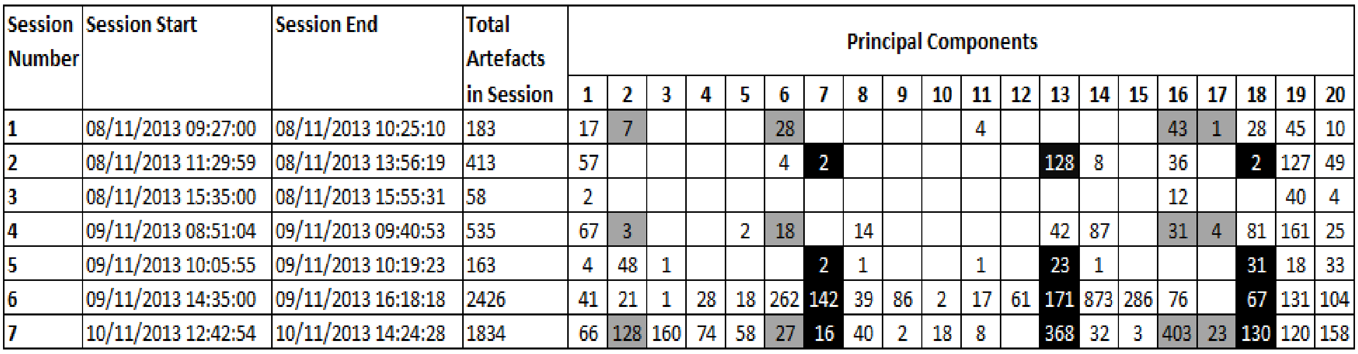
It can be seen that some Principal Components appear very regularly in the sessions. In data set 2, in Table 2, we can see that Principal Components 1, 19 and 20 occur in every session. In data set 1, Table 1, the effect is less pronounced, however it can be seen that Principal Components 10, 14 or 15 appear in every session.

The Principal Components per session-level view provides a compact way to view like-for-like comparisons of sessions. Some sessions seem broadly similar in composition of the constituent Principal Components to others, such as can be seen in Table 1 where sessions 4 and 9 have broadly similar numbers of artefacts in the session and the Principal



Session Number	Session Start	Session End	Total Artefacts in Session	Principal Components														
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	13/11/2009 01:46:24	13/11/2009 01:55:49	48										17				13	18
2	13/11/2009 17:28:26	13/11/2009 17:28:26	1														1	
3	13/11/2009 22:02:57	13/11/2009 22:03:00	2										1					1
4	16/11/2009 18:47:58	16/11/2009 19:05:52	147	8	1		2				42	10	36	16	7	7	14	4
5	16/11/2009 19:28:21	16/11/2009 19:45:10	40							1			1	21	16			1
6	16/11/2009 20:56:12	16/11/2009 21:08:15	84			4				3	47		8	1	3	2	16	
7	17/11/2009 00:00:14	17/11/2009 00:09:24	97			1				7	55					5	1	28
8	17/11/2009 21:49:25	18/11/2009 00:12:06	74						4				61			1	4	4
9	18/11/2009 17:59:16	18/11/2009 20:01:19	134	1	4		5			1	16		57	9		4	28	9
10	18/11/2009 21:05:23	18/11/2009 21:08:07	2						1								1	
11	19/11/2009 17:08:59	19/11/2009 19:55:57	60						1				43				15	1
12	19/11/2009 22:58:06	19/11/2009 22:58:44	2										2					
13	20/11/2009 17:07:12	20/11/2009 17:58:59	136			5		5		1	101		11			4	6	3

Table 1 - Data Set 1 Session Analysis of the Principal Components



Session Number	Session Start	Session End	Total Artefacts in Session	Principal Components																			
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	08/11/2013 09:27:00	08/11/2013 10:25:10	183	17	7				28					4					43	1	28	45	10
2	08/11/2013 11:29:59	08/11/2013 13:56:19	413	57					4	2						128	8		36		2	127	49
3	08/11/2013 15:35:00	08/11/2013 15:55:31	58	2															12			40	4
4	09/11/2013 08:51:04	09/11/2013 09:40:53	535	67	3			2	18		14					42	87		31	4	81	161	25
5	09/11/2013 10:05:55	09/11/2013 10:19:23	163	4	48	1				2	1			1		23	1				31	18	33
6	09/11/2013 14:35:00	09/11/2013 16:18:18	2426	41	21	1	28	18	262	142	39	86	2	17	61	171	873	286	76		67	131	104
7	10/11/2013 12:42:54	10/11/2013 14:24:28	1834	66	128	160	74	58	27	16	40	2	18	8		368	32	3	403	23	130	120	158

Table 2 - Data Set 2 Session Analysis of the Principal Components

Components are very similar, although there are conspicuous differences between sessions 4 and 9 in the length of the sessions.

In Table 1 we do also see other complex patterns, such as sessions 7 and 13 and also to some lesser extent sessions 5 and 6. However in Table 2 we do not see the complex patterns that can be seen in Table 1, rather we see simpler shorter patterns. In Table 2 we have shown a possible pattern in sessions 1 and 4, another possible pattern in session 2, 5 and 6 and finally in session 7 we see the possibility that both of these two patterns are overlapping.

Although it is possible to do basic pattern analysis on sparsely populated sets of data as can be seen in Table 1, an enhanced approach to analysing the sessions in greater depth would appear to be a next step in the research, especially when dealing with the modern, more densely populated sessions that we can see in Table 2.

IV. Discussion

A. *Data Sets and the Case Study*

Although the two data sets were not selected or designed to represent any specific pattern or behaviour, our approach does reveal patterns over a period of time. This would suggest that this approach may not be applicable to all kinds of forensic investigations such as incident response where there is only a short specific period of time and the holistic view of the system and the users and typical usage is less of a concern. An overview of a system would potentially be very interesting where there is habitual behaviour that can be extracted over a period of time, which might be the case in examinations where there is lawful access to a system to perform unlawful activities, the classic case of indecent photographs investigations.

The principal difference that can be seen between data sets 1 and 2 is the much higher density of artefacts in the session within data set 2, and consequently sessions 6 and 7 on Table 2 are broadly meaningless for creating patterns due to the large number of artefacts in those sessions. This strongly supports the need for low-level patterns at the next stage of the research.

B. *Enhancing the Approach*

A more sophisticated low-level event modelling approach may be desirable for the analysis of the Principal Components, similar to the approach shown in Abraham [1]. As such it is possible to build patterns of Principal Components and of the Intervals between the components. In Figure 3 it can be seen that there are three Principal Components (PC_1 to PC_3) point events on a timeline which are separated by Intervals (i_1 and i_2). The intervals may play a crucial part of a pattern, such as longer intervals at certain times of day etc. The reduction of the data into Principal Components suggests that pattern analysis such as proposed in Abraham along with the time boundary intervals such as Gladyshev and Patel [4] noted may be useful. Further research will hopefully provide a quantification of the regularity or the uniqueness of any low-level patterns on the system.

The approach proposed in this paper for Internet history analysis of the timelines using Principal Component Analysis allows an investigator to identify points in the timeline that are of potentially greater interest and should be followed up specifically through interviewing, keyword matching or file matching using known signatures. For example, an investigator identifying the download of an unauthorised or illegal file to a system that is exhibiting a regular and complex pattern is unlikely to be a one-off user. An investigator could determine the overlap of components that contain 'identification information', information that could reasonably belong to a specific user of the system, and components that contain suspect material.

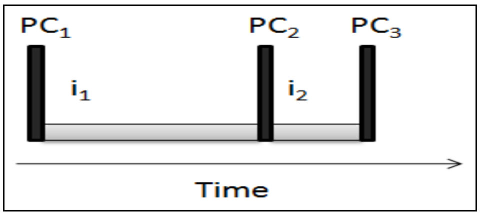


Figure 3 - Principal Components and Intervals on a Timeline

C. Kaiser-Meyer-Olkin Sampling Adequacy

The Kaiser-Meyer-Olkin (KMO) sampling adequacy test [8] literature highlights that data that produces KMO of less than 0.5 is unsuitable for Factor Analysis. In this case study, data set 1 has a KMO value of 0.334, and data set 2 has a KMO value of 0.503. This would suggest that from a statistical point of view that PCA may not be suitable or a barely sufficiently powerful approach to the analysis of Internet History. However, experimentally it can be seen that PCA does appear to be successfully grouping related artefacts that are regularly occurring in the Internet history but with a low statistical power rating due to the large quantity of the data that is 'one off' or infrequently occurring, which is still potentially interesting data but might under normal circumstances be considered outlier data.

D. Variance in the Data & PCA Sampling

In data set 1 using 15 Principal Components approximately 35% of the variance is accounted for, which is to say that 60 variables are being adequately represented by 15 components, which appears very effective with Principal Component 1 having three variables with correlation values of 0.955, 0.947 and 0.706 respectively, however Principal Component 14's maximum values are 0.269 and 0.253 respectively. Of the 60 domain name variables selected in the data set 1 case study, 21 of the variables have maximum correlations that are less 0.2 when choosing to use 15 Principal Components in the analysis. In data set 2, 40 of the 139 variables have maximum correlations that are less than 0.2 when choosing to use 20 Principal Components. It may be necessary to identify and suppresses variables that have a correlation with a Principal Component of less than X, where X is a value that further research will have to establish.

V. Conclusions

This paper proposes a novel method of analysing Internet history artefacts typical to Digital Forensic investigations using the unsupervised statistical learning technique Principal Component Analysis. The findings we show in this research paper are promising for identifying, reducing and modelling the Internet history of a user's behaviour.

After demonstrating a high-level tabular view of simple patterns for the analysis of the Internet history, further work includes the development of low-level rule-based analysis on the Principal Components, more sophisticated methods of analysis and additional types of interactive user logs, such as chat logs, file system and operating system events related to the behaviour of the system or user.

References

[1] T. Abraham, “Event sequence mining to develop profiles for computer forensic investigation purposes,” in Proceedings of the 2006 Australasian workshops on Grid computing and e-research, 2006, vol. 54, pp. 145–153.

[2] C. Boyd, P. Forster, "Time and date issues in forensic computing - a case study", Digital Investigation (2004) 1, 18-23.

[3] R.B. Cattell, "The Scree Test for the Number of Factors", Multivariate Behavioral Research, Vol.1, Issue 2, 1966.

[4] P. Gladyshev and A. Patel, “Formalising Event Time Bounding in Digital Investigations,” International Journal of Digital Evidence, vol. 4, no. 2, pp. 1–14, 2005.

[5] K. Guðjónsson, “InfoSec Reading Room Mastering the Super Timeline With log2timeline.” p. 84, 2010.

[6] C. Hargreaves and J. Patterson, “An automated timeline reconstruction approach for digital forensic investigations,” Digital Investigation, vol. 9, pp. S69–S79, Aug. 2012.

[7] R. K. Henson, J. K. Roberts, "Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice", Educational and Psychological Measurement 2006; 66; 393.

[8] G.D. Hutcheson, N. Sofroniou, "The Multivariate Social Scientist: Introductory Statistics using Generalized Linear Models", Sage, 1999.

[9] A. Marrington, “Computer Profiling for Forensic Purposes,” Queensland University of Technology, 2009.

[10] J. Oha, S. Lee, S. Lee, "Advanced evidence collection and analysis of web browser activity", Digital Investigation 8 (2011) S62-S70.

[11] K. Woods, C. Lee, S. Garfinkel, D. Dittrich, A. Russell, K. Kearton, "Creating Realistic Corpora for Forensic and Security Education", 2011 ADFSL Conference on Digital Forensics, Security and Law.

[12] "ACPO Good Practice Guide for Digital Evidence", Version 5, Association of Chief Police Officers, 2012. http://www.acpo.police.uk/documents/crime/2011/201110-cba-digital-evidence-v5.pdf

[13] Digital Corpora Project, http://digitalcorpora.org/corpora/scenarios/m57-patents-scenario

[14] 'plaso', http://plaso.kiddaland.net/

[15] 'Aftertime', http://www.forensicswiki.org/wiki/Aftertime