Towards Web Usage Attribution via Graph Community Detection in Grouped Internet Connection Records

David W. Gresty, George Loukas, Diane Gan and Constantinos Ierotheou
CSAFE Group
Computing and Information Systems
University of Greenwich, UK
Email: D.Gresty@Greenwich.ac.uk

Abstract—**Internet connection records can be very useful to digital forensic analysts in producing Internet history timelines and making deductions about the cause and effect of activity. However, the available data may include only a subset of the data that would be available from physical extraction. For example, the new UK legislation allows the collection of host website details, time of access and subscriber details, but not the specific uniform resource locator visited. Here, we investigate how to process data from Internet connections records to extract the websites, and construct the sessions of activity that are likely to be idiosyncratic the individual users, from the set of multiple possible users. We demonstrate how to display Internet history sessions as a network and perform graph community detection, showing a scheme for breaking up the component parts of the Internet history sessions into groups. We also introduce the use of websites' relative popularity for identifying websites that are likely to be meaningful to particular users of particular devices, further improving the accuracy of attributing a particular activity session to a particular user at a particular point in time.**

I. Introduction

Recovered Internet history record artefacts are particularly useful in digital forensics. By placing them into a timeline or a "super timeline" [7] including files and pictures, a forensic analyst can make deductions about the cause and effect of a suspected user's behaviour. Artefacts from the Internet history are also easy for the general public to understand and are regularly reported upon in the press. For example, during and after the trial of Vincent Tabak for the murder of Joanna Yeates [24], [25], it was reported that he had typed relevant search terminology, viewed map locations that corresponded to the location where the body of Joanna Yeates was recovered and viewed pornographic pictures that were described as resembling its appearance and condition. Such artefacts can be persuasive to a Judge and Jury as they may be able to relate to the artefacts and their everyday use of computers.

Physical device extraction can be costly in terms of resources, causing substantial backlogs [10]. Also, due to the use of "private browsing", as well as the normal overwriting of information, particularly on smaller capacity devices, it does not guarantee full availability of Internet history. National legislation may allow law enforcement to access a suspect's Internet connection record history as retained by the communication service provider. For example, recently introduced legislation in the UK [27] allows access to retained host website details, the time of access and subscriber details (but not the full Universal Resource Locator (URL) of the exact page that was visited). An investigator may be able to acquire and analyse Internet activity before a physical device has been examined, seized, and even potentially before a suspect is first interviewed. Whilst this new legislation was discussed within the UK, Lord Keen of Elie, The Advocate-General for Scotland stated that Internet Connection records "should not be acquired for trivial purposes" and that they should be available for "the investigation of any offence where the sending of a communication is an integral part of the offence: for example, offences related to stalking, cyberbullying and harassment which can, if not investigated, quickly escalate to more serious offences" [26]. Internet Connection Records would not be restricted to serious and major crimes, counter-terrorism or matters of national security, but rather they would be relevant and available to the day-to-day work of front-line detectives and investigators, as part of modern communication-enabled crime investigation.

Much as itemised telephone billing can only tell which number called which number and cannot tell who was on those telephones or what was said, Internet connection records can tell which machine contacted which website, but not who was sat at the keyboard or indeed which pages or content was actually viewed. Unlike the Tabak trial, Internet connection records will not be able to tell which search terms were used, pages visited and whether relevant maps were viewed. What Internet connection records can show are "sessions" of activity [6], that are periods of Internet access and activity that are delimited by sufficiently long idle periods. These sessions can be used to provide context for evidential artefacts that will be used to persuade court to make a particular decision, but can also be used to drive investigative lines of enquiry. As evidential context, finding what websites were being visited at the time that a suspicious artefact's metadata says it was created onto a device is important for showing the intent of a user. Alternatively, the context of the session can show whether the activity at a particular time was routine, commonly occurring or if it was a 'one-off' event. As a driver of investigative reasoning, context can allow interviewers to ask targeted questions about "who uses the computer in the afternoons?" or "who likes this football team?" if they have identified some time or feature which they can relate or associate with the crime or artefacts.

Within this paper, we demonstrate the usefulness of session-to-session analysis on the type of data an analyst has access to from the Internet connection records, including dates, times and host-name addresses. These kinds of records could belong to any of the users on any of the devices at an address and consequently, we need to presume a multi-user dataset where there is no clear demarcation between who is the user of a device at any time. This paper's contributions are the following:

- We investigate how data from Internet connection records can be processed to extract the websites and sessions that are likely to be idiosyncratic to one individual user (out of a set of candidates)
- We demonstrate how to display Internet history sessions as a graph and perform graph community detection, showing schemes for breaking up the component parts of the Internet history sessions into groups.
- We introduce the use of websites' Relative Popularity for identifying websites that are likely to be meaningful to particular users of specific devices.

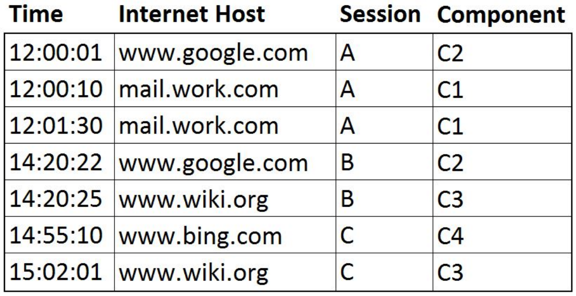
II. Related Work

The artefacts and potential evidence contained within Internet Connections Records and temporal history recovered from devices is substantial. However, the ability to process and visualise the data is not trivial. There has been much research on analysing temporal sequence data from the file systems, operating systems and applications. Zeitline, was introduced by Buchholz and Falk in 2005 [3] and was updated in 2014 with new features added by Inglot and Liu [11]. Its purpose was to reconstruct artefacts to enable an investigator to create complex events from the individual artefacts, using searching and filtering to populate and analyse a timeline.

Different applications, such as the different web browsers and operating systems leave individual footprints of their activity. The approach by Khan and Wakeman [15] is to determine the footprint of applications on a system based upon the typical artefacts that are created in normal usage. These features are then used to train a neural network which could be used during a forensic examination to attempt to reconstruct a timeline of events showing when applications were used. In 2009, the Cyber Forensic TimeLab (CFTL) tool was proposed by Olsson and Boldt in [20]. CFTL searches for known artefacts to produce a histogram timeline. It does not automatically analyse the artefacts, but requires the analyst to make a visual correlation of different timelines overlaid to display clusters. A significant tool for organising and examining artefacts that can be arranged in a temporal sequence is log2timeline, which was reported in [7], and the subsequently updated version plaso [28]. This tool creates a super-timeline, arranging all the File System, Operating System and application logs into a monolithic list which can then be processed using filers, rules etc. This approach was highlighted by Carbone and Bean in the review of timeline creation utilities, but in their view too many irrelevant files are included [4]. Hargreaves and Patterson developed a tool to reconstruct high-level events from low-level activity using temporal proximity pattern matching [9].

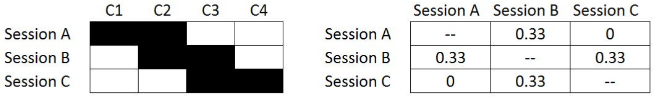
The cause and effect nature of event reconstruction has been studied, and James and Gladyshev [13] have defined action instances, a state transition model where an action produces a trace. If traces can be identified, then actions can be implied because of the causal nature of certain state transitions on computer systems. The idea of cause and effect can also be seen in Marrington's computer profiling [18], where the system provides interpretation of events and the identification of unknown events. Khatik and Choudhary have developed a timeline visualization tool [16], which integrates log files from web servers, searches for known patterns of activity that may be of significance and uses this to reconstruct the timeline of the system's operation.

The above timeline analysis tools are effective for assisting in forensic investigations, but are limited to the identification and presentation of known patterns of events or to highlighting patterns that stand out as not belonging to known patterns of events. Outside of this traditional area of digital forensics investigations, there is interesting research into event reconstruction, management and display. Large event sequences can be reduced and simplified for viewing as can be seen in Kiernan and Terzi [17]. This allows an analyst a global view of the activity but allows detection of suspicious activity. The authors have proposed techniques for the analysis of large-size audit logs, which need to be digested for display to an investigator. Eagle and Pentland [5] asserts that people have structures, routines and patterns of behaviour, which when spatially, temporally and even socially contextualised can be easily identified. The authors term these underlying principal component-like behaviours as *eigenbehaviours*. Schaefer et al. [21] describes event sequences and makes some notable distinctions between the time-synchronous events, and between aggregate events. The authors show different ways to visualise clusters of events and highlight gaps and show event information, which do not need to be timelines. Al Awawdeh et al. [1] show a real-time agent approach for recording data as it happens, which differs from traditional forensic approaches which are post-mortem style, post-event forensics. The authors discuss the problem of verbosity, which is the issue that unimportant details can be over-reported in logs and salient details can be reported but are not given adequate prominence. Hamid et al. [8] describe events as the interaction between "animate and inanimate objects" and highlight that the area of *activity discovery* is for the identification of repetitious patterns within sequences of data. The authors show that sequences of behaviour can be constructed from a variety of timeline data, such as sensors within a home showing someone moving from the kitchen to the stairway etc. These patterns can be indicative



Time	Internet Host	Session	Component
12:00:01	www.google.com	A	C2
12:00:10	mail.work.com	A	C1
12:01:30	mail.work.com	A	C1
14:20:22	www.google.com	B	C2
14:20:25	www.wiki.org	B	C3
14:55:10	www.bing.com	C	C4
15:02:01	www.wiki.org	C	C3

Fig. 1. Internet History showing the Component and Session details



	C1	C2	C3	C4
Session A	■	■		
Session B		■	■	
Session C			■	■

	Session A	Session B	Session C
Session A	--	0.33	0
Session B	0.33	--	0.33
Session C	0	0.33	--

Fig. 2. The Jaccard similarity coefficient between three sessions

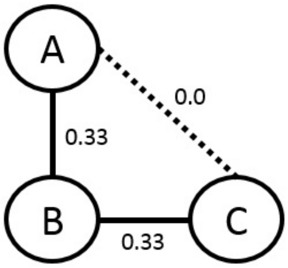


Fig. 3. The value on each edge is the Jaccard similarity coefficient value for each pair of sessions. The dashed lines illustrate where the the value is below a threshold t.

of the time of day and the individual. Minnen et al. [19] describes *motifs* as sub-sequences within a longer sequence of data. The principal problem with motif discovery is that the length, shape, size and scale of them are not known in advance.

III. Session-to-Session Visualisation

Gresty et al. [6] have shown how an Internet history timeline can be broken up into sessions, facilitating session-to-session analysis, where the forensic analyst looks at whether the same website appears in two different sessions, or intra-session analysis, where the focus is on how frequently or in which order particular websites have been accessed during a specific session. Our focus here is on session-to-session matching, where the presumption is that if activity in one session is sufficiently similar to activity in another session, this is likely to be a consequence of the web usage behaviour of the same user. In situations where the user has a broad range of options in how to behave, high similarity between sessions will be the result of the choices made by the user, and as such those two sessions will have a greater likelihood of having been made by the same user. We examine below how reasonable it is to presume that different users behave sufficiently differently given a web browser and no direction.

To measure the session-to-session similarity, we have used the Jaccard similarity coefficient [12], which is the size of the intersection of two sets divided by the size of their union. In this context, it is the sum of the components shared between two sessions divided by the total number of components in either session.

For example, taking Internet history in figure 1 and converting this to the session and component table in figure 2, if session A contains components C1 and C2, and Session B contains C2 and C3, then the overlap is C2, which corresponds to a Jaccard similarity coefficient of 0.333. For sessions B and C, it is again 0.333, but note that this cannot lead to any inference of similarity between sessions A and C, which share no components between them.

We can visually represent session-to-session comparison as nodes connected by undirected edges, as in figure 3, where the 0 value between A and C is shown as a dashed line. We can illustrate this by showing with a dashed line any relationship with Jaccard similarity below a threshold value t. We can see on figure 4 that the lower the value of t the greater the number of incorrect matches, but there are incorrect matches at high values of t, including exact 1.0 matches. Therefore only using a simple threshold value is insufficient for reducing error. We therefore propose a method of grouping our data which we refer to as Relative Popularity.

IV. Relative Popularity

Without any prior knowledge about the users, their interests or the types of websites they like to visit, we can compare the Local Popularity (LP) of components to external Global Popularity (GP) metrics. The LP is the total number of sessions that a component appears in. GP measures some level of impact assessment, which could be link-based algorithms that identify how well referenced sites are by other sites, such as PageRank, HITS and CLEVER [14], or the impact assessment could be based upon the analysis of the volume of web traffic, such as in Alexa Internet [22]. Here, we use the GP rank metric from the Alexa Internet Traffic Rank, which provides a global metric for a substantial number of websites. The data in our experiments is UK-based. Regional differences are described in the future work section of this paper.

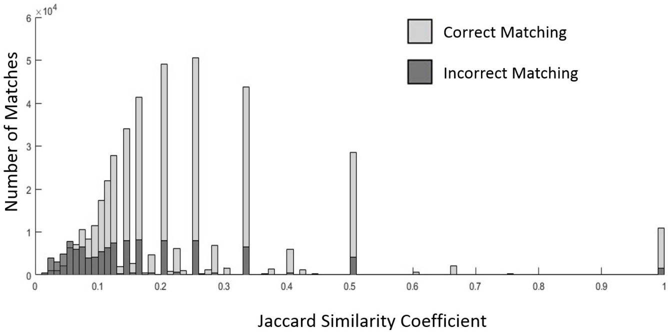


Fig. 4. Correct and Incorrect matching Session-to-Session comparisons

The GP data provided by Alexa Internet is already provided in rank order and the LP data can simply be converted to rank order. The difference between LP Rank and GP rank can be computed to determine whether the result has a Low or High difference between the Local and Global popularity. For simplicity, we have used numerical difference of the rankings, but one could also experiment with ranking ratio or more complex metrics. Four basic conditions can be inferred from this about the components:

1) *High Difference: Low GP, High LP*. This is the potentially idiosyncratic websites that are sufficiently niche that they have low GP, but are visited by a user with sufficient frequency that they immediately stand out as interesting to the analysis.
2) *High Difference: High GP, Low LP*. These are sites that are Globally popular but a user has rarely visited them. This condition would be typified by someone that has rarely used a particularly popular service, such as a user not having a significant social media footprint, but occasionally following links onto a social media site.
3) *Low Difference: High GP, High LP*. This condition is where a user on the device is a regular user of a globally popular website, such as search engines or social media sites. This condition however is not irrelevant as it may be that in a scenario where multiple users have access to the same device, one user may have preference to the use of one social media site while the other person is a user of a wholly different social media site, or even not at all.
4) *Low Difference: Low GP, Low LP*. This corresponds to infrequent viewing of fairly niche websites.

We can see from the above four conditions that high LP is always significant, principally because session-to-session analysis is an analysis of repetitive behaviour, and the more repetition of behaviour the better. Conditions 1 and 3, a High and a Low difference conditions are both therefore likely to be significant in the analysis of the Internet history, but they both represent different types of behaviours. Because of low LP, conditions 2 and 4 do not occur with enough frequency to provide a substantial number of patterns for identifying behaviour.

Condition 1 components are interesting because they indicate regularity of access to sites above the norm for the global population. We term these as ‘idiosyncratic’ giving the investigator clues to the users’ interests, hobbies, type of work, etc. These kinds of activities may overlap, or indeed may be mutually exclusive. A person may have various modes of operation, for example their ‘work mode’, ‘social media mode’, ‘pornography viewing mode’, etc. These modes may be considered part of a pattern of life for the user in that they are distinct activities that can occur at different times (and places).

Condition 3 components are not insignificant or uninteresting although they are not ‘idiosyncratic’. The components matching this condition may contain behaviour which is more difficult to distinguish from user to user because all the users of a device may overlap, for example, using the same search engine or social media site. Combining these groups with other aspects of the pattern of life such as time of day, day of week, location, duration etc. an analyst may be able to distinguish the behaviour.

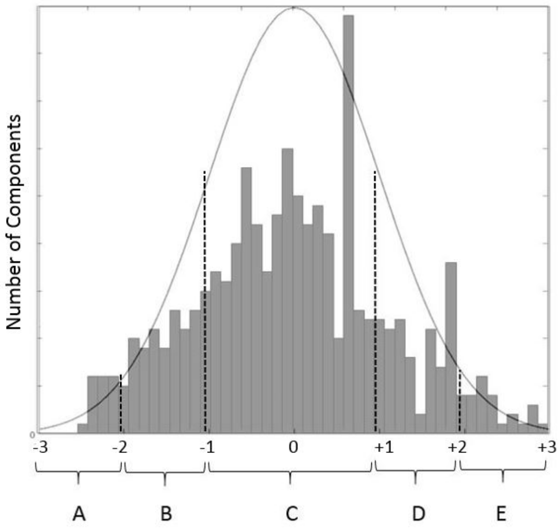


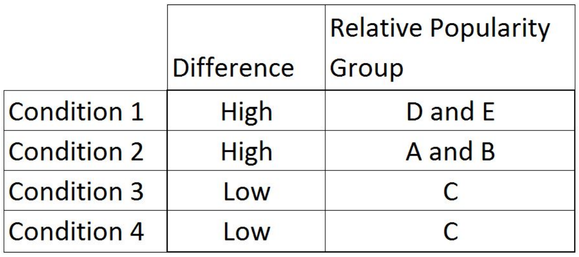
Fig. 5. The D1S1 Dataset plotted against the Standard Distribution curve to illustrate the relative similarity.

We can illustrate relative popularity by plotting the difference between the LP and GP ranks on a histogram, such as seen in figure 5, and then group based upon some threshold value. We have performed our experiments using standard deviation to group the Relative Popularity data. We can see in figure 6 that conditions 1 has been categorised into groups D (+1 to +2 Standard Deviations) and E (+2 or more), where E is the greatest difference between the LP and the GP ranks. The results for condition 2 are similarly divided for the negative high difference in groups A and B. There are notable spikes in the D and E groups and to some lesser extent some above-the-curve behaviour in the A and C groups. Therefore this suggests there may be interesting idiosyncratic data contained in D, E and perhaps also in A and some components in C.

In this paper we have not investigated the difference between group D and group E, but rather considered them together as group DE, which is the grouping for the high difference Condition 1. Similarly for Condition 2 it is considered as group AB, and Conditions 3 and 4 as group C.

V. Community discovery

We can graphically show our findings and we can perform community detection on those graphs to find the Internet history that is clustered together. A community is where nodes in a network can be grouped into clusters such that each set of nodes is densely connected internally, with sparser connections between other groups. We use the Modularity detection implemented within the Gephi software [23] based upon the algorithm presented in Blondel et al. [2]. Modularity



	Difference	Relative Popularity Group
Condition 1	High	D and E
Condition 2	High	A and B
Condition 3	Low	C
Condition 4	Low	C

Fig. 6. The Relative Popularity Conditions and their relationship to the grouping from Figure 5.

measures the density of edges inside a community and the density of the edges outside the community.

Community discovery for small interconnected clusters of sessions, such as can be seen in figure 12, is straight forward. Indeed, the majority of nodes in large interconnected clusters clearly belong to one particular community as long as they do not sit on the boundary between two clusters of nodes. For example, in figure 2, if we had to associate Session B with either Session A or Session C, then there is no clear clue where the boundary would be decided. The reality in figure 2 is that those three nodes would be placed into the same community, and as such it illustrates the problem that if sessions are in the same community then there is a relationship, but caution must be used by an analyst when saying they are 'the same'.

A. *Utilising Relative Popularity*

In this paper, we illustrate the Relative Popularity method. The 'D1S1' dataset is Internet history data of approximately 1500 sessions, 800 sessions belonging to user 1 and 700 sessions to user 2. This approach can be used for multiple users, however for simplicity we have used only two for these experiments. The set has been constructed from two different sources, such that experimentally we know the ground truth of which session-to-session matches are correct. There is a 12.89% overlap in components between the two users. The sessions are constructed using the variable-length approach and a 15 minute idle time to delimit sessions (as per Gresty et al. [6]). Both users are from the UK, both taken from combined home/work computers and there is no relationship or special reason to expect the two users would have shared common interests (i.e. the original users are unknown to each other). We have not attempted to profile the individuals based upon personally identifiable characteristics such as age, gender, education etc., although this may be an interesting line of future research inquiry.

For the experiments we illustrate in this paper, we have used the threshold of $t = 0.5$ for the DE, AB and C groups, which is to say if there is a Jaccard similarity between the two sessions (a correct or incorrect match) of 0.5, an edge is drawn to connect the two session nodes. For all these experiments non-repeating components (i.e. website hosts that only appear during a single session) were removed.

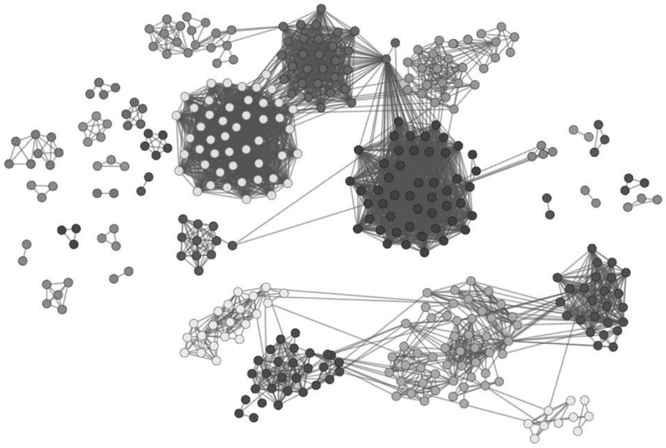


Fig. 7. The DE group two large interconnected communities top and bottom show the segregation in the behaviour of User 1 and User 2, with many smaller distinct communities left and right

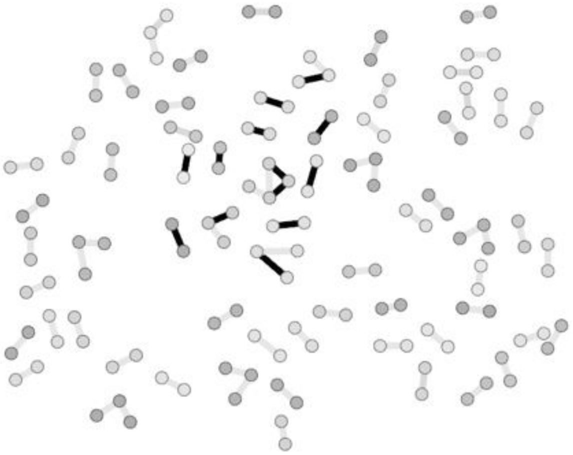


Fig. 8. The AB group - note the incorrect matching edges can be seen as the darker colour

B. *Colouring and Interpreting the Graphs*

On our graphs, we can apply a colouring scheme to the communities, which greatly simplifies the analyst's ability to quickly identify (and if necessary rearrange the layout) and determine the sessions/nodes of interest within the communities. In figure 7, we can see that for the DE group, there are two large interconnected clusters of sessions, comprising several communities and there is a collection of smaller communities which are not connected to the large clusters. These large multi-community clusters correspond to the distinct behaviours of User 1 and User 2. This kind of clear demarcation between the users' behaviour is not noticeable in figures 8 and 9 for the AB or C groups of data.

C. *Performance of Community Discovery*

The overall performance for the community detection with the D1S1 dataset can be seen in figure 10.

The AB group can be seen in figure 8. There is a large number of small communities, predominantly showing as

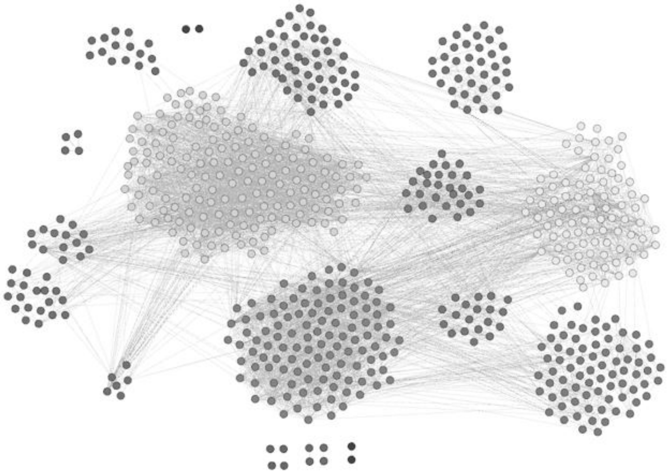
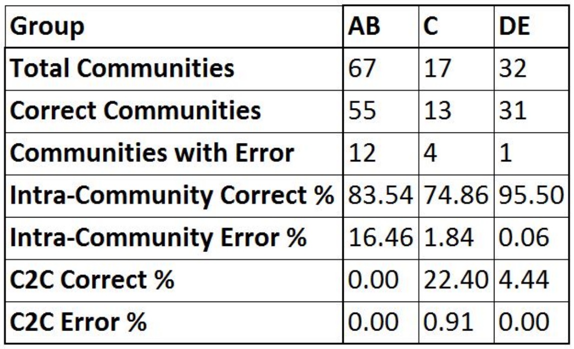


Fig. 9. The C group data showing large highly connected communities



Group	AB	C	DE
Total Communities	67	17	32
Correct Communities	55	13	31
Communities with Error	12	4	1
Intra-Community Correct %	83.54	74.86	95.50
Intra-Community Error %	16.46	1.84	0.06
C2C Correct %	0.00	22.40	4.44
C2C Error %	0.00	0.91	0.00

Fig. 10. The Number of Communities and Correct connections between communities

paired relationships. With 16% intra-community error within the AB grouping this was clearly the least effective of all the grouping schemes.

The DE group (figure 7) provided a high degree of intra-community accuracy, whilst also having community-to-community (C2C) accuracy. Overall this grouping scheme provided high accuracy and extremely low error, with meaningful C2C connections.

Within the C group of data the total intra-community error, 1.84%, does not affect all of the communities equally. Consequently, within the C grouping, the four communities that contain error contain the majority of the intra-community error and are the source and end points of the and C2C errors.

VI. Investigative uses of session-to-session graphs

Having shown that behaviour and local popularity can help identify matching sessions, we can use them towards addressing the case of "it wasn't me, it must have been someone else". If it is believed that there is a single user and what we are trying to identify is if a particular event, or set of events, could be the result of an unknown second user, then we are trying to model as much possible Internet history so as to show the regularity and normality of the activities on that device, at the times of interest. In this case, identifying whether a session is an outlier, not belonging to any community is important.

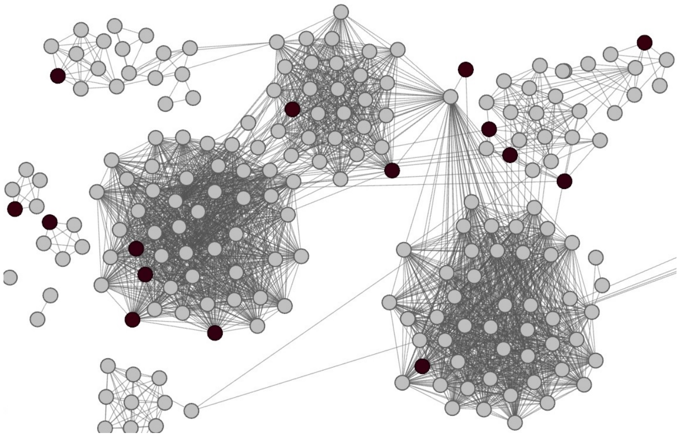


Fig. 11. The Early Morning Sessions on the DE group

If we have a situation where we have two or more sessions which we are trying to show have close association, we are less concerned with modelling as much data as possible as we are of providing the highest degree of assurance that those two sessions accurately belong to the same user. This is the case where there are sessions showing personally identifiable information, such as log-ons to private email accounts, and other sessions involved. The ways we can provide assurance that the two (or more) sessions are related is by providing as much context and showing how many groups the sessions overlap, and if any other pattern-of-life features are relevant.

In figure 11 we have shown the early morning sessions (which we have classified as starting before 0700 hours) as black-filled nodes, for the DE group. An interesting note here is that only one of the two large community clusters had early morning sessions (i.e. we identified that only one of two the users was an early riser). We can see the very left-most side of figure 11 shows two smaller five-session communities and consequently the time of day information might indicate the association with the same user as the other early morning sessions.

We can identify which of these early morning sessions belong to which communities such that figure 12 shows that we can extract the directly connected sub-graphs from within the communities. Session 60 belongs to the same community as Session 584, they are directly connected (at a strength greater than our threshold of $t = 0.5$), and in addition to that they both occur in the 'early morning' period. The same can be seen for the other two sub-graphs.

Using these directly connected sub-graphs based upon additional pattern-of-life information greatly adds to the assurance that these sessions were created by the same individual.

We can also then look at the breakdown of the components

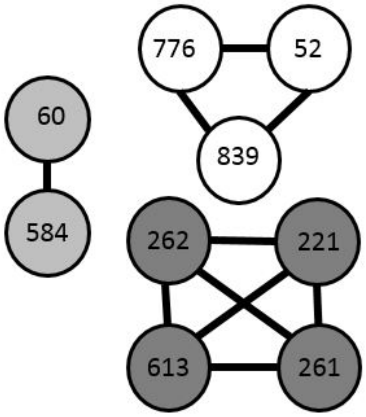


Fig. 12. The Early Morning Sessions sub-graph of directly connected communities

in figure 13. We see that component C3 was required for Sessions 60 and 584 to be members of the same community in the DE group, but both of those sessions also share component C5 from the C group. The larger sub-graph comprising sessions 221, 261, 262 and 613 all shared component C1, but we also see some sharing of C4 and C6 (and C6 also appears frequently in the next sub-graph). Additionally, the temporal sequence of Session 261 and then 262 also strengthens the relationship between those two sessions.

This now allows the analyst to focus in on a small number of interesting components that may show the identity of the user during the early morning period. By extension, as these sessions are part of a clearly interconnected group of communities, this handful of components provides us with supporting evidence for the identity of the user during a large amount of the Internet history.

VII. Further Work

A. Regional Popularity as a Consideration

Whilst assigning the external 'global' ranking it is worth considering the difference between a regional ranking versus a generic global ranking if such data is available. As can be seen at the time of writing, the overall globally most visited sites, such as the 'Alexa top 500 sites' [22], contain many regionalised versions of the same websites, such as the Google search engines or large-scale ecommerce sites such as Amazon.

A history recovered from a machine in the UK region is likely to contain artefacts relating to both the US/Generic version of a website (e.g. amazon.com) and the UK regional version of that site (e.g. amazon.co.uk). We would however expect to see the regional version to be considerably more popular on the local machine. The regional consideration becomes even more pronounced when dealing with websites for organisations or companies that exist only within the region and are not international represented. This means the overall global ranking can be considered unpopular, yet when considering the ranking within that region the site can be considered popular.

Take for example The University of Greenwich's website, 'gre.ac.uk', with ranking data from Alexa Internet: Global ranking of 45,496, UK Regional Ranking of 1,707, USA Regional Ranking of 221,930, Malaysia Regional Ranking of 3,509.

The UK regional ranking is considerably more popular than the global ranking, and if you were to view the website from another geographic region such as Malaysia or the USA there would be quite different rankings to the overall Global ranking.

Is the difference between the regional ranking of 1,707 and the global ranking of 45,496 as stark as it initially seems? We have performed correlations on datasets between the UK region data for the visited websites in this experiment and the Global regional rankings and have produced correlations of 0.98. The likely result of not using the regional GP value we have observed during our experiments is to shift the regionalised components to the right if plotted on a curve such as seen in figure 5. The sites that are popular within the region but relatively unknown globally (for example, high-street stores, or local news sites) will have a greater difference between the regional LP and the GP and as such may appear in the D group, whereas one would expect them to appear in C.

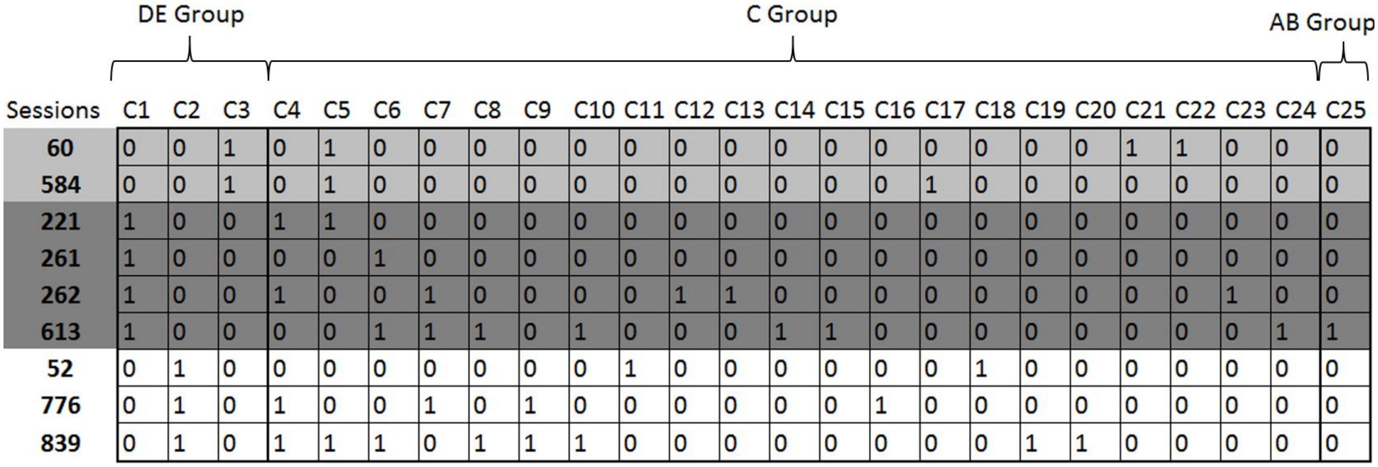
An interesting consideration is that the differences in our data based upon Regional variation does appear interesting but the high correlation between the local (UK) regional ranking versus the global (predominantly English speaking, at this time) region may be a linguistic, rather than a regional correlation. Further work in this area with Internet history for different regions and different languages would be desirable.

VIII. Conclusions

We have shown how to make use of the type of Internet connection records that will become available to UK law enforcement investigators. Indeed any Internet history record that has been recovered through traditional digital forensic examination can be processed using session-to-session analysis to provide context about the regularity of access to websites and to provide detailed associations between two or more sessions.

When trying to establish if an event or artefact present within or during a session is anomalous or a 'one-off' to support a statement that "it wasn't me!" then it is beneficial to model as much data from the Internet history as possible to see exactly how common are the components/websites visited.

If, on the other hand, the investigator is trying to establish or prove the user of a device, they will want as many supporting characteristics and they will want them at a high level of assurance. An investigator will want the data broken up to identify the idiosyncratic websites to examine (such as the high difference D and E groups of data, which we show in this paper appears to easily perform best at showing this), will want to know if that behaviour is supported by the popular frequently visited sites (low difference C groups) and will then want to start looking at whether additional patterns-of-life can be used to sub-graph any detected communities to show that times of



	DE Group			C Group																					AB Group
Sessions	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25
60	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
584	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
221	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
261	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
262	1	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0
613	1	0	0	0	0	1	1	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1
52	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
776	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
839	0	1	0	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Fig. 13. The components that comprise the Early Morning Sessions and the supporting Artefacts from the C and AB groups

day, website types or duration of session can further support the hypothesis that an event was caused by a particular user. By selecting increasingly large thresholds of Jaccard similarity and only comparing sessions within the same communities, we increase the the level of assurance that there will be low error.

We have shown that the Relative Popularity method for grouping data performs well at identifying with low error what we have determined Idiosyncratic websites. For the types of investigation where an analyst is seeking to associate two or more periods of time because one period might be related to an offence and another might contain the personal data indicating who the user was, we believe the approach in this paper is a powerful and reliable tool for assessing, visualising and analysing extensive quantities of Internet Connections and History Data.

References

[1] S. Al Awawdeh, I. Baggil, A. Marrington, F. Iqbal, "CAT Record (computer activity timeline record): A unified agent based approach for real time computer forensic evidence collection", Eighth International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE), pp. 1-8, IEEE. 2013.

[2] V.D. Blondel, J.L Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks", Journal of statistical mechanics: theory and experiment, 2008(10), p.P10008. 2008.

[3] F.P. Buchholz and C. Falk, "Design and Implementation of Zeitline: a Forensic Timeline Editor", DFRWS, 2005.

[4] R. Carbone, C. Bean, "Generating computer forensic super-timelines under Linux", SANS Reading Room, 1-136, 2011.

[5] N. Eagle, A.S. Pentland, "Eigenbehaviors: Identifying structure in routine", Behavioral Ecology and Sociobiology, 63(7): 1057-1066. 2009.

[6] D.W. Gresty, D. Gan, G. Loukas, C. Ierotheou, "Facilitating forensic examinations of multi-user computer environments through session-to-session analysis of Internet history", Digital Investigation, 16, pp.S124-S133. 2016.

[7] K. Gudjonsson, "Mastering the Super Timeline With log2timeline", SANS Reading Room, 2010.

[8] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, C.L. Isbell, "Unsupervised activity discovery and characterization from event-streams", arXiv preprint arXiv:1207.1381. 2012.

[9] C. Hargreaves, J. Patterson, "An automated timeline reconstruction approach for digital forensic investigations", Digital Investigation, 9: 69-79, 2012.

[10] B. Hitchcock, N.A. Le-Khac, M. Scanlon, "Tiered forensic methodology model for Digital Field Triage by non-digital evidence specialists", Digital Investigation, 16, pp.S75-S85. 2016.

[11] B. Inglot, L. Liu, "Enhanced Timeline Analysis for Digital Forensic Investigations", Information Security Journal: A Global Perspective, 23: 3244, 2014.

[12] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Socit Vaudoise des Sciences Naturelles 37: 547579. 1901.

[13] J. James, P. Gladyshev, "Automated Inference of Past Action Instances in Digital Investigations", International Journal of Information Security. Cryptography and Security. 2014.

[14] A. Jain, R. Sharma, G. Dixit, V. Tomar, "Page ranking algorithms in web mining, limitations of existing methods and a new method for indexing web pages", In Communication Systems and Network Technologies (CSNT), 2013 International Conference on (pp. 640-645). 2013.

[15] M.N.A. Khan, I. Wakeman "Machine Learning for Post-Event Timeline Reconstruction", First Conference on Advances in Computer Security and Forensics, Liverpool, UK. 2006.

[16] P. Khatik and P. Choudhary, "An Implementation of Time Line Events Visualization Tool Using Forensic Digger Algorithm", JCSE International Journal of Computer Sciences and Engineering, 2(4), 2014.

[17] J. Kiernan, E. Terzi, " Constructing comprehensive summaries of large event sequences", ACM Transactions on Knowledge Discovery from Data, 3(4), ACM. 2009.

[18] A. Marrington, "Computer Profiling for Forensic Purposes", Queensland University of Technology Report, 2009.

[19] D. Minnen, T. Starner, I.A. Essa, C.L. Isbell, "Improving Activity Discovery with Automatic Neighborhood Estimation", IJCAI. Vol. 7. 2007.

[20] J. Olsson, M. Boldt, "Computer forensic timeline visualization tool". Digital Investigation, 6: 78-87, 2009.

[21] M. Schaefer, F. Wanner, F. Mansmann, C. Scheible, V. Stennett, A.T. Hasselrot, and D.A. Keim, "Visual pattern discovery in timed event data". Visualization and data analysis, SPIE, San Francisco, 24 - 25 January 2011.

[22] The top 500 sites on the web. http://www.alexa.com/topsites

[23] Gephi. https://gephi.org/

[24] "Vincent Tabak 'researched killings and sentences after Joanna Yeates's death' ", 19th October 2011. https://www.theguardian.com/uk/2011/oct/19/vincent-tabak-joanna-yeates-death

[25] "Joanna Yeates murder: Vincent Tabak guilty of 'dreadful, evil act' ", 28th October 2011. https://www.theguardian.com/uk/2011/oct/28/joanna-yeates-murder-vincent-tabak

[26] Hansard, HL Deb 19th of July 2016 Vol. 774

[27] Investigatory Powers Act 2016, C.25. http://www.legislation.gov.uk/ukpga/2016/25/contents/enacted/data.htm

[28] Plaso Homepage. https://github.com/log2timeline/plaso/wiki