DIGITAL FORENSIC SYSTEM PROFILING USING CONTEXT ANALYSIS

David William Gresty

Centre for Cyber-Security, Audit, Forensics and Education
Dept. Computing and Information Systems

University of Greenwich

Thesis submitted in partial fulfillment of the requirements of the University of Greenwich for the Degree of

Doctor of Philosophy

March 2018

Declaration

"*I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me, and is not concurrently being submitted for any degree other than that of Doctor of Philosophy being studied at the University of Greenwich. I also declare that this work is the result of my own investigations, except where otherwise identified by references and that the contents are not the outcome of any form of research misconduct*"

STUDENT: ....
. DAVID W. GRESTY

SUPERVISOR: ....
. DR. DIANE GAN

SUPERVISOR: ....
. DR. GEORGE LOUKAS

SUPERVISOR: ....

. DR. CONSTANTINOS IEROTHEOU

For Elizabeth and Sarah - all my love for your tireless patience.

For all Analysts and Investigators everywhere - for your tireless patience as well.

Acknowledgements

I came to this project with the kernel of an idea based upon a number of operational forensics investigations that I had performed into people's use of computer devices. I approached Dr. Diane Gan and Dr. George Loukas with the idea that ultimately lead to this research project and I am incredibly grateful to the whole supervisory team and CSAFE group. Throughout this project, indeed throughout all of my time at Greenwich, Diane and George were interested, knowledgeable, available and a source of inspiration throughout this quite isolated and difficult project. I would also like to thank my mentor from John Moores University, Dr. Mark Taylor, who was also a great source of guidance when I started my academic career.

I would like to thank my colleagues from QCC Global and the Analysts and Detectives I have worked with during my Forensics career. You have challenged, educated and gave me a sense of perspective that was invaluable during this research and constantly made me question 'is this realistic?'.

A special mention should also go to Eritia Bosmans for her encouragement and keeping me sane during the lengthy and challenging process of producing this thesis.

List of Author Publications

Manuscripts published

- D.W. Gresty, D. Gan, G. Loukas. *Digital forensic analysis of Internet history using principal component analysis.* 15th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, 237-242, Liverpool, UK. June 23-24, 2014.
- 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:S124-S133, Elsevier, March 2016.
- D.W. Gresty, G. Loukas, D. Gan, C. Ierotheou. *Towards Web Usage Attribution via Graph Community Detection in Grouped Internet Connection Records.* CEWE, IEEE CPSCom-2017, 2017.

Abstract

Conventional digital forensic investigations search digital devices for specific events or specific artefacts that indicate a crime has occurred. This does fulfil the investigative need to identify a crime, but it does not attribute the user of that digital device when the crime occurred. If a crime occurs frequently, such as accessing unlawful pornography, or is an isolated event but is co-located in time with other frequently occurring events, such as the one-off sending of a harassing message, then there may be investigative value in processing the history of the device to determine if there are patterns of repetitive behaviour present at the times of interest.

This research project investigates the habitual use of a digital device by analysing the Internet history that can be recovered from the physical digital device, or from logs that are retained as the device is connected to a firewall or service provider. The presumption in this project is that there is zero-knowledge of the content of the web history, page content or even an accurate classification of the nature of the sites that are visited. We propose in this research that the patterns of usage themselves are a significant indicator of who the user is, or the type of usage that is being performed.

We define context analysis as the investigation not of what is contained within the artefacts, but rather the investigation of the meta-data relating to that artefact and any other similar artefacts within a proximity, be it temporal, spatial or potentially spatio-temporal. Specifically, we show in

this thesis that given suitable feature selection the context analysis we define is effective at identifying patterns of habitual behaviour, as evaluated in the case of Internet history artefacts.

We present as our major contributions: the methods of analysing periods of Internet history in contextual groups of sessions; the novel approaches to feature selection for the Internet history sessions; and the display of the results on a network graph such that techniques such as community detection can be used to automatically cluster the Internet history.

Contents

List of Figures

Chapter 1

INTRODUCTION

"Ei incumbit probatio, qui dicit, non qui nega"
"The (onus of) proof lies upon him who affirms, not upon him who denies"

Digest of Justinian (22.3), 6th Century A.D.

"Ei incumbit probatio, qui dicit, non qui negat; cum per rerum naturam factum negantis probatio"
"The proof lies upon him who affirms, not upon him who denies; since, by the nature of things, he who denies a fact cannot produce any proof."

Black's Law Dictionary, 1910

1.1 KEY CONCEPTS IN DIGITAL FORENSIC SCIENCE

Digital Forensics is widely considered the investigation and presentation of results from digital devices that could be used in a legal proceeding, to allow a decision maker to make informed decisions about the matter that is being investigated. We must therefore define the key concepts for the Digital Forensic Science, and the environment within which the concepts exist. The descriptions in this thesis focus upon the application of this research in the Common law, in the criminal courts of England and Wales or other similar jurisdictions (the general concepts are the same, although the specific acts vary in each from jurisdiction). This setting is an Adversarial rather than Inquisitorial legal adjudication, which is to say that the adversarial legal proceedings are led by a party such as the 'prosecution' in a criminal case on behalf of the state, whereas in an inquisitorial system the proceedings are instead led by the judge. The common law is based primarily upon laws and upon judge-made legal precedent, and is seen across the Commonwealth of Nations, former countries of the Commonwealth and within the USA.

Forensic Science can generally be considered the application of scientific methods and techniques to matters under investigation by a court of law. The basic principles or scientific laws of Forensic Science can be broken into laws relating to the natural world and the laws relating to the forensic analysis of artefacts.

Within the natural world, we have the Law of Individuality [58], Principle of Exchange (the Locard Principle) [29] and Law of Progressive Change [120]. The natural world laws/principles differ from the 'digital world' in that digital forensic artefacts can be exact duplicates as it is data, there is no bi-direction exchange of matter when writing a data artefact to a storage medium and although digital evidence is fragile and can easily be modified if handled without care, it is not subject to degradation over time in a realistic sense.

The 'natural world' of digital forensic science is a Computer Engineering problem when dealing with the technology, hardware, networks etc., it is a Computer Science problem when dealing with the data structures, applications, operating systems and file systems, and it is an Information Systems, indeed even a social science problem when dealing with how applications are used, why applications are used and the preferences that users exhibit with their digital devices.

The scientific laws of analysis appear to transcend the purely natural world and appear directly relevant to digital forensics, and are the Law of Comparison (or Comparative Judgment) [98], Law of Probability [97], Law of Circumstantial Facts [80].

The laws of Comparison emphasise that only like-for-like artefacts can be compared and such an analysis is only as good as the quality of the sampled artefacts. The law of probability states that identification of an artefact, definite or indefinite is made on the result of probability, and practically within digital forensics this can relate to the analyst stating that artefacts are present as the result of a particular application being used, rather than as a result of a computer virus, such as can be seen in the R v Caffrey case which involved a hypothetical Trojan Horse virus defence [11]. The law of Circumstantial facts/evidence, is to say the results of forensic analysis, have as much weight as direct evidence from a witness, as witnesses can and do make mistakes when recalling an event.

Therefore, we could define Digital Forensic Science as a subset of Forensic Science, but perhaps it would be more accurate if we consider that Forensic Science is the application of analytical methods and techniques to matters under investigation by a court of law, and there are branches underneath this with 'Natural World' and 'Digital' as specialisms within the general science.

In a general sense, a crime consists of four parts: at a time, a particular individual (or individuals), with a sufficiently criminal mindset (*Mens Rea*), perform or attempt to perform an act which is criminal (*Actus Reus*). There are exceptions to this general statement such as Strict Liability offences where the *Mens Rea* is not required to be proven, or conspiracy offences where an offender agrees to participate in some fashion, but the criminal act is performed by another individual.

The *Mens Rea* is a historic concept discussed by many authors and surveyed in [28] but is formally defined within English law under section 8 of the Criminal Justice Act 1967 [108] as:

> A court or jury, in determining whether a person has committed an offense,
>
> (a) shall not be bound in law to infer that he intended or foresaw a result of his actions by reasons only of its being a natural and probable consequence of those actions; but
>
> (b) shall decide whether he did intend or foresee that result **by reference to all the evidence, drawing such inferences from the evidence** as appear proper in the circumstances.

The emphasis added above highlights the importance of evidence and why it is crucial for determining the mental state of a defendant. Digital forensic evidence is therefore important in cases where the crime is recorded by a device, planned using a device, performed using a device, or even if a device is used after the offence has occurred to discuss or research the outcomes of the crime.

There are a range of offences where the totality of the offending occurs on the digital device. The possession [109] and distribution [111] of indecent photographs of children occurs wholly electronically between one or more parties. Offensive communications such as a string of well publicised harassments that have occurred using micro-blogging website 'Twitter' occur only on digital devices [7]. The classical example of hacking a computer to gain unauthorised access [107], where the victim of the offence is the digital device, can be seen in many well cited cases [94].

Sexual Grooming of Children over the Internet [112] occurs if a person (A) communicates (on at least one occasion) with a child (B) and "A travels with the intention of meeting B in any part of the world or arranges to meet B in any part of the world". Given that a device could be used to research places to stay, and book travel arrangements online we can see all the necessary preparatory behaviour along with the communications to demonstrate the act and mental state for such an offence.

There are instances of a computer being used to research the outcome of a crime, such as during and after the trial of Vincent Tabak for the murder of Joanna Yeates [117], [118]. He 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 had viewed pornographic pictures that were described as resembling its appearance and condition.

The temporal component of a crime is quite possibly the simplest and scientifically most rigorous part of Digital Forensic Science as substantial numbers of the artefacts that forensic investigators rely upon are timestamped, and these timestamps can be examined for correctness. With the notable exception of cases such as [10] where an analyst failed to take into account the timezone of an artefact, and wrongly concluded that Law Enforcement officers had intentionally placed artefacts on a machine after seizing it, unless some form of intentional obfuscation is used to destroy or confuse the timestamps, they are reliable forms of evidence. However, within the natural world, the reconstruction of sequences and events are non-trivial and may require the investigator to establish the precise order of the artefacts. In the well-publicised trial of the Olympic athlete Oscar Pistorius who murdered Reeva Steenkamp [114], a significant moment in the trial involved evidence about whether a pair of Jeans trousers were lying on top of a duvet bedsheet, as the prosecution suggested that bed sheets had been thrown off and that Miss Steenkamp had been trying to dress, or as the defence case asserted the Jeans were originally elsewhere and put onto the duvet by the Law Enforcement officers as an act of contamination of the crime scene. Ultimately the provenance of the jeans was disputed, and this significant sequential artefact was disregarded by the trial judge, but it does highlight the importance of the placement of an artefact can drastically change the interpretation of the events that led up to that placement.

Although we have described the *Mens Rea*, mental component of a crime, within a digital forensic science environment the most significant challenge is determining who the person (or persons) using the device was at the time of the offence. The requirement for the court or jury to infer the mental state implies we know who the user was at any time. Physical control of a small personal device, such as a smart phone, suggests that the owner of the device is the user of the device, but if investigating a one-off offence then it is perfectly reasonable, within the standards of proof of 'reasonable doubt', that the device was left unattended, unlocked for a brief period of time.

A simpler task for digital forensic science is the investigation of offences that occur over a period of time, such as the single offence of Sexual Grooming which we noted above requires at least two instances of communication, or multiple related offences such as the collection of Indecent Photographs of Children over a period of time. No attempt is made during this thesis to discuss the psychological aspects of certain crimes, compulsive behaviour and the technology that allows these crimes to be committed, however we note it here as an interesting area of digital forensic science investigation. For investigations of 'habitual' offending, the reliability of the temporal components coupled with a weight of circumstantial facts relating to specific acts being repeatedly performed easily implies there was a *Mens Rea* to commit the acts, for example, a large collection of unlawful pictures present on a device with the time and date information that shows that the pictures were made onto the device on a number of occasions, suggests that someone actively and repeatedly created those pictures. Habitual behaviour does not conclusively show the identity of a person, but it does show regular access to a device, and if that can be coupled with personally identifiable behaviour then that is a significant circumstantial fact.

The standard of proof that evidence will be judged at varies from the type of court, and who is presenting the evidence. In a criminal prosecution, the common expression "beyond reasonable doubt" (BRD) is commonly used, and within a civil court it is the "balance of probabilities" or the "preponderance of the evidence" which Lord Denning described as meaning "More probable than not" [73]. A defendant that is presenting an affirmative defence, i.e. presenting facts that support the defendant's case rather than the prosecutor's case, will need to meet the lower preponderance of evidence test in a criminal trial and the Digital Forensic Scientist must be aware of that when testing the affirmative defence. Reasonable Doubt (RD) has been quantified as a certainty by the jury 90% (or 0.9) [30], and consequently for a jury to make a decision beyond reasonable doubt, BRD, they must be certain to 91% (or 0.91) or more. This is equivalent to Blackstone's ratio which states "It is better that ten guilty persons escape than that one innocent suffer" as highlighted by Lundrigan [67]. Consequently, we can consider Denning's judgement of "More probable than not" as a probability of 0.51.

Within the England and Wales jurisdiction, case law has decided that electronic evidence is considered Documentary Evidence [95] and as such the principle of Best Evidence Rule [77] requires that the original evidence should always be available for examination, which led to the Association of Chief Police officers (ACPO) issuing guidance [105] based around four principles, with the first ensuring investigators work on copies of the best evidence where at all possible.

Rule 3 of the ACPO guidance also requires the replication of the same initial artefacts, tools, techniques and procedures be able to produce the same output evidence. This is interesting in that it specifically ensures that no 'black box' tool based upon an unknown set of training data can be used to produce evidence for court.

1.2 CHALLENGES WITHIN DIGITAL FORENSIC SCIENCE

Beyond the legal and scientific challenges presented above, research in the area of Digital Forensic Science is affected by the operational needs of Digital Forensic Investigators/Analysts, the technological constraints of digital media and even the behaviour of the victims and suspects in the investigations. It should be noted that the terms 'Analyst' and 'Investigator' are used in this thesis somewhat interchangeably, with the difference that an analyst is a technology facing role, while an Investigator tends towards people- and case-facing roles.

There is a tension within Digital Forensic Science research between the tools and techniques that Academics are producing, versus the needs of practitioners. Baggili et al. [6] surveyed 500 papers in the period between 1992 and 2011, and the authors found that only 10% of the research projects involved academia and industry acting in collaboration. Al Fahdi et al. [4] highlights that practitioners were concerned with anti-forensics and encryption as future challenges whilst academics worried about tool capability and social networking aspects.

Few surveys have been performed to ascertain the perceived needs of the Digital Forensic Community. The 'cyber forensics needs analysis survey', initially performed by Rogers and Seigfried in 2004 [87], and subsequently reperformed with more extensive statistical analysis by Harichandran et al. in 2016 [44] highlighted the following as the top issues or challenges within computer forensics in 2004:

1. Education/training/certification
2. Technologies
3. Encryption
4. Data acquisition
5. Tools
6. Legal justice system
7. Evidence correlation
8. Theory/research
9. Funding

10. Other

When performed in the 2016 survey identified the following concerns:

1. Education/training/certification (ETC)
2. Technologies
3. Tools
4. Evidence correlation
5. Theory/research
6. Encryption
7. Legal/justice system
8. Data acquisition & Funding (tied)

Overwhelmingly, the results in 2004 and 2016 showed that Education/Training/Certification was listed as the highest priority by practitioners. The majority of these challenges are operational, such as Funding and ETC, which is beyond the wider scope of our thesis, whereas our thrust can be considered 'Theory/research' and 'Evidence correlation'.

Although throughout the ACPO principles for electronic evidence [105] there is the implication that the analyst/investigator should be suitably qualified, it is explicitly noted in principle 2 when an analyst should be able and competent to explain the implications of the actions that they take when interacting with a live, changeable system. The expectation in the ACPO Manager's guide [106] is that it takes 2-3 years for an analyst to become suitably knowledgeable to be considered competent, although no specific benchmarks are identified in that guide. The language of the manager's guide also is indicative of the cost centric view of Forensic Science: "The costs associated with running a specialist investigative unit within a law enforcement agency in terms of personnel, equipment and training is a significant drain of resources but the overall value for money represented by such an asset can often be overlooked".

Irons et al. [50] note that "the implicit expectation is that digital investigators should be competent before undertaking any digital investigation duties". The concern that analysts should be competent, indeed that they should be considered 'expert', before undertaking analysis is potentially the reason why Rogers, Seigfried and Harichandran et al. identify the practitioner's concern about not enough training. The analyst being 'expert' is not always the reality. Gogolin [37] highlights that many investigators in law enforcement have little to no digital forensic science training before starting, and even after they are in role, "[o]nly 34% of [digital forensic] investigators [surveyed in Michigan, USA] received formal training in laboratory forensics, with the majority being trained 2 weeks or less".

Casey [21] states that "too little knowledge is a dangerous thing" with regards to digital forensic investigations. Casey claims that investigators (both internal to law enforcement or outsources specialists) may have an "over reliance on user-friendly or automated forensic software", and may "apply a form of pseudo-automation by rigidly following predefined protocols". Casey also states that, "Inexperienced individuals who do not critically review the results of a tool will inevitably misinterpret or completely miss digital evidence", which is an important point that emphasises that the recovery of artefacts is only part of the evidential 'production' and interpretation is just as, if not more so important.

Ultimately the analyst providing 'expert' testimony is as the ACPO [105] guidelines emphasises: "It is also the personal responsibility of any person working within the area of digital forensics to maintain their knowledge of the subject areas they are involved in. Formal training is just one route", which again can explain the training related apprehension illustrated in the 2004 and 2016 surveys.

The volume of material that must be examined during an investigation is not trivial and there is a body of literature examining practical ways that, usually law enforcement agencies, can manage the large volumes of digital equipment and the electronic evidence that they produce. Irons and Lallie [51] studied the steady annual growth in the number of forensic investigations, the amount of data being investigated, and the amount of data being investigated per case using the annual data published by the FBI from 2007 to 2011 and a UK regional police E-Crime unit.

The complexity of investigations is not purely the volume of material that must be examined, but the complexity of technology, such as dynamic web pages or cloud computing, which may mean that an analyst may no longer be able to fully reconstruct the activity of a system. We can see this illustrated in a quote from Garfinkel [35] "Without developing fundamentally new tools and capabilities, forensics experts will face increasing difficulty and cost along with ever-expanding data size and system complexity. Thus today's digital detectives are in an arms race not just with criminals, but also with the developers of tomorrow's computer systems". This is not necessarily an issue of anti-forensic technologies, such as encryption or tools for obfuscating activity, but an issue of the evolution of the technology, enabling normal everyday usage than is far from trivial for an analyst to understand or reconstruct.

Al Awadhi et al. [3] describes a problem of operational Digital Forensics as a trade-off between the number of person-hours spent on investigation, which needs to be kept to a minimum whilst also paying close attention to the authenticity of the evidence. Lillis and Scanlon [66] describes that traditionally, information retrieval effectiveness is evaluated using the potentially conflicting measures of *Precision* and *Recall.* A system with high precision avoids returning documents that are not relevant, whereas a high-recall system aims to ensure that all available relevant documents are returned to the user. Digital forensics is typically seen as a recall situation, but high recall inevitably leads to a higher rate of false positives, this is tolerated due to the requirement to find all available evidence. Within this thesis, we repeatedly return to the problem of Precision and Recall (although we tend to refer to this as Availability, as we are modelling as much of the historical events as possible).

James and Gladyshev [54] make the argument for well-planned, careful use of automation that allows for a more efficient and effective use of automation in digital forensic investigations while at the same time attempting to improve the overall quality of expert investigators. They state that there will be no "dumbing down" of the profession when automation is used at the correct stages of investigation.

Triage is the term used within Digital Forensic Science in a broad sense to mean a process of selection of the available evidence to limit the volume to a quantity that can be managed within a reasonable period of time. There are differences in how triage is used, one of which relates to time sensitive cases where devices must be looked at in the field, and this can be seen typified by modern mobile phone triage as outlined in literature such as Rodgers et al. [86], Casey et al. [22] and Mislan et al. [74] where the advantage of triage is highlighted, such that that it provides investigators with automated, fast, in the field intelligence gathering.

Alternatively, triage can be described a process in which devices or processes are ranked in terms of importance or priority to a case, and is applicable to the analyst or investigator within the forensics laboratory setting. James and Gladyshev [54] highlight the benefits have been examined within a UK high-tech crime unit in Goss and Gladyshev [38], which showed a reduction in the quantities of seized computers and suspect data which needing full in-depth analysis. Goss also compared automated triage performance with manual investigation, and found that triage gave comparable examination results in a fraction of the time for specific case types where in-depth knowledge is not required, such as indecent photographs of children detection. The ACPO Manager's Guide [106] recommends a Triage Officer at laboratories for filtering cases, requirements and evidence based upon the operational procedures in place at the Agency.

Organisational issues also impact law enforcement agencies' ability to process and investigate digital equipment as they have centralised their forensic science capability. This leads to solutions where organisations will distribute out some analytical and acquisition capabilities into the wider agency, such as the solution illustrated by the Royal Canadian Mounted Police in [45] which deals with vast areas of geographical area that are sparsely populated. An alternative is to use distributed technology where multiple people, which can be located in different areas can interact through cloud technology to a centralised repository for that case, such as can be seen in the Netherlands Forensic Science [99] Digital Forensics as a Service paper.

While automation and triage have been shown to have benefits in some specific cases, and with some technology that lends itself to 'push button' analysis, there are challenges, such as investigator training, potential missed evidence and verification of the best evidence that needs to be addressed.

As we noted above, the 'natural world' of Digital Forensic Science has traditionally involved the manipulation of data on hard disk drives, memory cards and removable media. Given the rise of storage within networks or 'The Cloud' as we see in Smith [92] that Cloud computing was identified as the "most hyped concept in IT". This has led to not only the technological difficulties of acquiring a copy of data that is consistent with best evidence principles in these remote locations, but also as Taylor et al. [96] comment "in legal terms, cloud computing systems will make it potentially more difficult for the computer forensic analyst to acquire and analyse digital evidence to the same standards as that currently expected for traditional server based systems, due to the difficulty in establishing what data was stored or processed by what software on what specific computing device". This emphasises that there is a legal component to future investigations, whereby the analyst may not have the authority to copy data, or there are substantially different jurisdictional differences between the location of the suspected crime, storage of the evidence and location of the investigation.

We see therefore that a forensically sound approach should be compliant with the following:

- Analytically sound: Law of Comparison, Law of Probability and the Law of Circumstantial Facts.
- Related to the four elements of a crime: Time, Person, Mental state, Criminal act.
- The digital forensic analysis is proportionate and provides results appropriate to the device's participation in the offence: the offence was recorded by a device, was planned using a device, performed using a device, or the outcomes of the crime were researched.
- The results of the analysis, the circumstantial facts, are compliant with the Rules of Evidence.
- Ultimately the conclusions of the analysis reach a standard of proof, required by the court.

We highlight that the most significant technical challenge within Digital Forensic Science is the attribution of an action to individual, which is the most basic elements of a crime. This is not an insurmountable challenge, if it were then no digital forensic evidence would ever be presented. How this challenge is dealt with during an investigation is by the use of witness testimony about the physical access and control of the device. If a defendant makes preposterous claims about how the device could have been used by anyone, the inference of non-truth may be made by people in the court. If a defendant makes affirmative defence, that are testable by a digital forensic analyst, then the analyst will either agree or say that they cannot agree with the defence.

1.3 OBJECTIVES

There may be a need to model the overall activity on a system when investigations identify that a series of individual actions on the machine form a pattern of identifiable behaviour which may indicate who the user of the computer was during those times.

We can see that there are three general times such modelling would be useful to an investigation:

- Where there is an identified criminal event co-located in time with a body of Internet history, but not necessarily related to the Internet activity. Modelling of the activity could show that a regular user of the computer was present at the device, mitigating a possible defence that an unknown person was using the device. Examples of these co-located events could be the creation, modification or access of a file that is relevant to an investigation, such as viewing an indecent photograph of a child or modification for a document used in a fraud.
- Where there is Internet history containing the criminal activities: where there are accesses to websites known to contain unlawful material and we want to isolate all the sessions containing those acts and a) show if a regular user of the device is engaged in that kind of activity and b) establish the patterns of access relating to those acts so as to establish an investigative hypothesis relating to the identity of the possible suspects to establish which ones are the most likely offender.
- We may also see that there are cases where there are personally identifiable actions that take place in one or more sessions and these may have a similarity to other notable sessions, such as those that are co-located with notable events or directly containing unlawful material. The inference in these cases is that because there is some similarity between the session containing personally

identifiable actions and the notable actions that they are created by the same user.

We can therefore identify three profiles of offences that are of interest to our area of research:

- Single Events that are a crime
- Single Events that are a crime, that occur multiple times
- Multiple events that combine to form a single crime

We suggest that modelling repetitive behaviour would be helpful to an investigator in all of these cases as it could indicate who the user of the device was at that time, particularly where there are multiple instances. In the case of single events the activity at the time may also speak to motivation and the mental state of the user which is also necessary for showing the criminal intent. Example of these situations:

- Single Events that are a crime an example would be someone sending a harassing message over social media. The additional context could show the user of the machine was a regular user of the machine and not that it had been left idle and unlocked such that any stranger could have done it.
- Single Events that are a crime, that occur multiple times and example would be the accessing of unlawful material. The context modelling would show regularity of offending, allow an investigator to determine patterns (such as time of day) which could prove the identity of the user.
- Multiple events that combine to form a single crime an example could be grooming of children, or computer misuse-type offences. In these cases, the user has to research and scan for vulnerable targets and consequently the modelling can potentially show considerable evidence relating to the guilty mental state of the suspected offender.

We have highlighted here that there is a need within Digital Forensic Science to show how a device was used, not only at the time when the offence (or offences) occurred, but across the entirety of the device's usage.

This broader model of the system speaks to the probabilistic behaviour of the system, it may show mental state of the defendant, for example "the user visited X website numerous times before the Y act was performed", and if there are identifiable features within the behaviour, it may provide identification of the user, for example, "the acts occurred between 9 and 5 and as such it is reasonable to believe it was the normal work-time user of this device".

Our objectives therefore are to research the following:

- **Objective 1.** Identify the state of the art and challenges in event modelling in multi-user computing environments.
- **Objective 2.** Identify Internet history artefacts which would be typically present on a regular digital device that can be used to model human Internet browsing behaviour on the digital device.
- **Objective 3.** Evaluate feasibility and compare different approaches for aggregating multi-user Internet history sessions without prior knowledge of the user.
- **Objective 4.** Develop a method for grouping a computer systems Internet history without prior knowledge about its structure, so as to identify and extract idiosyncratic features with an accuracy beyond reasonable doubt, so as to be admissible in a criminal court.
- **Objective 5.** Visualisation of the grouped Internet history so that the results can be used for investigative reasoning and analysis of the aggregated history sessions.

The scope of this thesis is based upon an investigation into multi-user desktop environments that are connected to the Internet, where there is potentially weak user authentication or account sharing. This represents a realistic and challenging environment for digital forensic science.

1.4 THESIS SUMMARY

Chapter 1: Introduction

In this chapter, we have outlined the general features of Digital Forensic Science and have shown that it differs from the general case of Forensic Science in that the digital environment differs from the natural world environment. However, the development of a forensically sound approach should be compliant with the analytical laws of Forensic Science and extract reliable circumstantial evidence that can be evaluated probabilistically to a standard where a court or jury can make a decision based upon the relevant burden of proof.

We describe the key challenges that have been identified by other researchers and practitioners and note that the is a strong case for automation in the field of Digital Forensic Science as there are concerns about the volume of potential evidence that must be examine. Although there is a case for automation, there is also a counter point made that too great a reliance on automatic tools and procedures does not allow the analyst to challenge or test the findings of their tools as well.

We identify in this chapter that the greatest challenge within Digital Forensic Science is the discovery of the component of a crime that involves the identification of the individual liable/culpable for the actions. In our objectives, we propose that methods involving the grouping of data demonstrating regularity of behaviour can be used to extract identifiable features of the user which can be used to address this question.

Chapter 2: Context Analysis

As the objectives of our research are based around the investigation of specific actions or points in time, chapter 2 presents the literature review of event modelling from the perspective of the established Digital Forensics community, and other related material. We propose in this chapter that the investigation of events viewed as individual points in time does not facilitate any analysis relating to the mental state of the user, or the identification of who that user was, and as such we propose the framework of 'Context Analysis', where events are viewed from the point of view of collection and comparison of related artefacts.

Chapter 3: Session-to-Session Analysis

In this chapter, we show how the various types of Internet history data can be rendered as groups of activity, which we call sessions. We show that sessions can be variable-length which better matches the human interaction with the Internet, or fixed-length blocks which can better model the interaction and behaviour of the websites. We describe a simple method for comparing sessions to other sessions that is visually and computationally simple to understand and forms the basis of the approach used in this thesis. We demonstrate and investigate how the variables relating to the session selection can have an effect upon the results that can be produced, and we highlight there are a number of choices or 'dials' that we can adjust during our experiments to increase the availability of sessions we can analyse at the expense of accuracy and vice versa (the precision/recall problem).

Chapter 4: Zero-knowledge Internet History Session Feature Extraction

In Chapter 4, we propose two methods of sub-dividing the sessions data presented in chapter 3 into groups that are based upon characteristics either, of the data contained within the sessions, or the characteristics of the sessions. For example, short sessions are combined together, long sessions are combined together etc. We present a novel approach to grouping data based upon the relative popularity of the websites within the Internet history, i.e. we show that websites can be considered indicators if they occur frequently amongst the sessions within the Internet history, but are niche websites within the global popularity metrics.

The contribution of this chapter is that the approaches presented in chapter 3 do model the activity on the system, but they do not highlight the individual characteristics of the user and as such the contributions of this chapter facilitates the extraction of characteristics of the behaviour, which we propose can better identify the individual or their mode of behaviour during the session.

Chapter 5: Graphical Representation and Use of Session-to-Session Analysis

In Chapter 5, we show the results of the zero-knowledge grouping methods proposed in chapter 4 using network graphs of the Session-to-Session data. The sessions are the nodes of the network, and the similarity coefficient between the sessions is represented as the edges between the nodes. A network community detection algorithm is used to group the sessions with high similarity and this allows us to determine the accuracy and correctness of the different grouping schemes with different datasets.

We provide results from experiments from test data using the different grouping methods of grouping to investigate the performance of the methods proposed in chapters 3 and 4, further illustrating that there is a trade-off between reliability of the results and the availability of number of sessions we analyse. We test our data against a 'Beyond Reasonable Doubt' (BRD) value (as noted above, 0.91 accuracy) and show what the resulting graphs of the grouped data at the BRD level.

We also describe in this chapter a method for using sub-graphs based upon pattern of life information about the known facts in the case or the possible users' activities to attribute the network communities to a suspected individual. This approach may automatically suggest website or groups of websites for an analyst to investigate and generate lines of enquiry to assist in the identification of a user or the mental state of the user.

Chapter 6: Evaluation and Conclusions

We conclude the thesis with an evaluation of the Session-to-Session Context analysis approach which we presented in chapter 3, expanded in chapter 4 and utilised in chapter 5. This chapter concludes by highlighting the achievements of the research to date, and identifies possible directions of future work which can be used to increase the overall performance of the approach.

Appendices

Appendix 1: We provide a full set of results data performed during the experiments outlined in chapter 5 for assessing the overall impact of the grouping approaches.

Chapter 2

CONTEXT ANALYSIS

"Without context, words and actions have no meaning at all."

Gregory Bateson - Mind and Nature: A Necessary Unity, 1979

2.1 INTRODUCTION

Researchers and practitioners in Digital Forensic Science have proposed a number of frameworks and models to formalise the process of recovering artefacts and converting them into evidence before a court of law. We highlight a few examples of the notable models and research that were developed in this area:

Pollitt in 1995 [81] proposed a process of Acquisition, Identification, Evaluation and Admission as Evidence. A major framework proposed for Digital Forensics was the Digital Forensics Research Working Group (DFRW) model [78] in 2001, which consisted of the processes: Identification, Preservation, Collection, Examination, Analysis, Presentation and Decision. Reith et al. [85] in 2002 expanded upon the DFRW model and proposed a nine-part process of Identification, Preparation, Approach strategy, Preservation, Collection, Examination, Analysis, Presentation and Returning evidence. Carrier and Spafford [17] presented in 2003 a large model that contained the 5 major phases: Readiness, Deployment, Physical Crime Scene Investigation, Digital Crime Scene Investigation and Review. Carrier and Spafford's model was expanded in 2004 by Baryamureeba and Tushabe [8] with the addition of concept of the primary digital crime scene and the secondary physical world crime scene being investigated concurrently. Other researchers have presented frameworks such as Casey [20] identified a process of Recognition, Preservation, Classification, and Reconstruction, and Kohn et al. present the process of Preparation, Investigation and Presentation.

We can see therefore that the traditional views of the Digital Forensic framework can generally be broken down into the *Acquisition stages, Investigation stages* and *Presentation stages.* Depending upon which framework is used there is greater or less consideration to the physical 'crime scene', the investigation stage may require more, or less prior knowledge of the circumstances of the investigation and about what is explicitly sought, and the presentation stage is often focused specifically at 'the court'.

We present in this thesis *'Context Analysis'*, an analysis process that can be fitted into the investigative stages of an existing (or future) Digital Forensic Science framework, rather than proposing a new all-encompassing framework for investigations.

We propose a form of analysis of digital forensic artefacts that takes account of how a system is used rather than the traditional view of finding a specific artefact, such as a contraband file or picture. We call this novel form of analysis 'Context Analysis'. This differs from a traditional view of digital forensic artefacts which is highly content focused. Unlike content analysis, context analysis can be viewed as the "what", "where" and "when" characteristics associated with these artefacts. For example, content analysis may be the searching for words, patterns of phrases, skin tone or facial recognition features, whereas context analysis would focus on the location where the artefacts are stored, whether they were modified, artefact type, location and time of creation, modification, access etc. In general, context analysis uses artefact metadata to group or associate separate "point event" artefacts, temporally (which is the most common), spatially or based on artefact type.

It should be noted that there is a conspicuous absence of the “who” characteristic from the list of context analysis, as in “who was using the device”, because as we have noted in chapter 1, it is far from a trivial technical question to answer. Although many types of devices have some form of explicit access control or user accounts, the subject of a digital forensics investigation can easily claim that they were not the user of a device at a specific time, or even that if they were the user of the device that some background process was responsible for the artefacts. Consequently, the “who” characteristic is an outcome or goal of context analysis rather than a reliable contextual information.

We propose that context analysis consists of identification, interpretation, validation and activity analysis. Identification of artefacts is a well-known process within digital forensics, which lends itself well to automation, whereas activity analysis is very much human-driven and a state-of-the-art research question. Casey [21] highlights that currently “diligent human oversight” is required in automated processes that are applied to digital forensic investigations. The validation and interpretation stages constitute the diligent oversight, be they performed by an automatic or human-directed process.

2.2 THE CONTEXT ANALYSIS COMPONENTS

Here, we elaborate upon context analysis for digital forensics investigations and define the sub-divisions of the four components that form the analysis.

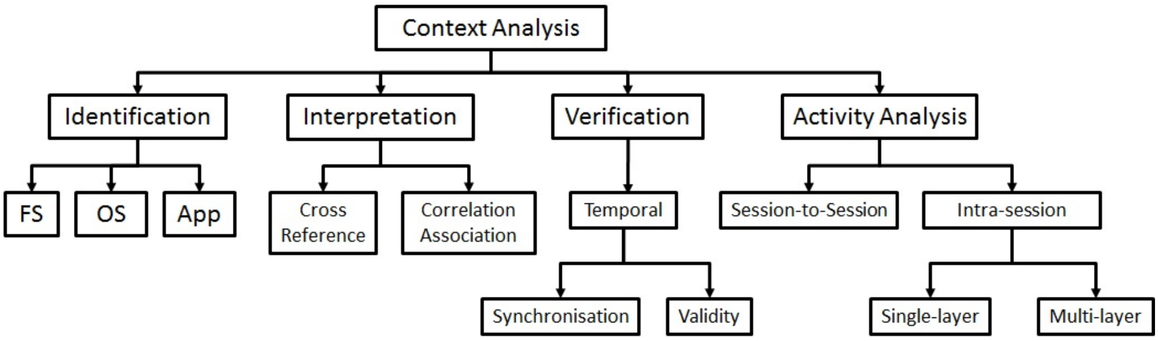


Figure 2.1: The characteristics of Context Analysis

Figure 2.1 shows the four major components of our proposed Context Analysis model. We will discuss this in more detail, then relate the Digital Forensics literature to this model in chapter 2.3.

2.2.1 IDENTIFICATION

The problem of artefact identification is well understood, albeit not straightforward. Artefacts can be present at one of three layers that we may classify as file system (FS), operating system (OS) and application (App). In some cases, the three layers can be highly dependent on each other, such as when an application writes the location of its configuration data, which is a file system address into a protected operating system area (e.g. the Microsoft Windows registry). Most of the time, however, the three layers can be seen as related but independent of each other.

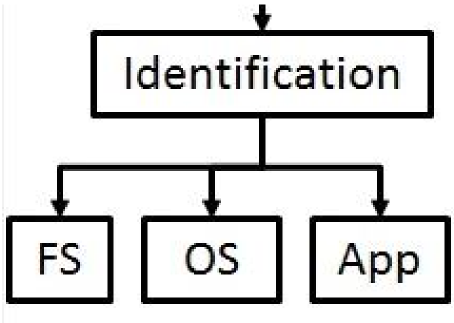


Figure 2.2: Identification within the Context Analysis Model

Identification - File System. Typically, the file system view is a logical hierarchical view of files, folders and directory structures built over a physical storage medium. The file system contains temporal data relating to creation, modification, access and destruction of artefacts, and may contain other metadata fields depending on the different file systems in use.

In a context analysis system, artefact identification can be particularly difficult when files are in a state of deletion or in some non-contiguous state, at which point it is crucial to identify the indexing system used to record the metadata relating to the files. In contrast, in a content analysis system, it is the ability to parse through and re-combine the individual storage units of the physical medium that is crucial.

Identification - Operating System. The operating system can span a range of power and complexity from a lightweight kernel system with a minimum set of features, to a complex tightly integrated desktop computing system or a distributed/cloud-based system. The operating system layer sits between the file system layer, where the storage of files occurs, and the application layer, where interactivity with the users occurs. The operating system layer, therefore, acts as gatekeeper, provider and monitor of resources, processing the logging, performance monitoring and management of files and applications. Within this layer we see paging files, hibernation files, links, pointers, Most Recently Used lists and other data structures that to some extent have their own file system.

An interesting and somewhat overlooked feature of analysis of the operating system layer is the nature that it is a dynamic system, patched and updated, with capacities that potentially change over the time span of the digital forensic investigation. When contextually analysing an operating system, the extent to which it will autonomously interact with the file system layer must also be considered, especially when involving automated defragmentation and backup procedures.

Identification - Application-level. The application layer sits above the operating system layer and often deletes or saves files from the file system layer. Applications may retain logs, history or relevant contextual information, such as the Internet history saved by a browser, the conversation logs of two users communicating with a chat application or the listing of files downloaded using peer-to-peer software.

2.2.2 INTERPRETATION

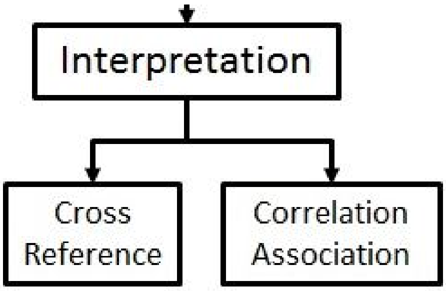


Figure 2.3: Interpretation within the Context Analysis Model

As noted above, there are few cases where purely the content of artefacts is so damning that they can stand by themselves without any technical interpretation. Typically, an investigator will want to know a history of what operations have happened to bring the artefacts into existence on the device, what operations have been performed (e.g., has it been modified or accessed?) and potentially what operations could be performed on the artefacts (are they accessible, visible etc.). Interpretation is a crucial stage of investigation and is normally performed by a skilled analyst cross-referencing certain known artefacts, or an automated statistical process identifying associations of patterns of significance.

Interpretation - Cross-referencing. Here, specific actions can be implied from the presence of two or more artefacts. For example, a picture file artefact and a link file artefact are recovered on the system. If it is known in advance that the link file is only created when a user clicks on the picture, then there is a cross reference to show that it is a human rather than an automated background process that accessed the picture. Cross-referencing requires prior knowledge of the system and the rules of where and how artefacts are created, modified, accessed and destroyed.

Interpretation - Correlation and Association. Here, artefacts that are in some kind of proximity to each other are assumed to have a relationship. The measure of proximity that is most commonly seen within digital forensics is time. Unlike cross-referencing, detailed rules and prior knowledge of the system is not required, but at the same time a certain volume of data is required before a reliable association can be made. For example, if a specific file is created during the time that a specific application is in operation, then there may be some general association between the two. However, if the specific file is of a particular type and those types of files are created during the operation of the same application, then we have a correlation between these two events.

2.2.3 VERIFICATION

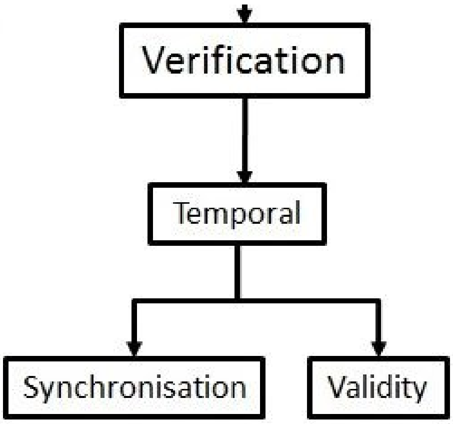


Figure 2.4: Verification within the Context Analysis Model

This is the process of establishing the correctness and equality between the artefacts. In this model, a single characteristic of contextual verification is used, namely the temporal characteristic, as this is a universal metadata characteristic across a broad spectrum of digital forensic artefacts. Note that some examinations, most notably in mobile computing, may have a spatial component such that there needs to be a time and space validity check. We have omitted Spatial verification from the model during the scope of this thesis.

Although we note that the temporal characteristic is broadly universal across digital forensic artefacts, it is by no means standard. A variety of levels of temporal precision exist on different file systems, operating systems and applications. Operating systems may be configured to operate in different time zones, and artefact synchronisation to a known reference time is not a given with any forensic examination. Consequently, reliable context analysis requires testing of data validity and synchronisation.

2.2.4 ACTIVITY ANALYSIS

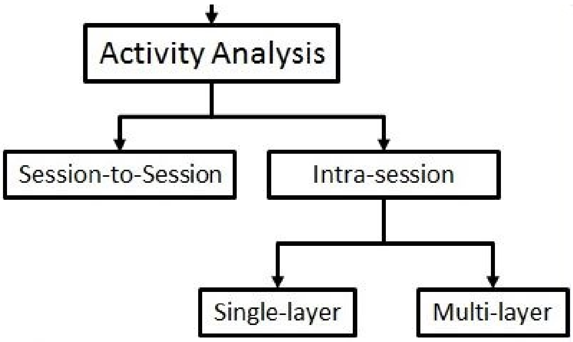


Figure 2.5: Activity Analysis within the Context Analysis Model

Activity analysis is the most complex aspect to the context analysis process in that it takes artefacts that are grouped temporally, spatially or using a combination of the two. Our focus here is on temporal grouping. Activity analysis presumes that there is a concept of a ’session’, a period of activity where there is a start, a number of artefacts and the end of the session, which is then delimited by an idle period before the start of the next session. In a section of the related academic literature, the entire data set is a single session. Other papers focus on a significant artefact and all other artefacts that are within a temporal proximity. For instance, when a USB stick is plugged into a computer all the artefacts created within the next ten seconds may be considered as part of the same session.

Activity Analysis - Session-to-Session. Aggregate data compared to other aggregates of data we have referred to as Session-to-Session analysis, and this type of analysis forms the bulk of the work from chapter 3 and onwards.

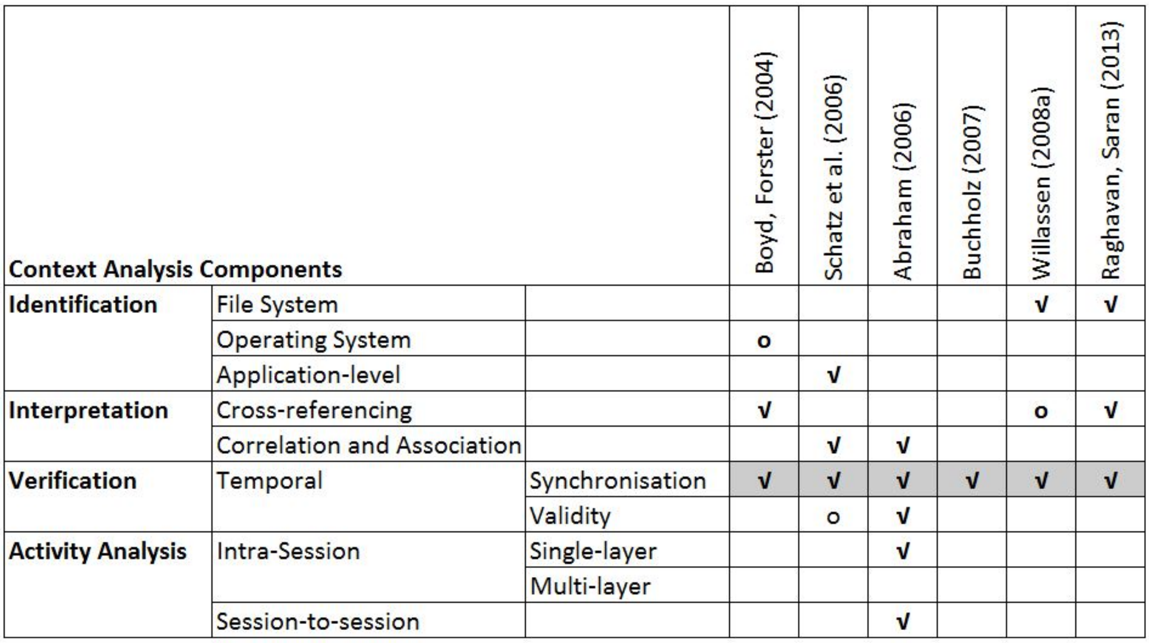
Activity Analysis - Intra-session. A sequential order of events that are spread over different levels such as the accessing of a File System artefact leaving a trace in the Operating System level. Systems that analyse these sequential Multi-level patterns we have referred to as ‘Multi-layer’ and sequential analysis of patterns that are present at a single level we have referred to as ’Single-layer’.

2.3 LITERATURE REVIEW

Here, we present the body of research that is applicable to Context Analysis in digital forensic investigations. The vast majority of related publications span more than one of the four categories identified. They may for example, mention their identification and interpretation approaches, but their focus is primarily on verification or activity analysis. So, to avoid replication we cluster all items of the survey based on their primary approach for example verification and activity analysis, and only briefly mention their identification and interpretation approach where relevant. We summarise this in a table for each section, where we display a '✓' for the primary approach followed and an 'o' for approaches that are noted in each paper but are not focal points.

2.3.1 TEMPORAL VERIFICATION SYNCHRONISATION

The consequence of failing to take into account the artefact time stamps compared to a standard reference time can lead to substantially different interpretations of the results of an investigation. It is common to encounter numerous synchronisation issues, such as changes in timezone caused by a shift to or from daylight saving time, or events that are recorded in both application logs and file system entries that have differences in time because the system and disks are recording at different rates. This problem is aggravated when there are multiple devices involved in an investigation.



Context Analysis Components			Boyd, Forster (2004)	Schatz et al. (2006)	Abraham (2006)	Buchholz (2007)	Willassen (2008a)	Raghavan, Saran (2013)
Identification	File System						√	√
	Operating System		o					
	Application-level			√				
Interpretation	Cross-referencing		√				o	√
	Correlation and Association			√	√			
Verification	Temporal	Synchronisation	√	√	√	√	√	√
		Validity		o	√			
Activity Analysis	Intra-Session	Single-layer			√			
		Multi-layer						
	Session-to-session				√			

Figure 2.6: Temporal Synchronisation Research Papers

Boyd and Forster [10] document a well-known (if somewhat notorious) case where law enforcement analysts were falsely accused of tampering with a computer that had been seized by police officers, because a computer examiner acting for the accused party had failed to take into account the difference of the time on the computer against the time of seizure by law enforcement. The paper provides a checklist approach to ensure that a practical examination clearly documents and accounts for substantially different timezones. Most importantly, the paper serves as a cautionary tale to the implications of a failure to synchronise evidence to a reference timeline.

Schatz et al. [91] discuss the problems of synchronising timelines from multiple sources, including different machines or different applications, such as Internet Explorer versus Google Chrome, and parts of the Operating System, such as the Most Recently Used list. The authors examine in detail the drift that occurs in clocks that have not been set to automatically synchronise. The experiments presented in the paper show the change of the system clock against the baseline time. They show that there is a correlation between artefacts when synchronisation is accounted for between different machines. Notably, they also explain that the significance of artefacts for the experiment depends on the granularity of the recording. For instance, cache files appear to be more useful than Internet history records. As a result, the authors emphasise that a forensic analyst should not assume that a perfect logging system is in place and that a complete and robust log of all artefacts is present. Data can be missing and the decoding of artefacts from a system can be incomplete.

An interesting approach to mining sequences of events is presented by Abraham [1], who defines how systems are formally used and identifies unusual occurrences. The author discusses how different types of profiles are useful for customer personalisation profiles. A number of issues are explored, such as the requirement for a unified timeline when using sources of data from multiple locations, defining events and sequences of the events. Sequence chains of events are defined as regular sequences, such as ABC, which may have one or more possible irregular events (say ABDC or ABEC) present. Sequences may require wildcard events to correctly identify a pattern in the system (e.g. AB*C). Also, whether repetitive or non-repetitive, profiling events is not straight forward. For instance, it is not always obvious whether an ABAB sequence is an instance of a single pattern or two shorter AB patterns repeated. Consequently, to profile complex patterns one needs rules for defining maximum length, minimum length, pruning and similarity, which depend on the scope and the complexity of the event chains. This particular paper includes an example of a door lock log. Although considerably simpler than say profiling an entire Internet history with multiple possible users, it is still not trivial to compile subject and sub-profile lists, which can themselves have sub-profiles within them. Through the detection of inconsistent behaviour or outlier detection, subjects may show multiple users with access to a single subject's account. The paper explicitly deals with validation of temporal events, talks about synchronisation, and presents a model pattern construction that can be used for both intra-session and session-to-session analysis.

Buchholz [13] follows on from Buchholz and Falk [12] by identifying that there are limitations with earlier models such as the one used by Stevens [93], as they do not compare well to a reference time, do not appreciate skew and there are some issues with certain time zones being given preferential attention while other times zones do not have a 1-to-1 mapping to reference time. To deal with these limitations, the author describes in detail clocks and events. Most importantly, he also notes that "an investigator needs to have the ability to actually adjust timestamps to the proper reference time and to create synchronized time lines as the final result". The result is a clock model. The particular paper shows a series of experiments where the effect of long term operation of the clock causes observed clock skew and the powering on and power off of the computer can change the skew. A "drift graph" to track the clocks drift away from reference time is mathematically defined and demonstrated. Its purpose is to facilitate tracking antedating.

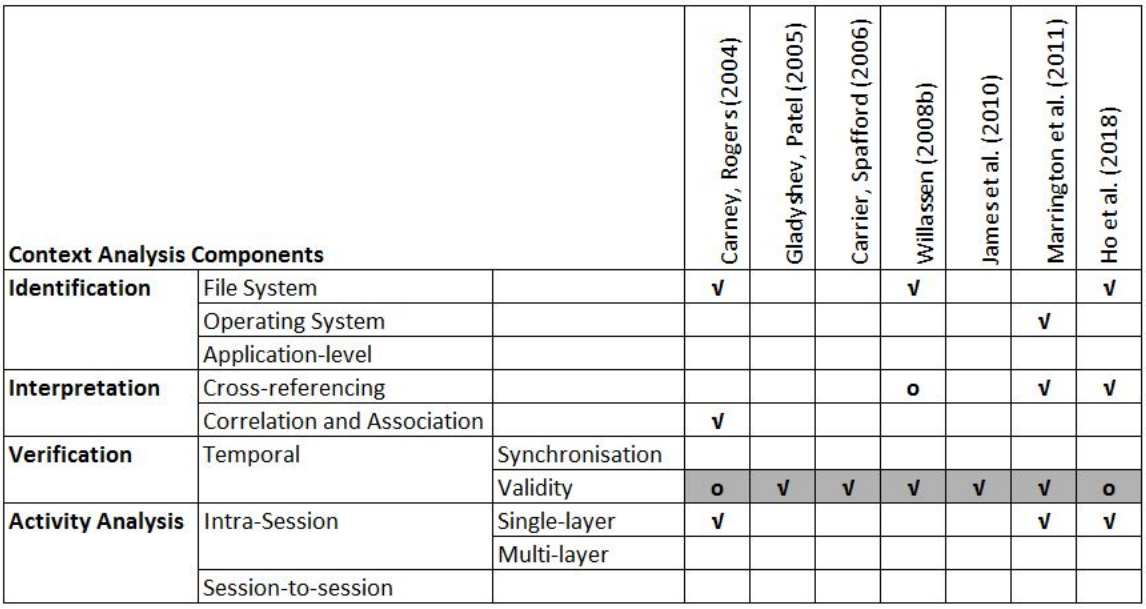
Willassen [101] suggests that any timestamp artefact is subject to hypothesis testing to ensure that the time recorded for the artefact is correct. This is because clocks not only drift but may also be altered by a user to a time that is not the objective "civil" real world time. The paper shows how actions that can change artefact timestamps can be listed, such as modifications to the file created, written and accessed times within a file system, as well as the sequences of actions that could have occurred to affect the state of an artefact.

Specifically for timeline analysis, Raghavan and Saran [84] propose a Provenance Information Model (PIM) which focuses on single artefacts that can only be credibly examined when present within a timeline. They highlight that the key challenges in creating timelines is to identify the syntax of the original time record, determine reference time and determine synchronisation to that reference time (skew and drift). Within a single computer there may be various homogeneous sources of time information. Therefore, the authors were motivated to create a provenance model that can synchronise with and between the various artefact types. To achieve this they have created a tool for creating a reference timeline which artefacts are enter onto and validated against regardless of the original timezone or local time variation that was present from the original provenance. The authors note in future work that predictive methods of determining skew and drift could be added to a PIM.

Within our thesis the issue of Temporal Synchronisation is largely accepted that the artefacts that we investigate will be correct. We note however that Abraham [1] provided very important motivation in our thesis, showing the complexity of sequential analysis and informing our choice of session aggregates. When we make comments that Session-to-Session analysis could be used for other types of data, such as sensors, such comments are informed in large part by Abraham.

2.3.2 TEMPORAL VERIFICATION VALIDITY

Validity tests aim to detect when individual events or artefacts have had timestamps modified, or when the system clock as a whole was modified. Unlike synchronisation issues which show a systemic variation to reference time, validity checks are principally aimed at identifying anomalous modifications or sudden unpredictable state changes to events, event sequences or the system as a whole.



Context Analysis Components			Carney, Rogers (2004)	Gladyshev, Patel (2005)	Carrier, Spafford (2006)	Willassen (2008b)	James et al. (2010)	Marrington et al. (2011)	Ho et al. (2018)
Identification	File System		√			√			√
	Operating System							√	
	Application-level								
Interpretation	Cross-referencing					o		√	√
	Correlation and Association		√						
Verification	Temporal	Synchronisation							
		Validity	o	√	√	√	√	√	o
Activity Analysis	Intra-Session	Single-layer	√					√	√
		Multi-layer							
	Session-to-session								

Figure 2.7: Temporal validity research papers

Carney and Rogers [16] have developed a specialist file system analysis technique based upon statistical differences between temporal metadata of the files. The technique proposed in this paper is to create hypothetical scenarios such as the user of a computer visiting a website that creates pop-up windows containing illicit material, and then metrics are calculated to show average time between file creations, saved items, viewed item etc. There are four reasonable example scenarios presented for the creation of unlawful pictures on a computer. The use of different applications, such as web browsing or "back door" remote control software, demonstrates variety in the approach. The analysis is limited to file creation times and to specific artefacts, such as the most recently used lists and the folder view thumbnails. The scenarios are of a somewhat limited scope, given the possible host of ways that unlawful pictures could appear on a system.

Even if the exact time of an event B is unknown, if it is known is that there is a causal relationship between an event before (A) and an event after (C), and that A and C have reliable timestamps, then B can be time-bounded between A and C. Based on this, Gladyshev and Patel [36] use a directed acyclic graph to represent causal connections between events. Because of time bounding, intervals of events and intervals between events can be calculated with some accuracy. However, the scope and scale of the manual processing involved to assign times to the events raise questions about the human readability of such graphs in a large scale environment and does rather indicate that it would possibly be more suitable for automated analysis or used to prove small-scale limited 'smoking gun' type events.

Carrier and Spafford [19] noted that forensic analysis of a system was the analysis of the current snapshot in time of the system or possibly previous states might be recorded. If an investigator needed to make logical inferences about states that are not recorded, or the information about that state are no longer present then there were no formal models of computer history to use. An example of a missing state could be when a file is updated numerous times, but only the last update is recorded in the files meta-data. The authors proposed two computer history models based on finite state machine (FSM) theory. The paper does also note that as static model of system capability is inadequate as devices are connected or disconnected, the capability would change and consequently a dynamic state model should be used to represent the changing capability of the device. Carrier and Spafford state that a primitive-level finite state machine model is not practical due to the vast quantities of state transitions that occur in a short space of time on a normal simple computer investigation. Hence the authors propose much higher-level history model based on a number of analysis classes, including abstraction, construction, reconstruction and materialisation.

Willassen [102] extends the work in Willassen [101] to show that in situations where there is a strong causal "this happens before that" logic, such as when files are sequentially numbered, then additional checks can be added to the action sequences to detect modification and further verify hypothesis testing of timestamps. The approach demonstrated takes existing functionality of the New Technology File System (NTFS) Master File Table (MFT) to create a logical check of whether data are in the correct time order based upon number sequences. Case studies are presented showing the results of users attempting to antedate data. The experiments are simple scenarios such as if a user without specialist knowledge did X, then would the approach detect X and the paper does not extend the analysis to the level of effort and knowledge that would be required to defeat the proposed analysis technique.

Extending previous FSM-based work, James et al. [53] show a model for defining states and transitions. They present an example where the claims in statements by different people under investigation can be verified logically. However, as the FSM diagrams enumerate all possible states and transitions, the relatively simple example presented in the particular paper may rapidly become unreadable. The authors acknowledge that an FSM based system is not a trivial matter for 'post-mortem' analysis of a computer system. The statement verification model does require some absolute and correct statements, and does not take into account error, uncertainty or the unreliability of witness statements.

Marrington et al. [72] extends the work on computer profiling in [70] to substantially include anomaly and inconsistency detection. The authors note that the predominate types of inconsistency that occur are related to the normal operation of systems where data is destroyed and overwritten rather than to deliberate tampering. The authors expand their event model which included recorded events and inferred events, to include missing events, which are pre-condition events that must occur in a system with well-known "happens-before" causality. A $\rightarrow$ B, B $\rightarrow$ C, transitively A $\rightarrow$ C. In an example where event C is a recorded event with a precondition that A must happen before B, and B can be implied from the recorded presence of C, then we can detect that A is missing. The authors have presented experiments to demonstrate the CAT Detect approach, notably collecting data from the start to the end of a session for a user logged into the computer's account. Large sessions do however run the risk that multiple users could use the same device during the logged in session and consequently attempting to profile individual users or types of activity could be difficult without the use of a smaller "temporal proximity" window.

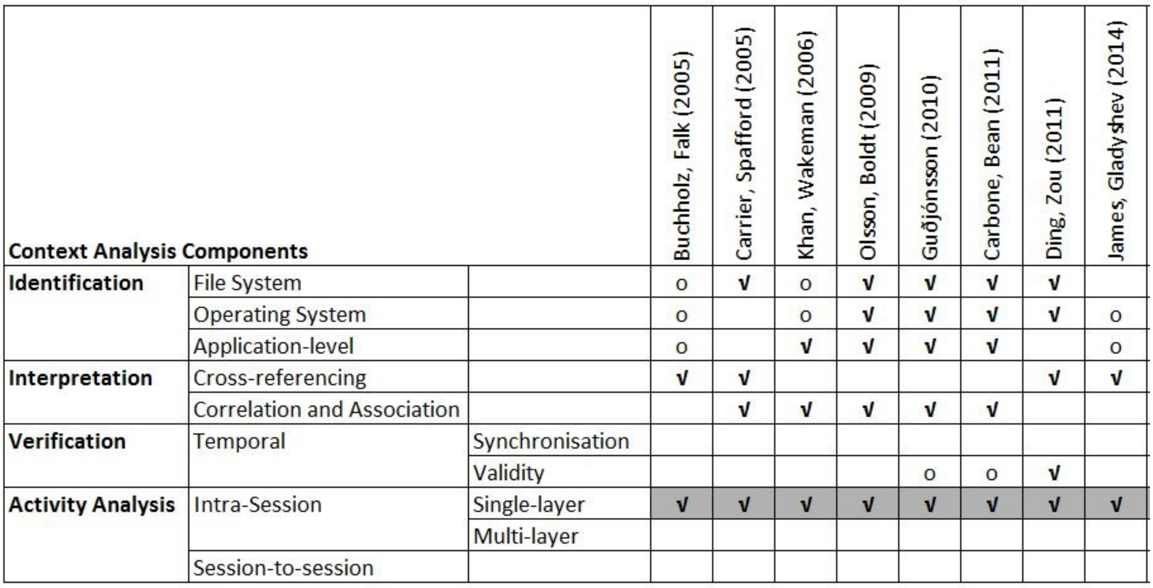
Ho et al. [48] in their paper show the problems in Context Analysis for ensuring the validity of file system artefacts as the are moved over modern cloud-based networks. The authors show experiments of moving a file from one type of file system, to a cloud to another cloud system and the observations of the impact of that transfer onto the metadata for the files.

This section was incredibly influential within our research, particularly Marrington et al. [72], although ultimately this work is about sequential correctness of sessions rather than the overall pattern of use. We could describe the type of work in this sections as being akin to cross-examining a witness, e.g. "So what did you do next? Where did you go? Which way were you facing when you saw the bag? How could the man have ran to the left if you saw the bag on right?". Ensuring that artefacts are presented in the correct sequence and identifying when there are missing elements is incredibly important, but we found during our specific research, sequential validity did not lend itself to the overview analysis of Internet history.

2.3.3 INTRA-SESSION ACTIVITY ANALYSIS - SINGLE-LAYER

The principal challenge for intra-session activity analysis is the grouping of a set of point events, such that they can be considered a contiguous period of operation. For example, if event X happened at a particular timestamp and Y happened at a different timestamp, concluding that the difference between point X and point Y is brief enough to say that the device was in continuous operation by the same user or process is not trivial.

Buchholz and Falk [12] have developed a graphical tool, called Zeitline, which allows events to be reconstructed from a number of different kinds of artefacts, logs and timestamps. In an analogy to a file system's browser with directories and files being the children of a root, the Zeitline's authors treat a timeline as a tree with a complex event as the root and a hierarchy of events as its children. It allows a user to create complex events from the records available, use search and filter capabilities, and populate and analyse timelines from those events. The approach requires known patterns and that the investigators select the artefacts that they consider pertinent.



Context Analysis Components			Buchholz, Falk (2005)	Carrier, Spafford (2005)	Khan, Wakeman (2006)	Olsson, Boldt (2009)	Guðjónsson (2010)	Carbone, Bean (2011)	Ding, Zou (2011)	James, Gladyshev (2014)
Identification	File System		o	√	o	√	√	√	√	
	Operating System		o		o	√	√	√	√	o
	Application-level		o		√	√	√	√		o
Interpretation	Cross-referencing		√	√					√	√
	Correlation and Association			√	√	√	√	√		
Verification	Temporal	Synchronisation								
		Validity					o	o	√	
Activity Analysis	Intra-Session	Single-layer	√	√	√	√	√	√	√	√
		Multi-layer								
	Session-to-session									

Figure 2.8: Single-Layer Intra-Session Activity Analysis Approaches

Carrier and Spafford [18] present an iterative four-stage crime scene and digital crime scene processing model. Within this model the authors state that the target definition phase is more of a heuristic process than the other phases which more closely resemble engineering processes. As a response to the heuristic nature of this problem, the authors present two different approaches. In the first approach, they show that cross-referencing the target definition based on known or predefined patterns allow detection of artefacts that are similar in time, location or content. The second approach detects outlier files or folders that have been buried within the file system structure so as to not appear conspicuous.

The approach by Khan and Wakeman [59] 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. The trained neural network can be applied to a forensic examination to attempt to reconstruct a timeline of events when the application was used. The authors do not consider whether using a training set from a wholly different source is desirable from a legal standpoint. An experiment is used in the paper to illustrate the approach and the authors highlight that there are performance considerations with particularly large datasets. The authors do not elaborate on the types of applications, the effect of different versions of the same applications or indeed if there are significantly different outcomes. They suggest that despite the complexity in describing the actions of machine learning to laymen, jurors and members of the legal profession, the value of their approach is worth considering for reconstructing events on a system.

Olsson and Boldt [76] present the Cyber Forensic TimeLab (CFTL) tool, which includes a scanner component for parsing a computer for a set of predefined artefacts from a variety of locations, and a viewer/presentation tool that displays the recovered artefacts on a histogram timeline. The authors show that the approach is extensible by adding additional parsers to the scanner component. The approach does not provide automatic analysis. Instead, it is a clustering and display tool for a number of individual timelines that are placed over each other so that a human analyst can make a visual correlation. The authors do demonstrate with a case study that the tool is beneficial when used side by side with a commercial tool, such as the Forensic Tool Kit [116].

Gudjonsson [43] suggests that timeline analysis carried out purely at the file system level does not provide adequate context for traditional file system artefacts. In response to this, the author presents the log2timeline tool which collects together file system dates and time but also parses log files and a variety of data structures for timeline information. All of the timeline information is put into a monolithic list or super-timeline. The paper details the locations and types of artefacts that can be gathered into the super-timeline. The author suggests that each timestamp is relevant, but it has to be understood within the context of the surrounding timestamps. The author does also discuss issues relating to temporal proximity, and analysis is by known keyword or by known pattern.

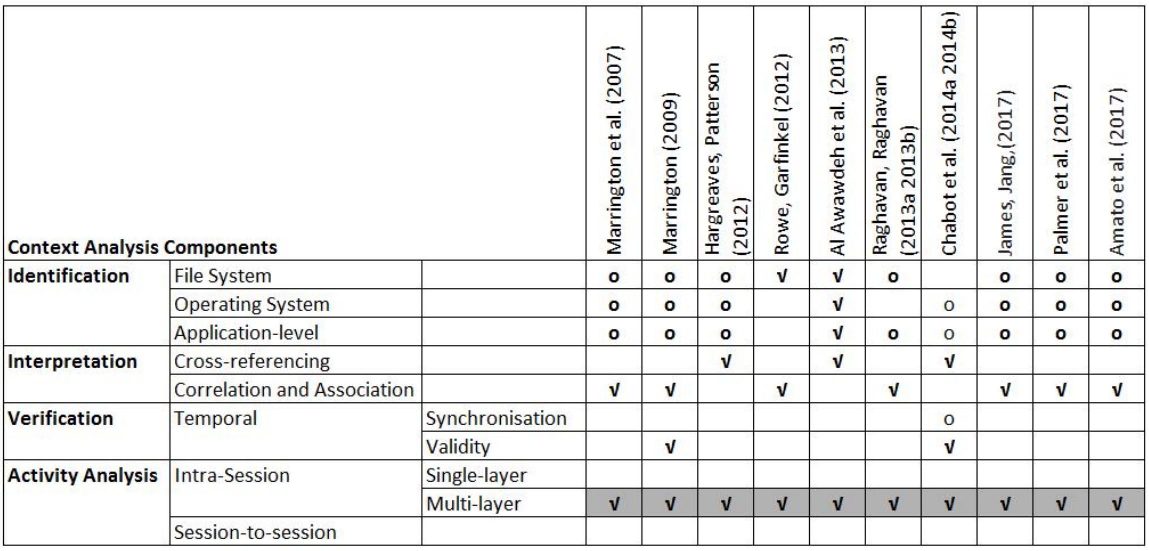
Carbone and Bean [15] review timeline creation utilities and competing formats for the storing and representing of timelines. The authors list the possible sources of timeline data that can be extracted with the log2timeline tool and make some cautionary notes about 'information overload' if too many files that are not pertinent are included. This paper, although extensive in its detail about available investigation tools advocates a flat super timeline. An interesting aspect to this paper is however that the authors provide detailed descriptions of file system permissions. Not only do they discuss the standard metadata times, such as created, modified and accessed, but they include descriptions of what could happen to the artefacts because of the file permissions.

Ding and Zou [32] detail the NTFS MFT operations that modify and create dates and times, and show how attackers can modify the MFT to obfuscate attacks on a system. The authors present an approach where suspicious files are checked in a three-stage cross-referencing process. They extract the appropriate temporal data, cross-check metadata within the operating system and the file system and then validate the accuracy of the timestamps based on rules related to their operation in NTFS.

James and Gladyshev [55] define the concept which they formally specify as action instances. This is 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 most recent action instance can be identified by applying the pattern across an entire computer system. Past action instances may also be reliably identified. The authors show experiments to highlight the use of this model using Internet browser artefacts and operating system logs. .

2.3.4 INTRA-SESSION ACTIVITY ANALYSIS - MULTI-LAYER

The multi-layer approach does not use the activity timeline as a monolithic list of point events but uses an understanding of the type and location of the artefacts. The multi-layer approach is more sophisticated than the single-layer method because it identifies the patterns in the artefact types and locations and presumes that concurrent patterns of different types of activity can be operating at the same time.



Context Analysis Components			Marrington et al. (2007)	Marrington (2009)	Hargreaves, Patter son (2012)	Rowe, Garfinkel (2012)	Al Awawdeh et al. (2013)	Raghavan, Raghavan (2013a 2013b)	Chabot et al. (2014a 2014b)	James, Jang,(2017)	Palmer et al. (2017)	Amato et al. (2017)
Identification	File System		o	o	o	√	√	o		o	o	o
	Operating System		o	o	o		√		o	o	o	o
	Application-level		o	o	o		√	o	o	o	o	o
Interpretation	Cross-referencing				√		√		√			
	Correlation and Association		√	√		√		√		√	√	√
Verification	Temporal	Synchronisation							o			
		Validity		√					√			
Activity Analysis	Intra-Session	Single-layer										
		Multi-layer	√	√	√	√	√	√	√	√	√	√
	Session-to-session											

Figure 2.9: Multi-layer intra-session activity analysis approaches

Marrington et al. [70] describes computer profiling as a method of forensic reconstruction analysis to determine a system's characteristics, behaviour and usage, that requires no prior knowledge of the system or initial direction to the sought artefacts. The approach presented in this paper is classifying a timeline of events into an object-orientated model and then performing correlation on the objects. The approach does identify that there are discovered events, which are directly extracted from the sources of artefacts (such as log files) and there are events that can be inferred from a known causal connection. For example, if event B is a discovered event and it is known that event B is always caused by an event A, then A must have occurred. Although computer profiling ostensibly does not require direction by a human investigator or prior knowledge of the events on the system, it is not to say that it is purely statistical, rather the object model and the inferred events do require detailed system specific knowledge (such as how inter-related artefacts such as link files and Registry files on a Windows system are for example), but the circumstances of the investigation, how many users etc. are not necessarily known.

Marrington [71] extends his research on temporal inconsistency detection initially presented in Marrington et al. [70]. Temporal inconsistencies include missing or overwritten data, clock skew, drift and also intentional modification. The research presents examples of inconsistency detection, such as misattribution of the ownership of the documents and modification of timestamps. The author shows that it is possible to perform automated detection for temporal inconsistencies when rules governing causal behaviour on the system are known.

Hargreaves and Patterson [47] propose an approach that first extracts timeline artefacts from a system, then performs low-level analysis of the artefacts and finally combines the low-level events into high-level events. For example low level events are simple records or log artefacts such as the record of a USB device being connected, or the record of an executable program running, whereas the high-level events are a known activity or action happening, such as a possible virus is introduced into a system through a USB device being plugged into a system. The two low-level events above performed in that sequence in close proximity could suggest the high-level pattern of an automated virus being introduced. The approach uses rules defined in advance of the analysis, be it extraction, low-level or high-level analysis. This approach is not about automated discovery of a systems behaviour, but about automating the identification of known types of activities or events. The approach is however distinctly modular and therefore the ease of adding and extending the established rules is noted.

Rowe and Garfinkel [89] compared the file systems on a variety of different drives and anomalous files are detected. Their approach is to compare a large quantity of hard disk drives which will by the nature of normal usage have broad similarities, with semantically predefined groups for the artefacts, such as pictures files grouped together, database files grouped etc. Drives with large quantities of pictures, media or application files will out stand as being anomalous.

The approach by Al Awawdeh et al. [2] is a real-time agent for recording data as it happens rather than post-mortem style forensics. The aim is to provide data for further academic research and to provide data for incident response in a large scale commercial deployment. The authors discuss the problem of verbosity, which is the issue that unimportant details can be over-reported in logs and salient details are not given adequate prominence even though they are reported. Within their experiments they show that there can be significant differences between the amount of data recorded by different operating systems.

The AssocGEN approach in Raghavan and Raghavan [82], [83] is a simple correlation between file system metadata and web-based application logs. The purpose is to provide the origin of a file and specifically to determine if that file has been downloaded from the Internet. The authors approach is to trace the metadata of a file to other activity on the system with similar metadata associations and times, and then ultimately to try to establish the website address in use at the time and confirm that the picture file did in fact originate at the location identified. The approach not only makes comparisons against log files but also against network packet captures, suggesting that the approach is an active monitoring tool rather than a 'post-mortem' forensic examination. It also has the requirement that the files cannot be overwritten, must contain metadata and are recent and accurate.

In Chabot et al. [24], [25], a holistic approach is proposed for gathering data relating to the circumstances of the investigation, provide a model of the investigation process, tools for extraction of heterogeneous data and to "provide tools to assist investigators in the analysis of the knowledge extracted from the incident". The approach requires a high degree of prior knowledge to model different scenarios with a method called Semantic Analysis of Digital Forensic Cases (SADFC). The authors review a number of forensic analysis techniques for comparison, using the criteria of auto extraction, heterogeneity, analysis, theory and data integrity. Excluding auto extraction, which is primarily a criterion to assess the tool's effectiveness, the majority of approaches had a capacity for dealing with heterogeneous data. All techniques scored rather poorly on automatic analysis, theory and data integrity. Analysis is discussed and is broadly based upon temporal proximity to events, correlation in time and heuristic rules that have been determined in advance. The authors argue that a digital forensic investigation needs rich knowledge representation and processes for consistency checking of data and filtering. With respect to events the authors describe the intervals of events and use time boundaries from Allen's logic, much as can be seen in previous works such as Gladyshev and Patel [36]. Interestingly, they point out that because of the nature of time intervals and the beginning and end boundaries potentially overlapping it may not always be possible to discriminate between event footprints".

The approach used by Kalber et al. [57] is to perform statistical clustering on a file system to identify what applications and files are closely associated in time. The approach presupposes no prior knowledge of the system, although to successfully interpret the results would require an analyst to review the clusters and confirm the associations that have been made. The authors note that events that happen at the same time or within a very short space of time can be differentiated because they will appear in different clusters. The paper presents an experiment where the approach is successfully used to identify the use of various applications across a file system but the author notes that investigation within the applications, such as examination of an email application is not possible with their approach.

James and Jang [56] propose a generic detection and general identification of events on a system without specific prior knowledge what the events are. The Action Instance model relates to events that update a number of traces and by detecting the traces this approach can make deductions about the events that caused the traces. The approach does require some domain-knowledge relating to the causal relationships between events and trace artefacts and the authors do note that if you were basing the traces on a location (as opposed to clusters of metadata) then it is possible for a user to save a file into that location without the normal event that would have occurred to cause a file to be there. They give the example of files in the Internet Cache are usually present as the result of a web page viewing event, but it is possible for a person to save into that location without it being caused by a web page view.

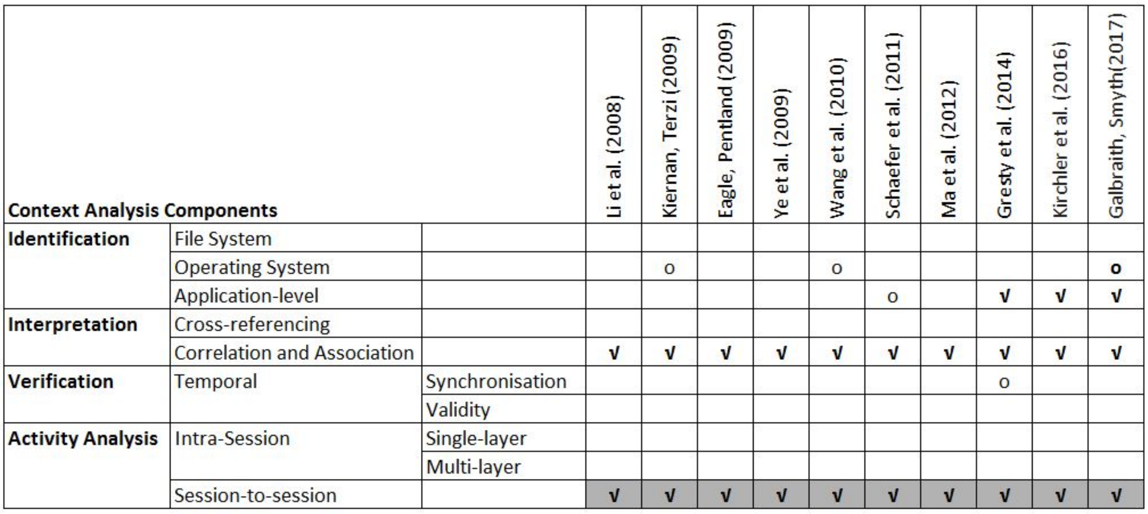
Palmer et al. [79] demonstrate in this paper how digital forensic evidence is difficult to reliably present. The authors assert that placing the digital forensic artefacts of note onto a graph and using an appropriate algorithm to draw associations between the nodes a correlation can be drawn between the artefacts, and the they give an example from memory analysis showing files and the software used to access those files can be correctly deduced.

Amato et al. [5] suggest that digital devices are increasingly likely in modern investigations involved as goal of the crime, medium or simply witness of a criminal event. They propose a framework for the analysis and reasoning of digital investigations, by adopting the practices and technologies of Semantic Web. Amato et al. propose that the use of such technology would provide advantages of Information Integration, Classification and Inference of evidence, Extensibility and Flexibility of resources and improved Search capabilities.

This section of the research was very influential in our thesis and shows the two big issues are cross-referencing for known patterns that are significant, and for correlation and association to discover the potentially interesting but unknown activities or events. Ultimately our research led to working with correlation and association style discovery of previously unknown features as we were particularly interesting in pursuing zero domain knowledge in this research.

2.3.5 SESSION-TO-SESSION ACTIVITY ANALYSIS

Where there are multiple sessions within a data set, there is the possibility of assessing the similarities and differences between sessions and inferring some session-to-session patterns. For example, sessions that appear at certain times of day or on certain days of the week may bear similarity to the corresponding sessions in other days or weeks.



Context Analysis Components			Li et al. (2008)	Kiernan, Terzi (2009)	Eagle, Pentland (2009)	Ye et al. (2009)	Wang et al. (2010)	Schaefer et al. (2011)	Ma et al. (2012)	Gresty et al. (2014)	Kirchler et al. (2016)	Galbraith, Smyth(2017)
Identification	File System											
	Operating System			o			o					o
	Application-level							o		√	√	√
Interpretation	Cross-referencing											
	Correlation and Association		√	√	√	√	√	√	√	√	√	√
Verification	Temporal	Synchronisation								o		
		Validity										
Activity Analysis	Intra-Session	Single-layer										
		Multi-layer										
	Session-to-session		√	√	√	√	√	√	√	√	√	√

Figure 2.10: Session-to-Session Activity Analysis Approaches

Li et al. [65] propose a geographic information system analysis based upon spatiotemporal similarity of users. The work in the paper proposes a hierarchical-graph-based similarity measurement, for determining the similarity between users, taking into account not only the geographic regions they accessed, but also the sequence that the regions were visited. The approach proposed in the paper is based upon data gathered from the global positioning system (GPS) locations and based upon the sequence of locations and the time taken at the locations, similarity of the users is first calculated and from co-existing patterns friends and a "community" of like-minded users can be inferred. The authors provide a case study with an example of using a simple threshold of time to identify the significance of a location. Is it a location where people choose to spend significant time or a place where people are caused to stop, such as waiting in a queue? The paper compares different users at different times, in the form of a session-to-session activity analysis.

Outside of the area of traditional digital forensics investigations, of interest are also the large event sequence files, such as those considered by Kiernan and Terzi [60]. They assert that large event sequences must be reduced and simplified to view, whilst at the same time give a global view of the activity and allow suspicious activity to be detected. Examples of this problem domain would be resource management, database optimisation and the authors do propose this technique for the analysis of large-size audit logs which need to be digested and displayed to an investigator. The authors show sequences of data that are broken up by type into what we can consider sessions where certain types of activity is more frequently occurring than other types. The authors present a segmentation and encoding scheme for the data, and once this has been segmented the results can be colour coded and displayed on a high-level timeline for simple viewing by an analyst. Different algorithms for the segmentation are shown and the results are demonstrated with each option.

Eagle and Pentland [33] assert that a person has structures, routines and patterns of behaviour, which when temporally, spatially and even socially contextualised can be easily identified. The authors term these underlying principal component-like behaviours as eigenbehaviours. In addition to identifying the components that represent the individual, they assert that a social group's behaviour can also be predicted based on how close or far the individual is from the social group. The premise is particularly interesting, but the experiments presented in the paper rely only on a sample of telephony data showing location, communication and levels of device usage of a selected group of individuals.

Ye et al. [104] propose a notation called life pattern normal form (LP-normal form) and a life pattern framework to determine and mine location-based data about individuals and their mobile computing habits. The authors of this paper propose that the patterns refer to significant places in an individual's daily life, but these must be extracted from raw GPS data, using "stay point" detection and clustering. The research identifies that there are sequential and non-sequential patterns within their non-conditional life patterns. For the conditional patterns these were considered life rules. E.g. at X location and only at that location, action Y occurs. The authors explicitly use the day as the unit of temporal granularity and work on life rules such as ‘work day’, ‘every Monday’ etc.

Wang et al. [100] propose that event summation, as can be seen in Kiernan and Terzi [60], does not reveal underlying properties of the patterns. The authors suggest that their approach is one of an educated guess at what processes are in play to produce the events and can be considered hidden concept learning, using Hidden Markov models. Their paper presents results of experimentation on synthetic and real-world data from an event log from a single computer, using different optimisation algorithms for comparison against their approach.

Schaefer et al. [90] discusses event sequences and makes some notable distinctions between the time-synchronous events, where a precise ordering of events is significant and between aggregate events, where a particular interval of time is important and the significant data is present during this period, but the precise order is not required. The paper presents different ways to visualise clusters of events, gaps and indeed shows representations which are not timelines, but only event information. Two examples are shown using fraud detection and keyword and content matching from news feeds. They show a visual analysis tool to present event data for auditing information based on rules.

Ma et al. [69] describe spatial location and behaviour pattern identification research. Locations and actions that are being performed on mobile computing devices can extrapolate user behaviour patterns and be used to identify other similar users operating within the same area. The authors note that there are problems with data sparseness so they propose grouping interactions based on location, such as home or work and types of activities such as email and games. The authors suggest that attempting to provide social context from precise location is not trivial. GPS signals may be turned off or unavailable indoors, and positioning based upon cell-tower locations is too imprecise especially in dense urban or sparse rural areas. The authors address this by classifying a number of possible social locations, and attempt a method to address the prior knowledge required to create the classified interactions for the users. Notably within the experiments presented, the data sessions are typically very long, over several months of data per user. General periods of the day are classified during the behaviour patterns, for example, does the entry occur between 0800 to 0900hrs, but explicitly breaking the patterns into sessions is not explored in this approach.

In Gresty et al. [40], a session-to-session comparison of artefacts from an interactive application, such as a web browser on a personal computer, allows statistical analysis for determining whether repetitive or habitual behaviour can be observed during the sessions of usage. The authors show that a user's Internet history can be processed to reduce a large data sample to a small number of principal components. These components can be clustered into sessions by temporal proximity to other principal components and then a like-for-like comparison of the sessions can be performed. The results of the experiments show that timelines of Internet history with large numbers of events can be digested into a simple table where an analyst can detect habitual behaviour by visually observing patterns and regularity to the data.

Kirchler et al. [61] address the problem of online privacy by demonstrating an approach to tracking user activity through Internet activity logs using behavioural fingerprints. The approach presented here does not use cookies or active click monitoring types of behavioural tracking, but rather this is an approach where sessions of network traffic activity are analysed. The session sizes used in these experiments are one day in length and the approach used to perform the analysis uses an unsupervised machine learning, which the authors assert does not require large volumes of labelled training data.

Galbraith and Smyth [34] use a statistical technique normally available in traditional real-world forensics to examine and compare two sets of sequential event data to determine if they originated from the same source. The experiments used in this paper tested real-world browser event streams and they used data from 28 individuals. The authors propose that further work and improvement of this approach can be done with suitable calibration and characterisation of the properties within the data

The approaches presented in this section are significant to our research, not least because our own early work in this area is present (Gresty et al. [40]), but because this work relates to comparing information from one period of time to other similar periods. The papers in this area are not necessarily drawn from what we would consider the traditional Digital Forensics "event reconstruction" field, and instead are based on research into spatial and/or temporal events.

2.4 RESEARCH QUESTION AND METHOD

2.4.1 The Research Issues

We have outlined that Context Analysis is ultimately about being able to consider artefacts together such that we can understand the behaviour of the system. So, given the motivation for this research outlined in section 1.3, we have to determine if there is a form of Context Analysis, and type of artefacts, that can be used to achieve these objectives? The overall approach we are presenting is a novel form of Analysis for Internet history artefacts. We presume that such types of artefacts are correctly identified and valid within the Context Analysis model.

We note that there is no approach within the 'traditional' Digital Forensics literature that takes a overall view of the actions on a system and shows how those overall actions may be useful to an investigation. Researchers such as Marrington [71], Hargreaves and Patterson [47] , James and Jang [56] etc. do note how the actions or traces of activity on a system can imply the technical actions that caused those traces, but they do not necessarily address the intent behind the actions.

We see in Amato et al. [5] that the authors make the same assertion we do in chapter 1, that increasingly the electronic device is a witness to a crime, but furthermore we assert that not only is the device an ‘eyewitness’ that can talk about the actions at a specific time, but it is also a ‘character witness’, like a spouse that can talk about long-term abuse, or the co-conspirator in a case of fraud. The credibility of such a witness is important but they provide very compelling testimony relating to the intent of an offender. Stepping beyond the eyewitness approach allows us to look at a much wider open field of research, which we see is important for our motivation and objectives (section 1.3), where we note that there are offences that a traditional ‘eyewitness’ would not be suitable: the single events that occur multiple times and multiple events that form a single crime. This therefore showed that there was a motivation to move beyond the traditional Digital Forensics body of literature and to incorporate other types of Activity analysis.

As part of our Context Analysis we have incorporated work such as Ma et al. [69] which deals with spatial location and behaviour, Schaefer et al. [90] deals with event sequences and the order of events and Eagle and Pentland [33] investigate the interaction between social groups and so on. These areas of research relate to comparing information from one period of time to other similar periods, and are not necessarily drawn from what we would consider the traditional Digital Forensics “event reconstruction” field. This widening of the research techniques is necessary to deal with the new types of digital forensics crime scenarios we present as motivation for our research.

2.4.2 Methods in this Thesis

In Chapter 3 we investigate the choice between an Intra-session type of Activity Analysis and the Session-to-Session approach. We describe in detail methods of aggregating data into ‘Sessions’, namely the fixed-length and variable-length approaches, and we show that by plotting the number of sessions whilst varying the fixed-length and variable-length thresholds, we can show appropriate settings for our session aggregation thresholds, without any prior knowledge. We also introduce the backbone of our research, the session-to-session comparison using the Jaccard similarity coefficient.

In Chapter 3 we also address the need for suitable data that can be used to test the appropriateness of our approaches, which does not simulate any specific ‘crime’, but is of a suitably large volume, that the data is marked as coming from one source or other and finally showing that the data is suitable to represent ‘normal’ Internet history, which we show using Spearman correlation to a measure of popularity for websites (which is properly explained in chapter 4).

Chapter 4 considers the problem that two or more users of a single device may have similar interests and consequently Activity Analysis can draw incorrect conclusions about the provenance of two periods of time. We describe in this chapter variables that can be adjusted in the Internet history session data to improve correctness, namely the *s-val*, *t-val* and *c-val*. We present novel approaches to breaking up Internet history session data without prior knowledge of the users or of the types of data by showing the length of the session can identify types of behaviour and the Relative Popularity method which uses a reference source and the difference from that reference point to the actions on the system dictate how the data is grouped.

In Chapter 5 we demonstrate that the approaches used in chapters 3 and 4 can produce overall sets of results with sufficient accuracy, which we call the Beyond Reasonable Doubt (BRD), at 91%. The results of the session-to-session comparisons (chapter 3) of the grouped data (chapter 4) is placed onto a graph and Louvain Community Detection is used to cluster and colour the graphs. We then present how an analyst or investigator could use such graphs for Pattern of Life detection, answering Investigative Hypothesis and varying the *s-val* and *t-val* variables to show strength of connection between two or more sessions.

2.5 CONCLUSION

We have shown that there is a broad literature in Digital Forensics that considers tools and techniques in a way that we describe as Context Analysis. Techniques that collectively examine forensic artefacts in conjunction with other artefacts, so as to describe, test the validity or determine the behaviour in time and or space of those artefacts, we would describe as contextual.

The alternative to Context Analysis is to test the metadata of the artefact in isolation, such as searching for a specific keyword, hash value, or to test the content of the artefact.

A system that uses context analysis need not provide all of the elements that we present here to be a context analysis technique, but the analyst must be aware that for example, in a multi-layer intra-session activity analysis technique, there must be trust that the correct artefacts have been identified, are valid and have been interpreted correctly, even if those elements are not explicitly tested contextually.

In the following chapters, we propose a system that uses Internet history artefacts as a Session-to-Session Activity Analysis that does not explicitly use any other form of contextual temporal validation, and the identification methods are addressed due to this being aggregated from application level.

Chapter 3

SESSION-TO-SESSION ANALYSIS

“I have now finished with the ungrateful task of criticizing, and I proceed to propose a system which it is hoped will be as severely criticized by others.”

Sir Richard Francis Burton - A new system of sword exercise for infantry, 1876

3.1 INTRODUCTION

In Chapter 1, we proposed that there are a number of challenges within Digital Forensics Science. In this chapter we are interested in investigating the following:

- Investigate ways in which the behaviour that is recorded on the device can be automatically grouped in such a way that it can be analysed that is both correct, and sufficiently simple that it is able to be transparently described to a court or jury.
- Show if there are specific events than an investigator is interested in, how patterns can be identified within the data relating to those events, which may assist in determining the *mens rea* of a suspect, if at all possible.
- Facilitate identifying features within the digital data that can be used to identify the user or users of a device

- Investigation into multi-user desktop environments that is connected to the Internet, where there is potentially weak user authentication or account sharing. This represents a realistic and challenging environment for digital forensic science.

In Chapter 2 we therefore explored the literature of what we described as Context Analysis, which considers event artefacts that can and should be collectively analysed together. Considering artefacts together is essential to building the weight of circumstantial evidence, that we noted in chapter 1 is the practical way in which Digital Forensic Science answers the question about the user performing the actions on a device.

In this chapter, and for our thesis, we have looked at a highly interactive set of event artefacts that would be present on our goal target system, and would likely contain personal and identifiable features: namely the Internet history artefacts that are left on a device as part of the normal use of browser software. A detailed description of the data sets used in this thesis can be found in section 3.9 of this chapter.

We assert that the different individuals who use a device will have characteristic patterns of behaviour that can be identified. Therefore, if there were specific events that comprise the *actus reus* of a crime, we could identify if there were personally identifiable patterns that intersected with the act, such that it could indicate the user at that time. If there were other notable periods of time that interacted with the act, they could show planning, performing or researching the outcomes of a crime, i.e. the *mens rea*.

3.2 INTERNET HISTORY ARTEFACTS

The Internet history records are single point events showing access to a resource/page, and they do not contain the content of the page/site/resource, but may be paired with the content stored elsewhere within the web cache of the device. Investigators typically can find Internet history artefacts from the unallocated area of previously used areas of hard disk drives, file slack, within page files, shadow copy file structures and any other forms of backup. Given that Internet history may also be recorded at the firewall, gateway or service provider it is quite possible, indeed in our experience it is extremely common, that a Digital Forensic Scientist will be analysing a corpora of Internet artefacts with only the point event data without any of the content to refer to.

From these individual points, which contain a date and time and the address to the resource that was being accessed, an analyst can imply a period of continuous usage by considering closely occurring point events, separated by short intervals. Figure 3.1, illustrates an example of such history drawn on a timeline with the point events as black lines.

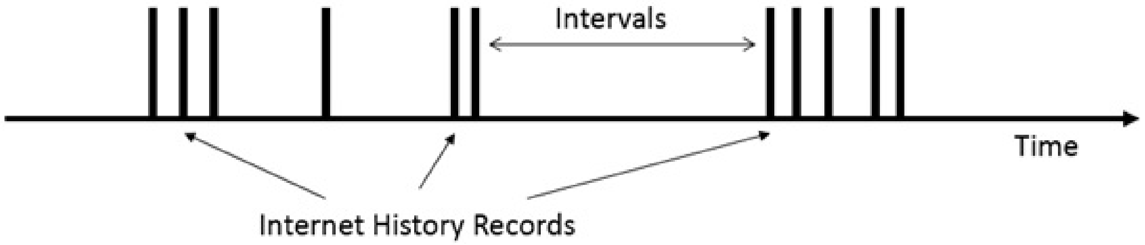


Figure 3.1: Internet artefacts are point events on a timeline

The address component of the Internet history records can have three levels of resolution and verbosity depending upon the source of the artefact: Host-level, Page-level and Element-level.

Host-level Resolution - At this resolution the only information that is retained is the address of the host where a page was accessed. The pathway to the individual pages or elements is not retained and this type of resolution is typical of the type of data saved at firewall/gateway logs and is the type of data that will be retained by Communication Service Providers as part of the Investigatory Powers Act 2016 [110].

Page-level resolution - This resolution of data is retained within a typical web-browser 'History' and displayed to a user when they view the previously visited sites. This level of data shows which page was accessed but it does not show details about how many pictures, what JavaScript code was executed or if there were sidebars on the page using data from other sites.

Element-level resolution - The low-level data in a web-browser 'Cache', that is not shown at the Page-level resolution is shown at the Element-level. This level retains the pathways to the scripts, hyper-text files and pictures that are needed to construct the pages that are being accessed.

Some pages may contain elements from hosts that are wholly different to the host that is calling the element. For example, if there is a page that is being accessed:

www.unknown-site.com/page.html

This page might contain pictures or content from other hosts that are known and are being blocked through conventional firewall or service provider Host-level filtering.

www.known-illegal-site.com/pictures/1.jpg

The approach used in this research project is to use a Host-level view of the data, but that view is constructed from any of the available Internet history. For example, if the analyst has an Element-level Internet history that has been recovered from a device, the trailing page or element details will be stripped off such that only the Host-level details remain.

This approach has the advantage that it largely complies with the scientific law of comparative analysis, such that all three levels can be compared against each other, which will enable the mixing and comparison of evidence sources to see if there is missing or obscured data, and 'cross drive analysis' types of problems where there is an attempt to identify similar artefacts across different media sources. There is a difference in that some websites and pages have a high degree of dependence upon each other, i.e. going to one web page will always access resources from another website. For example, A regional shop web page selling an item could have the page:

www.shopsellingitem.co.uk/washingmachines.htm

Which contains elements that refers back to the global headquarters website:

media.shopsellingitem.com/superdeluxmodel.jpg

As such when we take this Element-level history and render it down to a Host-level history we will see the 'shopsellingitem.com' and 'shopsellingitem.co.uk' components appearing together in all sessions. If an analyst was then to compare a second set of data that was natively Host-level or rendered down Page-level history, with the data that had been rendered down from the Element-level history, the analyst would see a difference as there would be one set of sessions that contained 'shopsellingitem.com' and 'shopsellingitem.co.uk' and another set of sessions that did not ever show the global 'shopsellingitem.com' component. Therefore, care must be used if recovering Internet history that has come from different levels.

3.3 SEQUENCE OR AGGREGATE

Schaefer et al. [90] outlined two useful methods of analysing temporal data, the sequence approach and the aggregate approach:

- Temporal sequence comparison. Patterns are identified within an ordered, typically long, sequence of data.

- Aggregate-against-aggregate comparison. A collection of grouped artefacts is compared against another collection of grouped artefacts.

In chapter 2, we have identified within the Activity Analysis section of our Literature Review that there are event modelling approaches that are sequence-based and there are aggregate methods. Within this project, we have investigated sequential analysis of Internet history data and noted that overall there are two main problems:

- Issues relating to how the webpages are technically implemented.
- Issues related to user interaction.

Web pages come in a variety of styles: large blocks of text, interactive 'flash' pages, short text blocks that require a user to navigate through pages that could easily have been presented on a single page, thumbnail gallery picture pages etc. Developing rules based upon number of pages visited, speed of navigation, pictures viewed etc. is in effect a profile not of the user, but a profile of the technical implementation of the websites. Whilst that could be an area of interesting future research, especially with regard to commonly used websites, it would necessitate the actual content of the websites visited available for comparison. It may be possible to profile individuals, how they viewed, used, accessed the pages, but such an analysis goes well beyond the scope of what is commonly available within Internet history listings and the scope of this thesis.

We have found it an extremely complex problem trying to create sequential rules that can identify repetitive behaviour based upon the realistic experiences of users interacting with their web browsers. A user that is a regular visitor to a particular sport/activity/interest website, i.e. a high degree of probability of visitation to that site, may still visit the site at the beginning, middle or end of the session. If there are a variety of sites that the user may or may not visit, and they may visit them in any order, and they may visit any number of other sites during the session, then the reliability of the rules that can be deduced and the amount of replicability of those rules is in our experience low.

Therefore, in this thesis, we have looked at a simple aggregate method: If a user visits a website during a session (sessions are defined in the next section), once, ten times in a row, or visited and returned to at a later time in the session, it is considered only a single time. Therefore, this approach ignores implementation issues, it ignores the variability of the person's behaviour and ultimately we will show that the sessions that share commonly visited websites can be simply identified and compared against each other to determine how similar the overall aggregate behaviour was at that time. In our further work, we propose that once we have highlighted sessions that appear similar to each other using this aggregate method, we can then use sequential analysis to further investigate the behaviour at those periods of time.

3.4 SESSIONS AGGREGATES

For the analysis of Internet history timelines, or for any meta-data context analysis within digital forensic investigations, we propose an approach where 'session' temporal aggregates are compared against other 'sessions' to identify to what extent any of the sessions contain matching members or components. Once sessions have been compared and the like-for-like sessions have been grouped together, then the process of intra-session sequence analysis may be performed if it is so desired to identify whether specific patterns of components appear. This session-to-session grouping itself provides significant macro-level contextual analysis about the use of a device at any time, and temporal sequential analysis after this analysis, substantially reducing the quantities of sequential data to be processed.

The selection of the session temporal aggregates is therefore fundamental. We identify two approaches to selecting sessions:

- Fixed length sessions. Fixed periods of time are selected in advance, for example all artefacts in a window of 30 s, 60 s, or 60 min.

- Variable length continuous activity sessions. If two artefacts are closer together in time than a predefined temporal threshold, they are considered to be in the same session. Otherwise, the second artefact is considered the start of the next session.

The variable length approach organically follows the activity from beginning to end of the session without artificially breaking up long sessions into smaller chunks. However, like-for-like comparison between sessions is open to some interpretation when using a variable length approach. Two sessions which could have the exact same component members and look at face value to be the same, could have very different characteristics. For example, one session being two or three times longer than the other and having quite different behaviour at the beginning and end of the session.

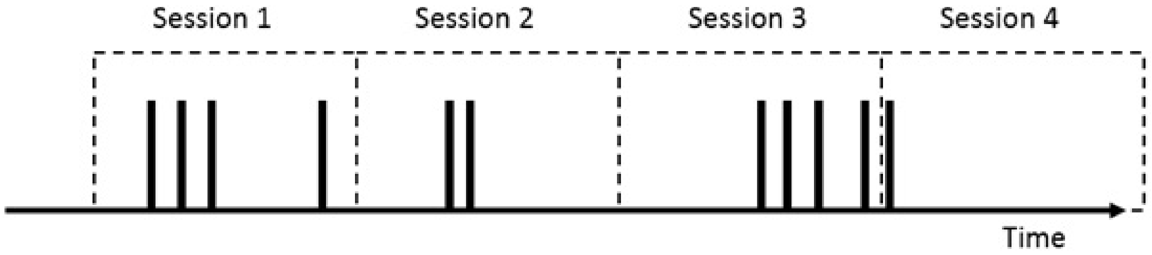


Figure 3.2: An illustration of Fixed-length sessions

The biggest disadvantage with using a Fixed-length approach can be seen in figure 3.2, where it can be seen at the boundary between Session 3 and 4 where there appears to be contiguous activity that is lopped into two different sessions, which could create a misleading pattern.

In the Variable-length approach, if two artefacts are closer together in time than a predefined temporal interval threshold, they are considered to be in the same session.

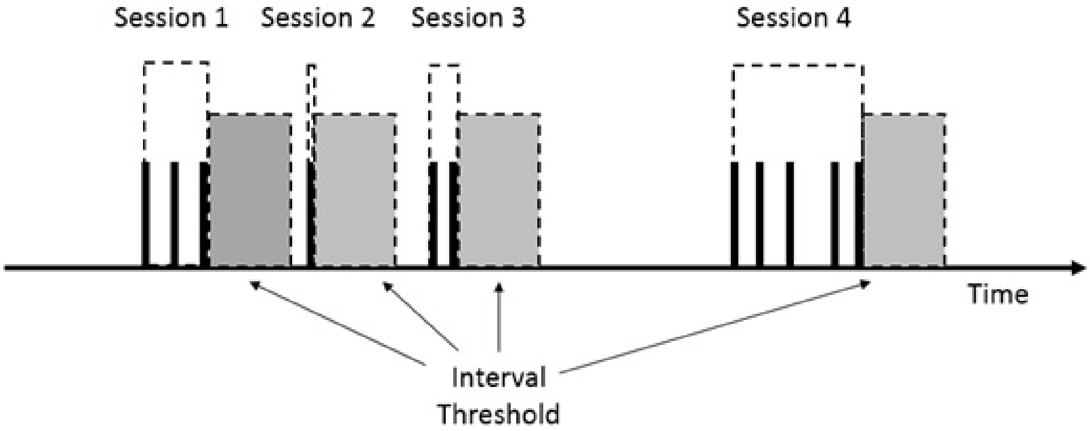


Figure 3.3: An illustration of Variable-length sessions

Figure 3.3 illustrates that if the fixed interval threshold (grey box) can fit between the point events then that is the point where the sessions are segregated. The advantage of this can be seen in that it does not arbitrarily cut-off the contiguous activity as can be seen with Session 4. The disadvantage is the opposite to the fixed-length sessions in that all of the sessions are different in length and have different start times etc.

The problem with using sessions is capturing the right amount of information that represents the 'behaviour' that is taking place at the time. The simplest example of this is where two users share the same user account on a computer, but each uses the computer for accessing very different website interests. Choosing a very large fixed-length size could easily capture the usage of the computer by both users if there is a short window of time where they swap over usage, when there is very likely a desire to try and isolate the different access habits. Figure 3.4 illustrates two large fixed-length sessions where there are two different users creating Internet history, and because of the large size selection of the session aggregation, the first session contains mixed User 1 and User 2 artefacts.

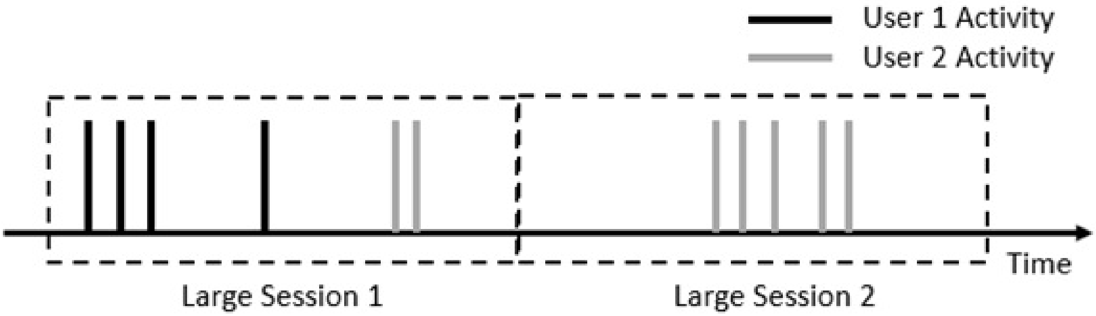


Figure 3.4: An illustration of incorrect session selection

There are a variety of levels of precision when dealing with digital timestamps as noted in Oh et al. [75]. Oh's second-level of precision (i.e. log files are stored at the 1 second level of precision rather than at the microsecond level or larger) 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. During our experiments, tests have been performed on larger time window aggregates than the 1- second level of precision, such as cases that contain all the events within a 5, 10 or 30 up to 3600 seconds (60 minutes).

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 occurring before the web artefacts appear on the computer.

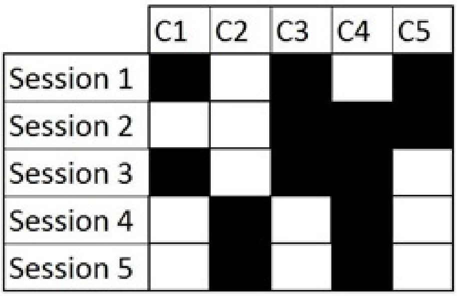
3.5 BINARY COMPONENTS VERSUS INTEGER COMPONENTS

Components are the individual distinct actions that are recorded within the sessions. In file system analysis, a component could be each directory or each file creation, modification or access to a file type. In a wider pattern of life analysis of a home automation system, the activation of lights, devices or other sensors could be recorded as components. Here, we focus on Internet history components and consequently the host-level website details are used as components. We can define two kinds of Components:

- Binary Components - if there is a visit to the website host once or multiple times during the session, the session is marked as a black box in our diagrams, an example is shown in figure 3.5.

- Integer Components each instance of the website host occurring during the session is recorded (the notation we originally used was a grey box with the integer inside the box).

The Integer components suffer from the implementation issues noted above in Sequence Analysis, and they are unsatisfactory when comparing session like-for-like behaviour. Individual components that create larger numbers of artefacts or require the user to navigate more will ostensibly seem more important or more frequently used than another site which could in actual fact be visited more frequently, but create fewer artefacts.



	C1	C2	C3	C4	C5
Session 1	■		■		■
Session 2			■	■	■
Session 3	■		■	■	
Session 4		■		■	
Session 5		■		■	

Figure 3.5: Five sessions containing five binary components, C1 to C5

As we noted above, the act of visiting certain sites would be more important than the number of times it is visited. For example, if trying to attribute a particular session to a specific user who is known to be a motorcycle enthusiast, the number of times that a motorcycle-related website is accessed is substantially less significant than the fact that the motorcycle website was accessed at all. Therefore, not only is recording an Integer component unhelpful, but it can also mislead. Take another example: During session 1, site A is visited once; During session 2 site A is visited twice. The similarity between session 1 and session 2 is 0.5 as there were only half as many visits during session 1 as during session 2. The problem is that the oversampling, which is again a challenge within the analytical Forensic Science laws, has completely obscured the important similar behaviour that the same website was accessed.

Consequently, for the aggregate Session-to-Session comparisons, our research has focused solely upon using Binary Components (which we refer to only as Components throughout the rest of the thesis) and we leave concerns about frequency of visits to a website to the Intra-Session or Sequential Analysis phase that we note in future work.

3.6 SESSION-TO-SESSION COMPARISON

The basic presumption for all our Session-to-Session comparisons is that if we match a session to all of the other sessions in the dataset, which contains data from more than one individual, the highest matching sessions to the one we are testing should have been created by the same user.

By creating a binary condition for components, a simple visual display can be made for the components per session as can be seen in Figure 3.5, which shows an example set of data containing five components (C1 to C5) and five sessions. Session 1 to 3 represent user 1, whereas sessions 4 and 5 represent user 2. Even with this example small set of data the repetitive pattern in sessions 4 and 5 and somewhat in sessions 1 and 3 stand out well visually.

The sessions, however, form a simple string which can have a pairwise distance comparison. For example, session 1 [10101] and session 2 [00111] can be calculated to have a distance of 0.5 using the Jaccard similarity coefficient [52] which is the size of the intersection of two sets divided by the size of their union:

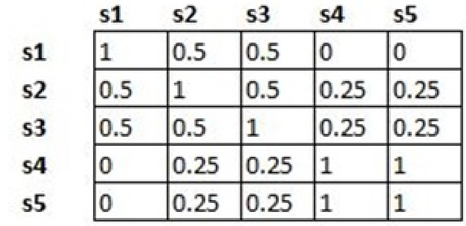
$$d_j(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Figure 3.6: Jaccard Distance

The advantage of using Jaccard is that it only considers the components in sessions 1 and 2 that they share and does not consider C2, which is 0 in both cases.

Other methods for performing pairwise comparisons are available, such as Hamming distance [46], Levenshtein distance [64] or Sorensen index / Dice's coefficient [31].

When dealing with some hundreds or thousands of components all of them 0's such as seen in typical, frequently used home computer Internet history, use of Hamming or Levenshtein distances are undesirable as they will produce similarity coefficients of 0.999 as all of the 0's will be considered exact matches. Sorensen index / Dice's coefficient are functionally similar to Jaccard, but as we are using binary components Jaccard is easy to understand and explain to a court/jury.



	s1	s2	s3	s4	s5
s1	1	0.5	0.5	0	0
s2	0.5	1	0.5	0.25	0.25
s3	0.5	0.5	1	0.25	0.25
s4	0	0.25	0.25	1	1
s5	0	0.25	0.25	1	1

Figure 3.7: Jaccard Similarity Table for figure 3.5

Patterns are constructed by identifying groups of two or more sessions that are above a Jaccard distance measure. Although any value above 0.0 is potentially useful, the number of loosely associated sessions significantly increases as the acceptable Jaccard value is lowered. For example, at a level of 1.0, one session pattern is created: Pattern 1 = [s4 s5]. At a Jaccard level of 0.5, two session patterns are created: Pattern 1 = [s1 s2 s3], Pattern 2 = [s4 s5].

3.7 SELECTING THE TIME THRESHOLD

Selecting the total length of the session when using the Fixed-length approach, and the temporal interval threshold when using the Variable-length approach has a significant effect on the number of sessions that are available for analysis, and the accuracy of the comparisons.

Figure 3.8 shows an example of experiments performed using the Variable-length approach on the seven individual sets of data that are combined together to form our test data sets (which are detailed in chapter 3.9). We plotted the number of sessions that were available using a 1 second Variable-length aggregate, up to 3600 seconds (1 hour). We can see that the rate of change in the number of sessions is on the whole more interesting that the final proportion of sessions that are available. The data plotted as Series6 was a small set of Internet history, and consequently we see that it falls off much less drastically than the data plotted as Series3, which was from a large set of Internet history over an extensive period of time.

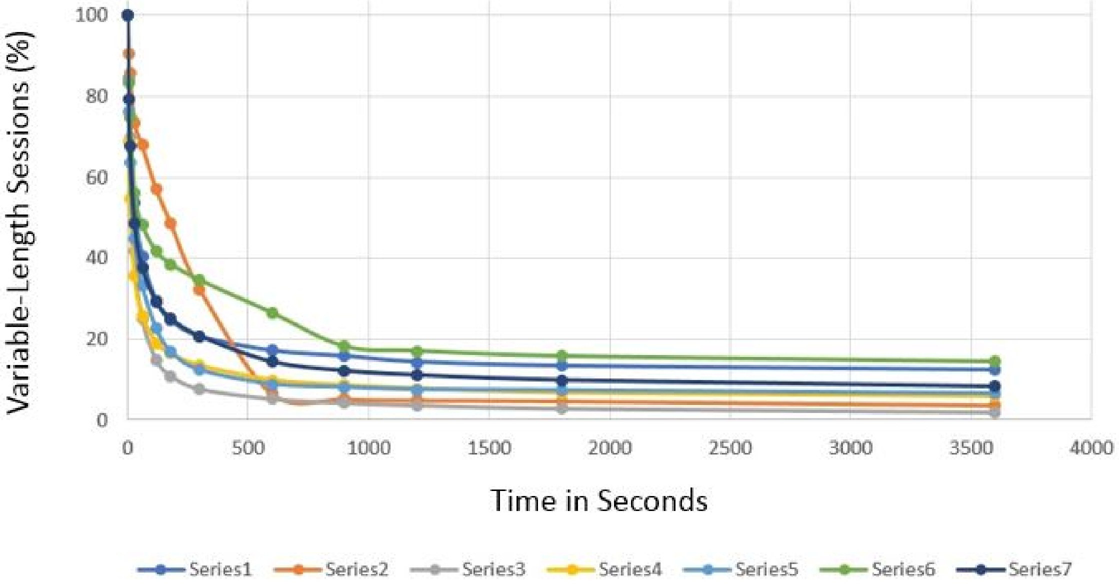


Figure 3.8: The overall percentage of sessions available for analysis based upon different interval times for the Variable-length analysis approach

The rate of change in the number of available sessions falls rapidly as one would expect, but depending upon the data we see from around 300 to 600 seconds (5 to 10 minutes) the number of sessions has largely stabilised, and at 900 seconds (15 minutes of idle time) we see that there is very little change. For our experiments, we therefore selected the 900 second threshold as this experimentally suited the mixing of data, without having to select too great a threshold. Operationally (i.e. without having to know the ground truth) an analyst could process the Internet history by looking at rate of change and where appropriate select a shorter threshold than 900 seconds. For example, if the analyst was processing data similar to that plotted as Series3 or Series5 then they might want to use a 300 second threshold, while if it was more like Series2 or Series1 they might opt for 600 second threshold.

To illustrate the difference in number between Variable-length sessions and Fixed-length sessions we can look at two examples using the 'ZR' dataset and the 'ZS' dataset. In [41], we showed two different datasets and performed the fixed-length and variable-length session selection. The 'ZS' set comes originally from the Digital Corpora project [103], [115] and is a set of user data that is based upon three small-size workplace computers using cache-level recovery of the Mozilla web browser. The 'ZR' dataset is created from two large-size real user's Internet history based upon home/domestic usage recovered from cache-level extraction of Internet Explorer. These sets of Internet history have been constructed using a method similar to those described later in section 3.9, i.e. from two sources in the case of the 'ZR' dataset, and from three sources in the case of the 'ZS' dataset.

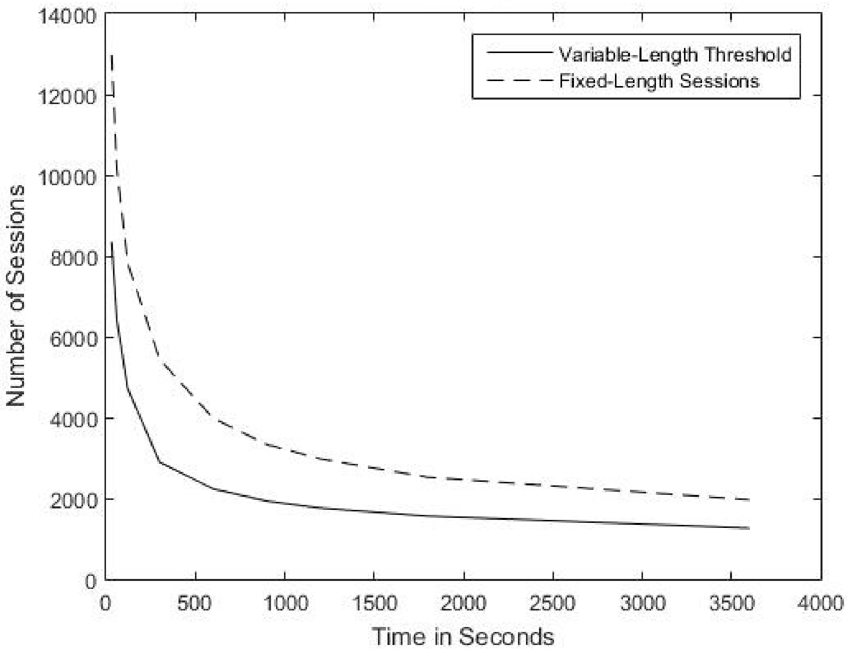


Figure 3.9: The ZR dataset showing the number of Fixed-length and Variable-Length sessions

Figure 3.9 shows the number of sessions that would be available for the ZR dataset if Variable-length and Fixed-length methods were each used to aggregate the data.

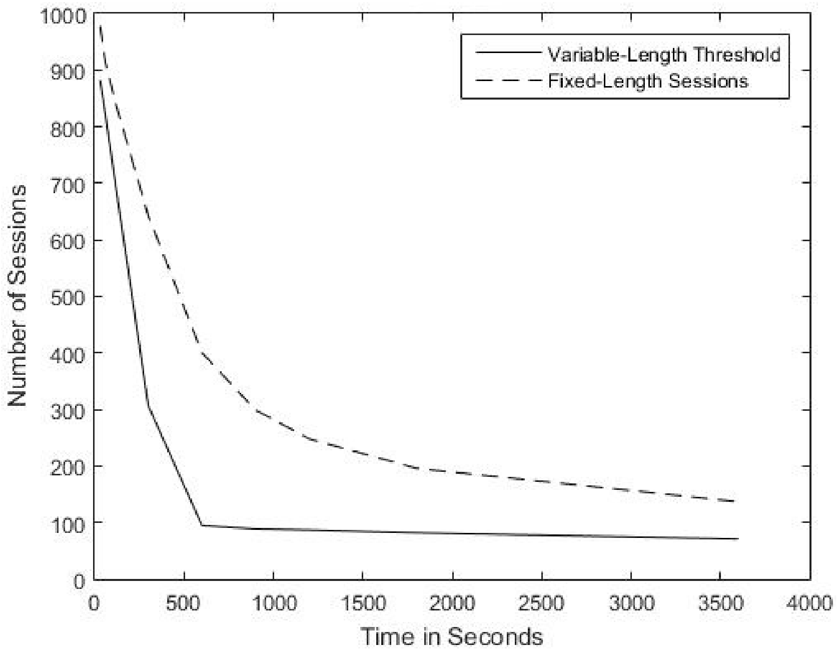


Figure 3.10: The ZS dataset showing the number of Fixed-length and Variable-Length sessions

In figure 3.10, we see the number of sessions that would be available for the ZS dataset if Variable-length and Fixed-length methods were used to aggregate the data.

For both the ZR and ZS dataset we see that there are more Fixed-length created sessions and although there is initially a steady fall-off in the number of available sessions, both the ZR and ZS cannot be said to 'stabilise' over the 3600 second windows. We can see that when using the Variable-length method, the ZS data set stabilises at about 100 sessions, which is much smaller than the 1800 or so sessions that the ZR dataset stabilises at when also using the Variable-length approach.

The Variable-length approach produces fewer available sessions, but these do stabilise, again both datasets show between 300 and 600 seconds is a fair point to consider Variable-length thresholds for this data.

3.8 ACCURACY OF THE SESSION SELECTION METHODS

Here, we show the results of experiments to compare the Variable-length approach and the Fixed-length approach using the two simple datasets (ZR and ZS).

The method used for these experiments was to create sessions using different temporal thresholds, perform a Session-to-Session Jaccard similarity analysis and find how many correct matches were created at a specified threshold level (which we define as *t-val* in section 3.10.1), i.e. how many correct matches occur if the session-to-session overlap threshold is 0.25, 0.5, 0.75 and exact matches of 1.0. We have presented the results of these experiments as two types of graphs:

- The **percentage** of correct matches in the Y axis, the temporal threshold in the X axis and the *t-val* session overlap threshold in the Z axis.
- The **total number** of correct matches in the Y axis, the temporal threshold in the X axis and the *t-val* session overlap threshold in the Z axis.

The ZR percentage graphs (figure 3.11 for the Fixed-length approach and figure 3.15 for the Variable-length approach) appear to show that as both the temporal and *t-val* thresholds are raised the level of accuracy increases dramatically, but we see in the total number graphs (figure 3.12 for the Fixed-length approach and figure 3.16 for the Variable-length approach) this is because the number of available sessions has dropped off to a very small number. As we raise the precision, we drastically reduce the number of sessions we have available to analyse, but we have much greater confidence that those sessions are correct.

We see in the ZS graphs (figure 3.13 for the Fixed-length approach and figure 3.17 for the Variable-length approach) a different performance to the ZR graphs, however we do still see an element that as precision is increased, availability of sessions is decreased. We see that there is generally a point between 600 and 900 seconds, and above the *t-val* 0.75 where there is good overall performance in the percentage of correct sessions, and reasonable availability in the number of sessions. This also corresponds to the temporal point in figure 3.8 where the rate of change for number of sessions has stabilised.

When looking at the ZS data (figure 3.13 for the Fixed-length approach and figure 3.17 for the Variable-length approach) the Variable-length approach does appear to perform better with a greater percentage of correct matches than the Fixed-length approach, and similarly for the ZR data (figure 3.11 for the Fixed-length approach and figure 3.15 for the Variable-length approach).

Therefore to simplify the numbers of variables for our experiments, the approach used for the remainder of this thesis is based around the Variable-length approach, for the reasons illustrated in this section.

Graphs plotting the Jaccard similarity value (Z axis), Time interval or threshold (X axis) and percentage or total number of correct matches (Y axis) - the graph is coloured in greyscale from low results (dark) to higher results (light).

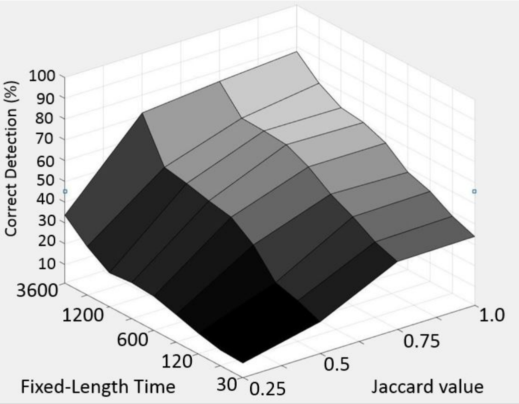


Figure 3.11: The ZR dataset showing the correct PERCENTAGE of matching patterns, using the Fixed-length approach

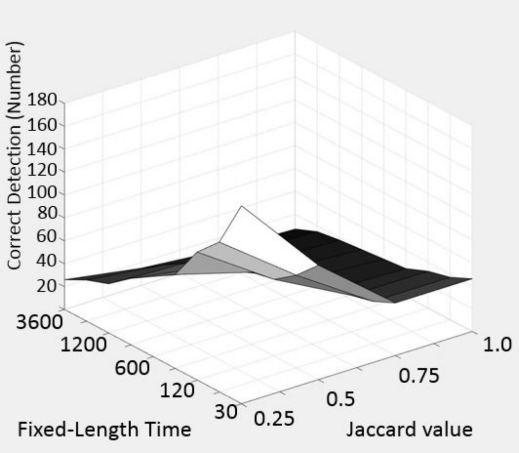


Figure 3.12: The ZR dataset showing the correct TOTAL NUMBER of matching patterns, using the Fixed-length approach

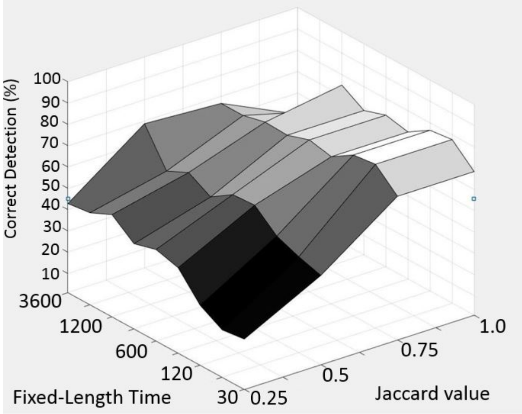


Figure 3.13: The ZS dataset showing the correct PERCENTAGE of matching patterns, using the Fixed-length approach

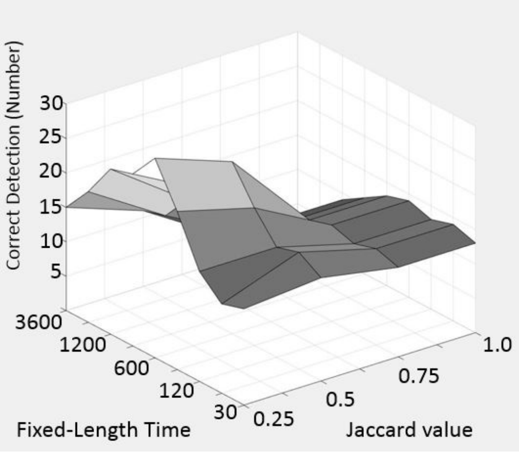


Figure 3.14: The ZS dataset showing the correct TOTAL NUMBER of matching patterns, using the Fixed-length approach

3.9 DATASETS USED IN THE EXPERIMENTS

The experiments to test the theories in this thesis and to illustrate the techniques that we propose were initially tested on the ZS and ZR test datasets. To ensure that

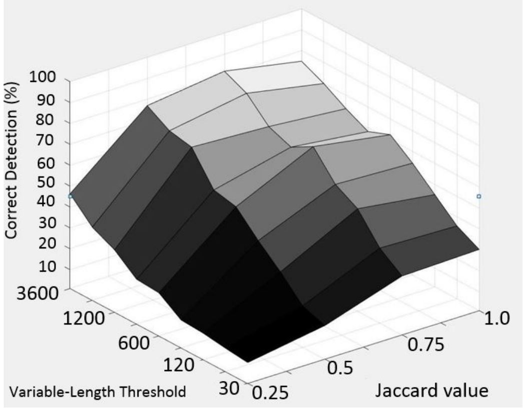


Figure 3.15: The ZR dataset showing the correct PERCENTAGE of matching patterns, using the Variable-length approach

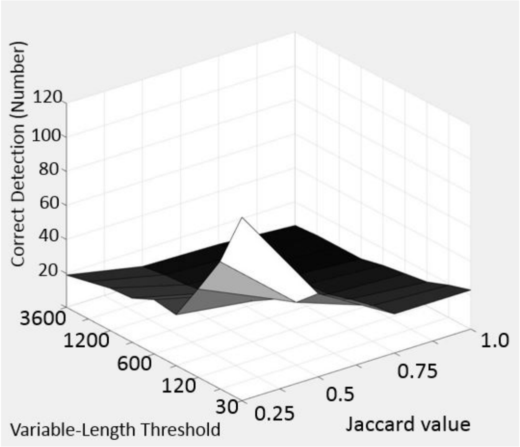


Figure 3.16: The ZR dataset showing the correct TOTAL NUMBER of matching patterns, using the Variable-length approach

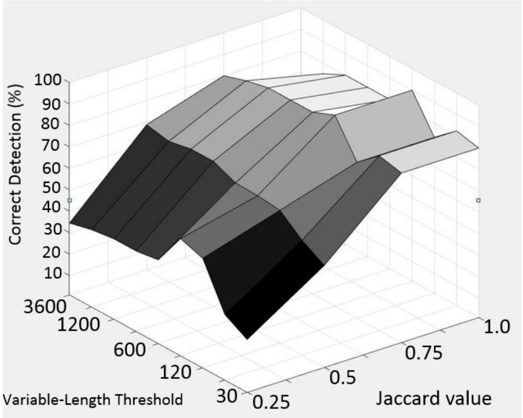


Figure 3.17: The ZS dataset showing the correct PERCENTAGE of matching patterns, using the Variable-length approach

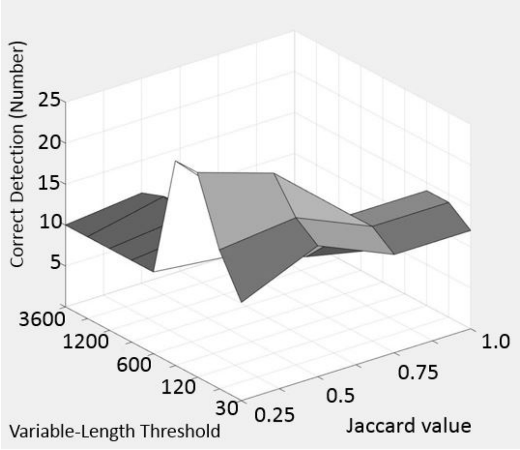


Figure 3.18: The ZS dataset showing the correct TOTAL NUMBER of matching patterns, using the Variable-length approach

the ground truth was known, these two datasets came from known sources. The ZS data was manufactured from test data provided from the Digital Corpora project, and the ZR data was taken from the hard drives of people that knew they were providing Internet history data for testing, but were asked to "behave normally". In these experiments we knew the ground truth of who had made the artefacts, but for the large-scale testing which we present throughout the remainder of this thesis, we selected new sources of data, where there was no element of bias, modification or construction of the data.

3.9.1 SELECTION OF SUITABLE DATA

We are not trying to simulate any specific scenarios, but we need data that is consistent with large-scale device usage and interaction, such that it would be relevant to the investigation of cases where it would be beneficial to model the overall activity, as outlined in section 1.3.

Grajeda et al. [39] surveyed the Digital Forensics field and examined how many datasets were publicly available for researchers to use, and also examined the need and impediments for the sharing of datasets. Grajeda et al. categorised datasets into 'computer generated', 'experimentally generated' and 'real world'. We can see from the datasets surveyed in this paper that there are no standard sets of data that can be used to explore Internet history analysis and consequently to overcome any concerns relating to researcher bias being introduced with 'computer generated' and 'experimentally generated' dataset, we have opted for 'real world' data.

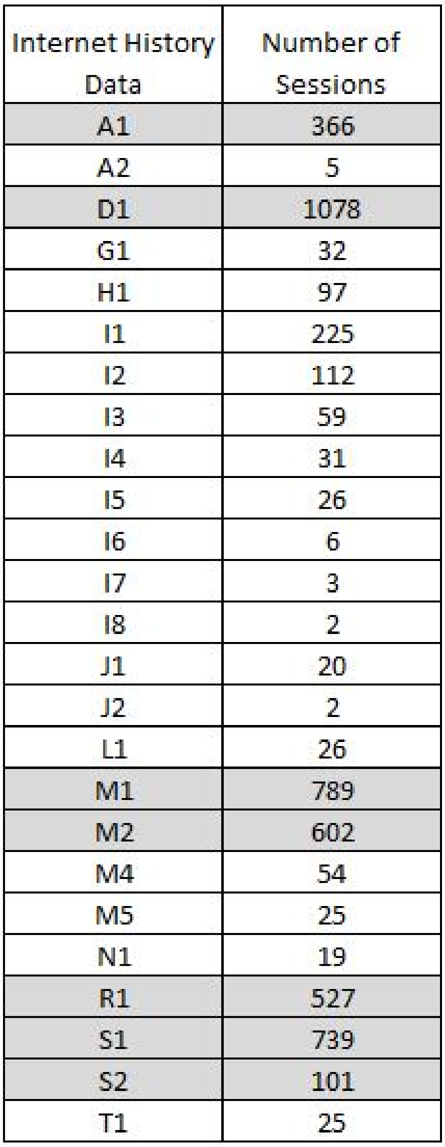
The test data must show the activity of the users' Internet History that was recovered from hard-disk drives (although in theory it could come from Internet Connection Records) using standard Internet History recovery tools. To ensure that the researchers have not introduced bias or expectation of what 'normal' web browsing should be, we have during data selection acquired the data from normal desktop and laptop computer hard disk drives and there has been no selection on the type of data contained with those devices, other than a suitable quantity of data for testing:

- No attempt to perform selection based upon any criteria within the dataset.
- No attempt to select on the number or type of host websites.

Because we have not performed any selection within the datasets we cannot determine how representative they are of 'normal behaviour', however we do not currently see any evidence within the research literature that there is a single type of 'normal' behaviour when it comes to usage, activity or users. We present a statistical analysis of the datasets below (see section 3.9.3) which indicates that the datasets do correlate with observable patterns of normal website access.

A selection of hard-disk drives were acquired, the Internet history was searched and cache-level Internet history records were recovered (see chapter 3.2 for a description of cache-level). In all cases, the Internet history records were for the 'Internet Explorer' web browser with the exceptions of D1 (which was Google Chrome), M4 and M5 (which were Mozilla Firefox).

Figure 3.19 shows the number of variable-length sessions (see chapter 3.4 for Variable-length description) that are created when using a threshold of 900 seconds (i.e. 15 minutes of idle time to delimit the sessions). In figure 3.8, we can see that 900 seconds is a general-purpose threshold, where there is very little change in the number of sessions after that point, which we selected for consistency across the experiments. The seven series of data plotted in figure 3.8 are the seven sets of data that we highlight here and were used to construct our test data sets.



Internet History Data	Number of Sessions
A1	366
A2	5
D1	1078
G1	32
H1	97
I1	225
I2	112
I3	59
I4	31
I5	26
I6	6
I7	3
I8	2
J1	20
J2	2
L1	26
M1	789
M2	602
M4	54
M5	25
N1	19
R1	527
S1	739
S2	101
T1	25

Figure 3.19: Number of Variable-length sessions available in the Internet History Datasets when using 900 second thresholds. The highlighted histories are used in our research

We can see that there are major sets of data with several hundred periods of time accessing the Internet, minor sets of data representing a small number of access up to a hundred accesses on a single machine and trivial sets of data that only represent a couple of access to the Internet.

The selection criteria when creating the experimental datasets was to identify the large sets of data and combine them in such a way that there was minimal chance of an inadvertent overlap between the original sources. For example, The S1 set may have come from the same source as the A1 set, therefore S1 was combined with D1 which was known to come from a different time and location. The use of the large data sets were determined to be most useful because the experiments were investigating repetitive behaviour and therefore using large quantities of data allowed us to look at substantial clusters of similarity.

The datasets we have used for the experiments address the three needs for modelling of the whole system that we highlighted in section 1.3:

- Where there is an identified criminal event co-located in time with a body of Internet history the important feature here is to consider the volume of available Internet history that the system can analyse. The 'Pattern of Life' approach presented in section 5.7 could be used to provide a pattern of the co-located events and cross-reference this with the impacted communities.

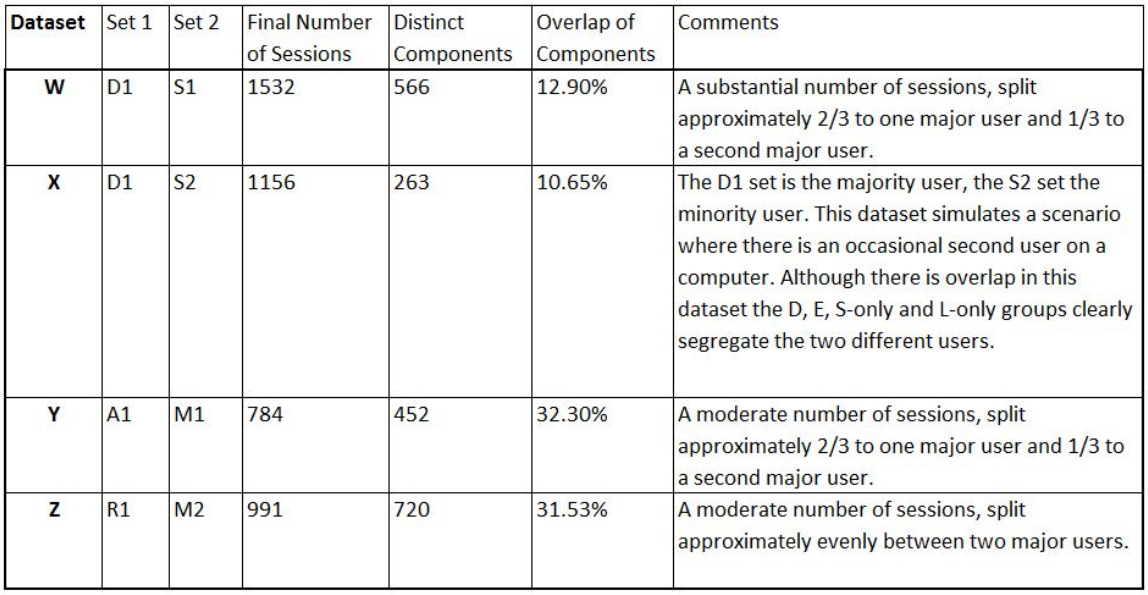
- Where there is Internet history containing the criminal activities The approach presented in this research lends itself very well to this analysis and if certain components within the data was notable to an investigator, then whole communities could be designated as notable.

- The Session-to-Session relationships we do look at this in in great depth as the whole approach used in this research compares the similarity of one period of time and deduces the confidence that they were both created by the same user. We present in section 5.2 an investigation of this, and we note the difficulties, and our results in section 5.6 show that within communities this is potentially a valid technique but becomes increasingly unreliable when it is a community-to-community analysis.

The techniques presented in this research do not however improve the ability for an investigator to perform traditional computer-based crime analysis such as keyword search or artefact identification, rather the approach is to place this into context with other artefacts such that they can be interpreted and presented to court.

3.9.2 THE DATASETS USED IN THIS RESEARCH

In figure 3.20, we can see that the six largest sets of data were combined together to form three large datasets: W, Y and Z. The S2 dataset, which was the most substantial of the 'minor' sets of data, was selected to be combined with D1 so we could test the reality of a second but infrequent user combined with a major set of Internet history. The major and minor set created the X dataset. The column 'Final Number of Sessions' is smaller than the sum of the Sessions in figure 3.19, as the sessions comprising only of non-repeating components are removed (see the reasoning for this in the *c-val* section, 3.10.3). For example, R1 + M2 = 1129 sessions, whereas the Z dataset is 720 sessions. This is a significant and interesting difference, but profiling the abnormal or non-repetitive behaviour is outside the scope of this thesis.



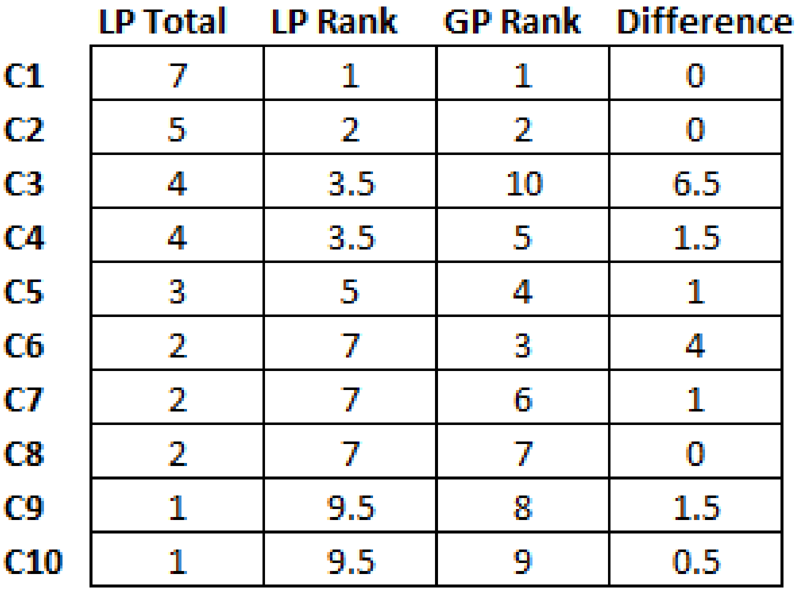
Dataset	Set 1	Set 2	Final Number of Sessions	Distinct Components	Overlap of Components	Comments
W	D1	S1	1532	566	12.90%	A substantial number of sessions, split approximately 2/3 to one major user and 1/3 to a second major user.
X	D1	S2	1156	263	10.65%	The D1 set is the majority user, the S2 set the minority user. This dataset simulates a scenario where there is an occasional second user on a computer. Although there is overlap in this dataset the D, E, S-only and L-only groups clearly segregate the two different users.
Y	A1	M1	784	452	32.30%	A moderate number of sessions, split approximately 2/3 to one major user and 1/3 to a second major user.
Z	R1	M2	991	720	31.53%	A moderate number of sessions, split approximately evenly between two major users.

Figure 3.20: The Experimental Datasets and details about the quantity of Internet History artefacts

3.9.3 STATISTICAL CORRELATION WITH THE DATASETS

We can see in Grajeda et al. [39] that there are presently no standard sets of Internet history that can be used to act as a model or reference point to assess analytical techniques against, so therefore we must address the question that any Internet history used for this kind of system model is representative of a normal web browsing history. The approach presented in section 4.4 uses the rank order of the Global Popularity of websites as a reference point and compares this to the rank order of the frequency of visit/popularity of the websites on the local machine.

We can see in figure 3.21 that we can calculate the popularity rank order of the components C1 to C10 and as long as we have the rank order of the Global Popularity of those sites we can determine the difference between the Local Popularity rank and the Global Popularity rank. We see in figure 3.21 that the Global Popularity of C3 was ranked the 10th, the least popular of the components, and we could call it a 'niche' interest website, whereas we can see that on the system this history came from that it was the 3.5th most popular website visited (shared in popularity with component C4). Therefore, component C3 stands out as having the highest difference at 6.5.

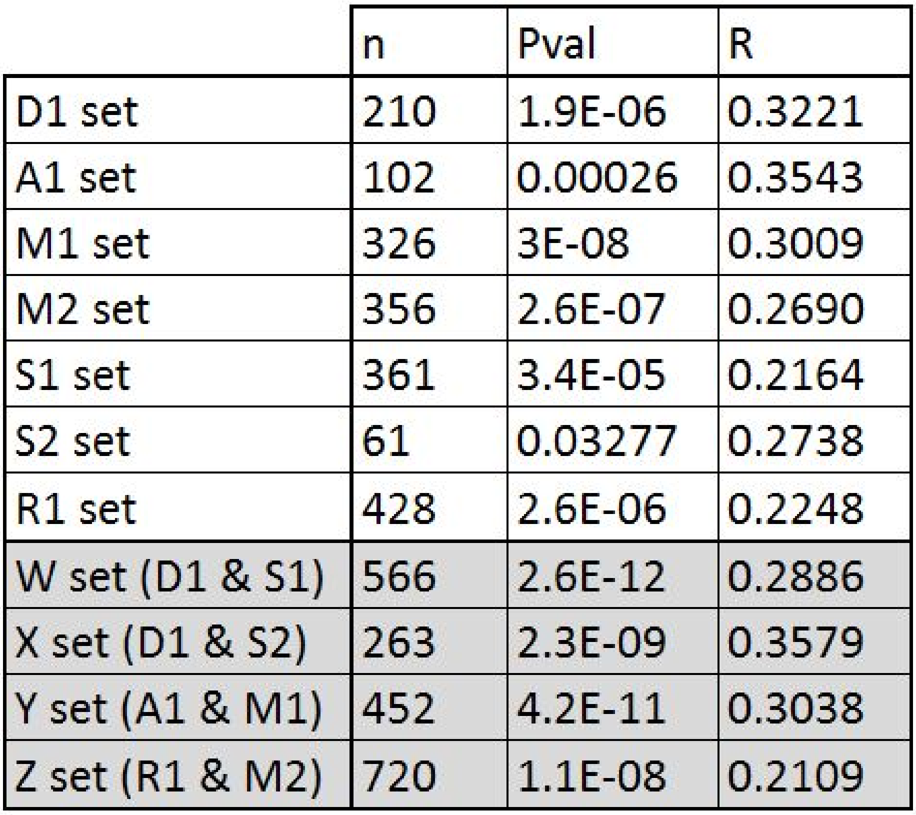


	LP Total	LP Rank	GP Rank	Difference
C1	7	1	1	0
C2	5	2	2	0
C3	4	3.5	10	6.5
C4	4	3.5	5	1.5
C5	3	5	4	1
C6	2	7	3	4
C7	2	7	6	1
C8	2	7	7	0
C9	1	9.5	8	1.5
C10	1	9.5	9	0.5

Figure 3.21: An example of the Local Popularity (LP) of figure 3.5 with the Rank order for the LP and Difference in rank order calculated from an example Global Popularity (GP) Rank.

This rank order data lends itself to a Spearman Correlation between the Local Popularity rank order and the Global Popularity rank order. Performing a Spearman correlation between the GP and LP ranks in figure 3.21 we can see that the correlation is 0.6 with a pvalue of 0.067, which would be close to significance, but with only 10 components we would likely not consider this sufficiently significant.

If the high difference C3 component was removed, the Spearman correlation would change to 0.87 with a pvalue of 0.0034, clearly demonstrating that the user's behaviour does approach the norm and the data become statistically significant when the niche interest websites are removed. We have performed a Spearman Correlation analysis of the individual datasets used in this research, and when they are combined together to form our two-user datasets, the results of which can be seen in figure 3.22.



	n	Pval	R
D1 set	210	1.9E-06	0.3221
A1 set	102	0.00026	0.3543
M1 set	326	3E-08	0.3009
M2 set	356	2.6E-07	0.2690
S1 set	361	3.4E-05	0.2164
S2 set	61	0.03277	0.2738
R1 set	428	2.6E-06	0.2248
W set (D1 & S1)	566	2.6E-12	0.2886
X set (D1 & S2)	263	2.3E-09	0.3579
Y set (A1 & M1)	452	4.2E-11	0.3038
Z set (R1 & M2)	720	1.1E-08	0.2109

Figure 3.22: Showing the results of the Spearman Rho Correlation (R) of the Experimental Datasets correlated with the Global Popularity ranking taken from the Alexa Internet service as described in section 4.4

We can see in figure 3.22 that due to the large number of non-repeating individual websites/components that appear in our datasets (the ‘n’ value) the correlations of 0.21 to 0.35 (the ‘R’ value) show statistical significance at these P-value levels.

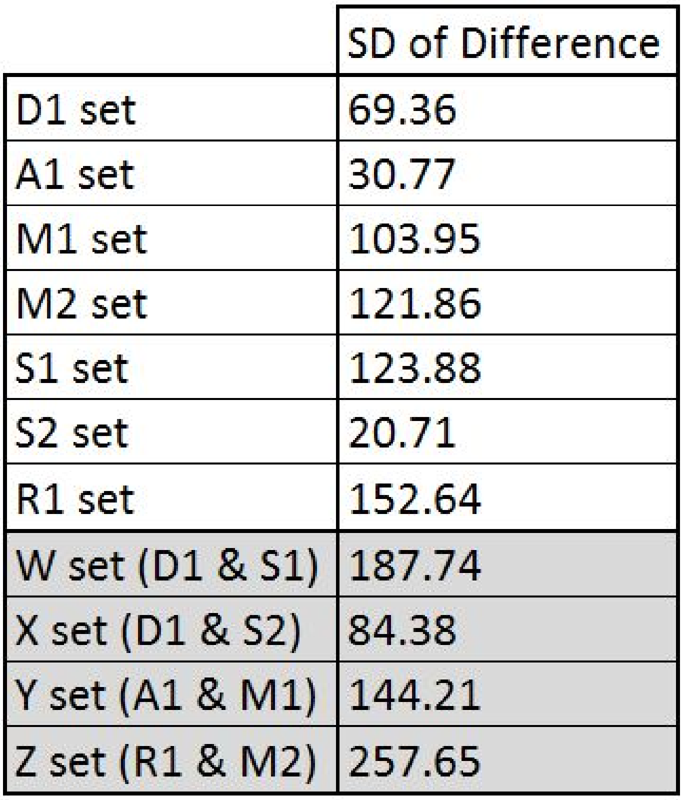
N.B. the S2 dataset individually would probably be rejected as insufficiently significant in our experiments, but how we use it in the X dataset is specifically as a minority user of the device, which we could expect to not have a comprehensive range of Internet access. Future work of our research will focus on the normality and significance of the low difference conditions of the history, such as highlighted in section 4.4.3 of this thesis.

We therefore conclude from this analysis that each of these possible datasets (with the possible exception of S2 taken individually) is representative of a 'normal' set of Internet history. When combining the individual datasets into known test datasets *W*, *X*, *Y* and *Z*, we see that these datasets although artificially combined, do retain statistical significance and we conclude therefore are valid for our purpose of modelling a body of Internet history sessions.

3.9.4 STANDARD DEVIATION AND 'NORMALITY' OF THE DATASETS

It is possible to plot for each website/component the difference of the Local Popularity (LP) rank and the Global Popularity (GP) rank. These plots will be either zero, where there is no difference between the rank order of the GP and the LP, or will be positive or negative of zero. The graphs shown in figures 4.13 to 4.20 are histograms showing the number of websites and their position relative to the zero difference at the middle of the graph.

It is not to suggest that the data inherently has a normal distribution form, rather these plots have discovered that there is normality to the distribution of the behaviour of the users. The full implications of this distribution are not presently understood, but does provide a practically useful model for segregating the components based upon the difference in rank order by using the Standard Deviation (SD) from the zero difference.



	SD of Difference
D1 set	69.36
A1 set	30.77
M1 set	103.95
M2 set	121.86
S1 set	123.88
S2 set	20.71
R1 set	152.64
W set (D1 & S1)	187.74
X set (D1 & S2)	84.38
Y set (A1 & M1)	144.21
Z set (R1 & M2)	257.65

Figure 3.23: Showing the Standard Deviation of the difference between the Global Popularity and Local Popularity rankings across all of the components in the datasets

We can see in figure 3.23 that the D1 and the S2 datasets individually and when combined into the X dataset have relatively small SDs when compared with the individual R1 and M2 sets, and when combined to form the Z dataset. The Z dataset has not only three times as many components as the X dataset, but it also has a much wider range of differences. We can therefore see in section 4.4.4 that we can segregate the data based upon the individual fitting of the curve to the data regardless of how broad or narrow that data profile is, rather than using either a hard, pre-defined threshold (e.g. +/- 200 difference) or an experimentally derived proportion of the data (e.g. x% difference). Further work in this area may address alternative schemes for effectively partitioning the data and we see in section 4.3 that the partitioning scheme is temporally based, rather than the difference in rank ordering.

3.9.5 DATASET DISSEMINATION ISSUES

Grajeda et al. [39] acknowledge reasons why researchers within Digital Forensics are unwilling to share datasets, particularly with respect to those generated from real world sources as there are substantial ethical and legal concerns about releasing data that contains personally identifiable information. The research presented in this project uses two methods of grouping, one of which is temporally based and would be possible to anonymise, whereas the second technique is not able to be realistically anonymised. It is possible to anonymise a dataset by applying a one-way hashing algorithm to the data, so it can be disseminated to other researchers. For example, the host address 'google.com' becomes '1D5920F4B44B27A802BD77C4F0536F5A' when using the MD5 hashing algorithm, and it is not possible to reverse from that hash back to 'google.com'.

It is possible for other researchers to however partially (or even wholly) reconstruct the dataset by using a 'rainbow table'-like approach, where a list of the most likely/probable hashes have already been computed and can be compared against the 'anonymous' dataset. The researcher would search for '1D5920F4B44B27A802BD77C4F0536F5A' and when it appears (it is likely to, as it is the most visited website in the world at the time of writing) replace that hash with 'google.com'. If the dataset contains only popular websites that could be pre-computed the dataset could be wholly reconstructed if a 'salt' value is not used (and if a 'salt' value is used to prevent someone pre-computing a hash table, then like-for-like comparisons cannot be performed between datasets).

There are two approaches we use in this thesis: the short/long approach lends itself to anonymous datasets as there is no additional contextual information other than the time intervals between the artefacts, whereas the Relative Popularity method requires contextual information about the significance or ranking of the websites. If a Researcher had the hash 'E4D965FCC60DD83C7FF8BA0CBC198EC1' and the

Global Popularity ranking of 45,496, even though this is a somewhat niche website that might not appear in a pre-computed rainbow table of hashes, it should be simple to cross reference and compute this hash belongs to the website host 'gre.ac.uk'.

Because of the data protection privacy concerns, we have therefore not provided our datasets either in plain-text or 'anonymised' format as the relative popularity method would allow the dataset to potentially be reverse engineered.

3.10 ACCURACY OF THE SESSION-TO-SESSION COMPARISON METHOD

In the previous section, we looked at Temporal Interval Threshold and the Session Aggregation method, which we noted across a range of data and temporal intervals. One of the criteria used was the Session-to-Session *t-val*, which we define in this section, and highlight two other variables, which we call the *c-val* and *s-val*.

3.10.1 THE *T-VAL*

The Jaccard similarity coefficient is the similarity between two session aggregates, whereas the *t-val* represents a minimum value that the Jaccard coefficient be above to be considered significant. In figure 3.24 we can see a plot of all the Session-to-Session similarity coefficient values. For example, if there were 10 sessions then we would see 100 plots on the graph (although removing 0.0 values is worthwhile). In figure 3.24 we see a large dataset where the overwhelming number of the session-to session matches are below 0.5, and consequently with the ground truth we can plot the number of correct and incorrect Session-to-Session matches.

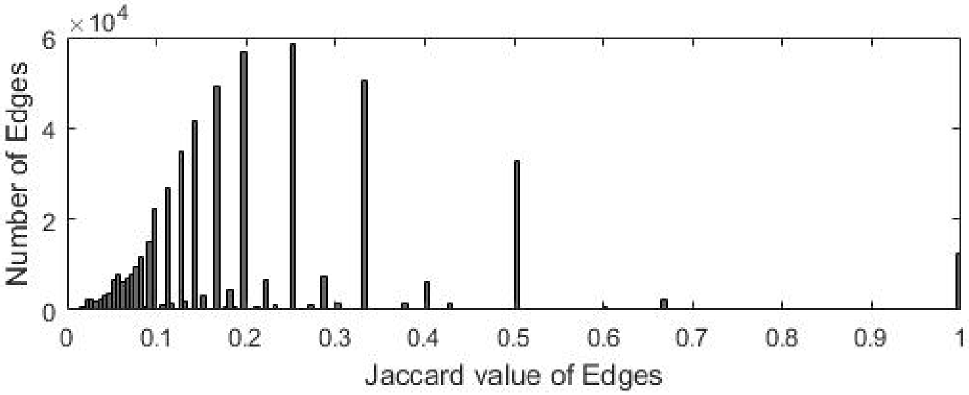


Figure 3.24: Graph showing the number of non-zero session-to-sessions comparisons (Y Axis) and the Jaccard Similarity value (X Axis) of those comparisons in the W dataset.

In figure 3.25, we see that the same data from figure 3.24 has been split to show all the correct user comparisons are represented in green and the incorrect comparisons are in red. This is important because it shows that after all the 0-value comparisons are removed, there are clearly a majority of correct user comparisons remaining, even at the low end of the plot. This is the support for our assertion that the like-for-like sessions do show the correct user the majority of the time. We can, however, see also in figure 3.25 that even high similarity comparisons are incorrect with a notable proportion of red comparisons at 0.5 and even 1.0 Jaccard Similarity.

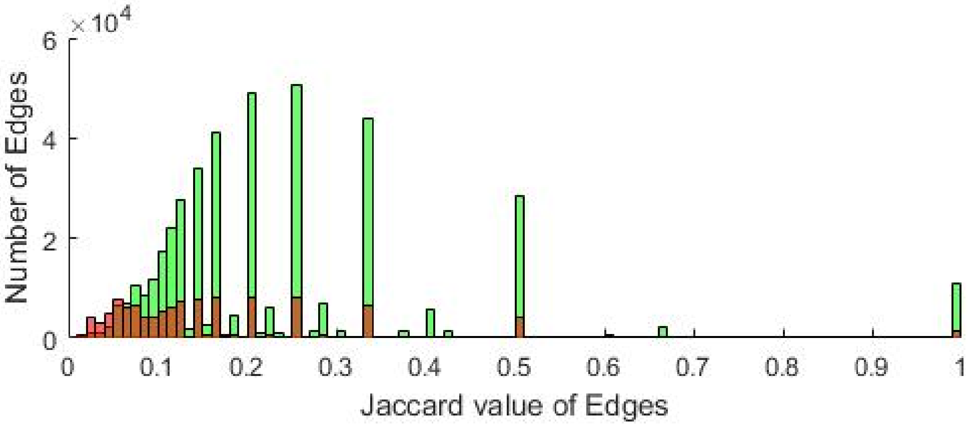


Figure 3.25: Graph showing correct user comparisons (Green) and incorrect comparisons (Red) from the W dataset, 900 s Variable-Length

Figure 3.26 shows a similar set of data to figure 3.25. The D1 data is the same, whereas the S2 data is much fewer in number than the S1 set, consequently there is less opportunity for the two individual dataset to collide, and much less error is seen in the figure.

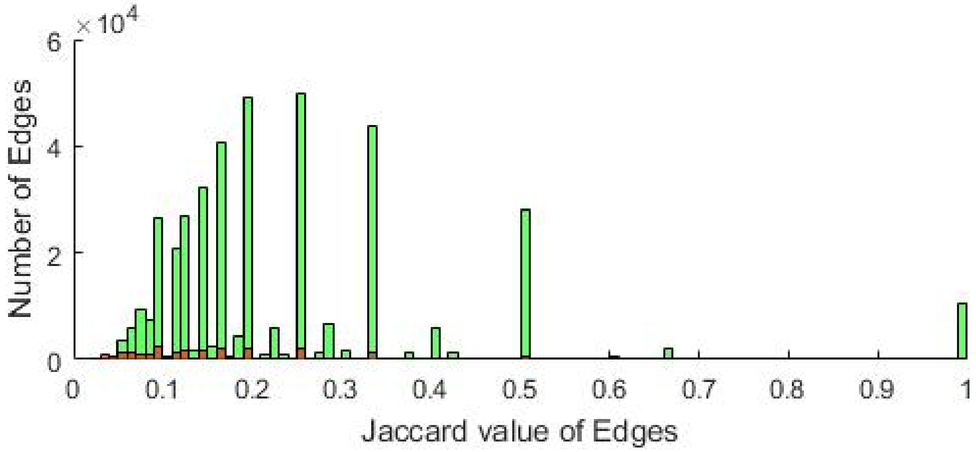


Figure 3.26: Graph showing correct user comparisons (Green) and incorrect comparisons (Red) from the X dataset, 900 s Variable-Length

Figures 3.27 and 3.28 both show the same type of plots with different sets of data. In these cases the high degree of red indicate similarity between the original source data, such that a reliance on the *t-val* alone, would lead to miss classification

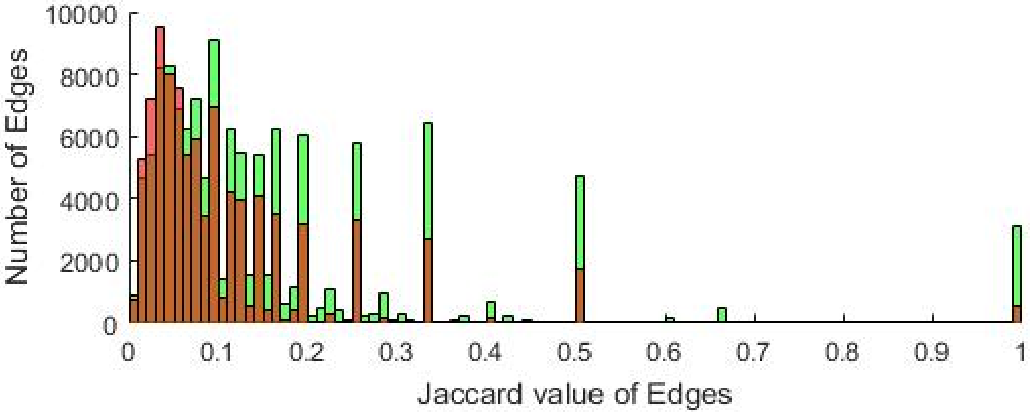


Figure 3.27: Graph showing correct user comparisons (Green) and incorrect comparisons (Red) from the Z dataset, 900 s Variable-Length

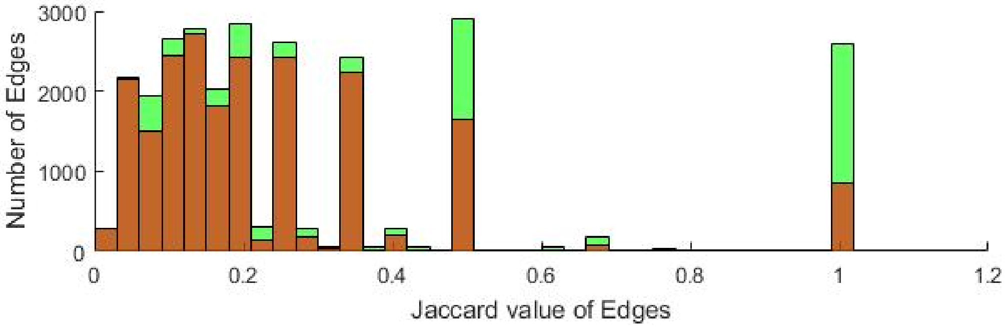
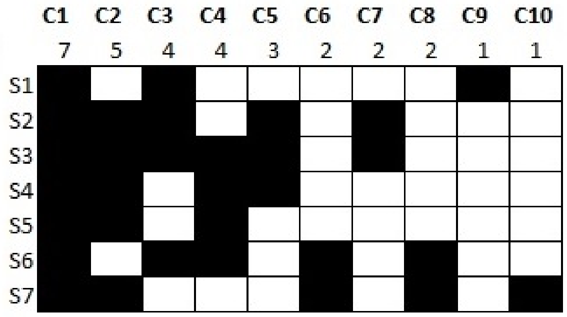


Figure 3.28: Graph showing correct user comparisons (Green) and incorrect comparisons (Red) from the Y dataset, 900 s Variable-Length

The experiments noted in the above section 'Accuracy of the Session Selection Methods' (section 3.8) illustrated in figures 3.11 through 3.18, suggest that relying upon a *t-val* of 0.75 can provide reasonable performance (for particular temporal thresholds) but what we see here in figures 3.25 to 3.28 is that there are potentially significant errors from *t-val* $\geqslant$ 0.50. However it is also notable that in the figures that there are substantial numbers of correctly matching sessions-to-session comparisons below 0.50. If our objective was to utilise as many sessions as possibly, to model as much of the activity on the device as feasible, then we would want to use as many of these matches as possible and therefore a *t-val* itself is not sufficiently precise to achieve this.

3.10.2 THE *S-VAL*

The *s-val* is the minimum number of components that must appear in a session for it to be considered valid.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	7	5	4	4	3	2	2	2	1	1
S1	■	□	■	□	□	□	□	□	■	□
S2	■	■	■	□	■	□	■	□	□	□
S3	■	■	■	■	■	□	■	□	□	□
S4	■	■	□	■	■	□	□	□	□	□
S5	■	■	□	■	□	□	□	□	□	□
S6	■	□	■	■	□	■	□	■	□	□
S7	■	■	□	□	□	■	□	■	□	■

Figure 3.29: Illustration of seven session (S1 to S7) containing ten components (C1 to C10)

Figure 3.29 illustrates Internet history that comprises seven session (S1 to S7) containing ten components (C1 to C10) comprising of Binary Components, which records any one or more visit to that website as a black box. We see by adding up the horizontal rows that the lowest *s-val*s for this set of data is 3, which can be seen in sessions S1 and S5, showing that these sessions contain visits to three websites. We can re-emphasise here that these sessions could be short or long in duration and the components could occur a single time or multiple times, potentially visited and revisited throughout the duration of the session.

3.10.3 THE *C-VAL*

The *c-val* is the minimum number of sessions that a component must appear in to be considered valid. A website that is visited once has a *c-val* of 1 and demonstrates no repetitive behaviour.

For example, consider a case where Session 1 contains three repeating components and five non-repeating, whereas Session 2 contains the same three repeating components and ten different non-repeating components. If the non-repeating components, the *c-val* = 1, were removed, a comparison using the Jaccard similarity measure would show the sessions as exact matches, 1.0, but only as 0.167 if including the non-repeating components.

In figure 3.29, we see that components C9 and C10 have a *c-val* = 1 and they are not repeating components.

3.10.4 THE AMBIGUITY PROBLEM

The presumption for all our Session-to-Session analysis is that if we match a session to all of the other sessions in the dataset, the sessions that are most similar to each other are likely to have been created by the same user. If we then look down the descending list of sessions and their matching values, we should eventually find matches that were created by a different user if there is some shared kind of similarity, but these will be lower than the highest match.

However, in figures 3.25 through to 3.28, we can see that there are sessions that exactly match (values of 1.0 similarity) incorrectly with the wrong user. In addition to wholly incorrectly matching, there can be a situation where there are two or more 'next highest' matches belonging to different users, and consequently we would not be able to identify which one of these belongs to the correct user. We have called this the *ambiguity problem* state.

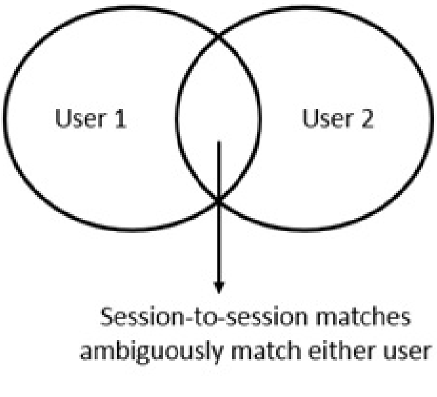
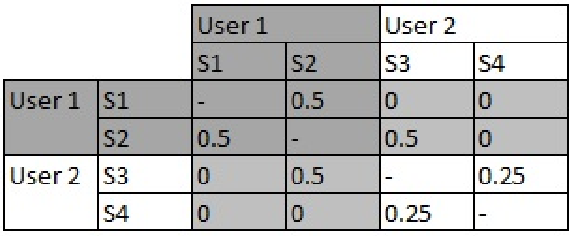


Figure 3.30: Where User 1 and User 2 have sessions that are the same we are unable to determine which user was responsible: the Ambiguity Problem

If we were to cluster sessions together that had either the highest matching value, or were even above a particular *t-val*, it is possibly that these sessions were created by one user, but also match against another user.



		User 1		User 2	
		S1	S2	S3	S4
User 1	S1	-	0.5	0	0
	S2	0.5	-	0.5	0
User 2	S3	0	0.5	-	0.25
	S4	0	0	0.25	-

Figure 3.31: Table illustrating an example of Session-to-Session results for two users, User 1 responsible for S1 and S2, User 2 responsible for S3 and S4

For example, in figure 3.31 there are four sessions (S1 to S4) and the session-to-session comparisons, for two different users. Session 1's highest value is 0.5 (an inexact match as this is not 1.0), which matches with Session 2, the same user and as such this would be a correct match. Session 2 however matches the highest value 0.5 with both Session 1 and Session 3, which would be an ambiguous match because we know that these sessions belong to two different users. Session 3's highest match is with Session 2 rather than Session 4 which belongs to the same user and as such this is an incorrect match. Session 4 is correctly matched with Session 3, but at potentially quite a low matching value of 0.25.

It should be emphasised that 'Ambiguous' matches may be correct, but there is no clear determination between which user is correct. We see for Session 2 that there are two matches, and one of these is indeed a correct match. Consequently, the ambiguous matches are the biggest problem with this approach to session-to-session analysis, more so than the clearly incorrect matches, as invariably users will share some components as they visit the same search engines, social media or news sites etc. We will see in this chapter that there are substantially more ambiguous matches than there are incorrect matches.

3.11 MANIPULATING THE 'DIALS'

We can test our hypothesis that the initial session is sufficiently unique, as the next highest match (an exact or inexact match) should be from the same user of the device, where there are two sets of data which can be compared to each other. This simulates a scenario where there are two users of a single machine that uses a single user profile. The results of these comparisons therefore fall into three categories:

- Correct Match: the highest match belonged to the same user.
- Incorrect match: the highest match belonged to a different user.
- Ambiguous Match: there were matches for both users that had the same match value.

We can therefore measure the correctness of manipulating the *c-val* and *s-val* 'dials' with respect to how many errors, ambiguous sessions matches and overall availability of sessions for analysis.

The graphs presented in this section show the number of Correct highest matches, Incorrect highest matches and Ambiguous highest matches for a number of different conditions of the *c-val* and the *s-val* (which we have for simplicity just called C and S): C=2, S=1; C=3, S=1; C=3, S=2; C=4, S=1; C=4, S=2; C=4, S=3.

These experiments have an effect of showing that as the manipulation is applied by removing the components and sessions, the proportion of sessions which are correct increases, but the overall number of sessions that can be examined are substantially reduced.

3.11.1 THE *W* DATASET

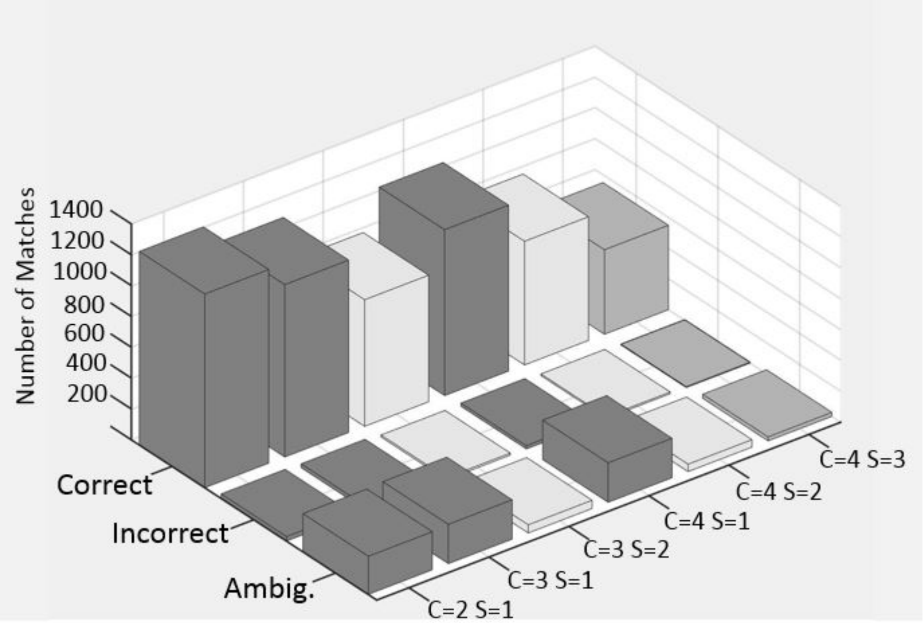


Figure 3.32: *W* dataset showing correct, incorrect and ambiguous matches at various levels of error reduction

In figure 3.32, we see a dataset constructed from two substantial sets of Internet history (explained in more detail in chapter 3.9). The C=2, C=3 and C=4 sets show a broadly similar number of ambiguous matches when the number of session are S=1. However, the number of correct and ambiguous matches significantly falls as the *s-val* is increased, whereas the *c-val* has a much smaller effect.

3.11.2 THE *X* DATASET

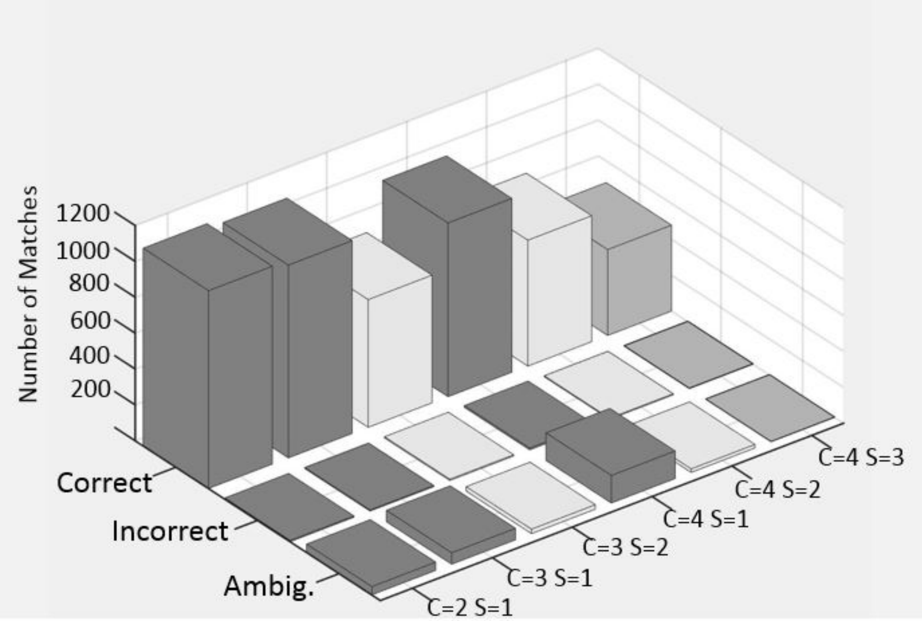


Figure 3.33: *X* dataset showing correct, incorrect and ambiguous matches at various levels of error reduction

In figure 3.33, we see a dataset constructed from a major set of Internet history and a much small second set (explained in more detail in chapter 3.9). These sets of experiments appear to show an unusual increase in the ambiguity problem as the *c-val* is increased without the *s-val* being adjusted. The *X* dataset is similar to the *W* dataset (figure 3.32), and consequently this is not a surprising increase at C=4 S=1, rather the C=2 and C=3 values are relative lower than they were in the *W* dataset.

3.11.3 THE *Z* DATASET

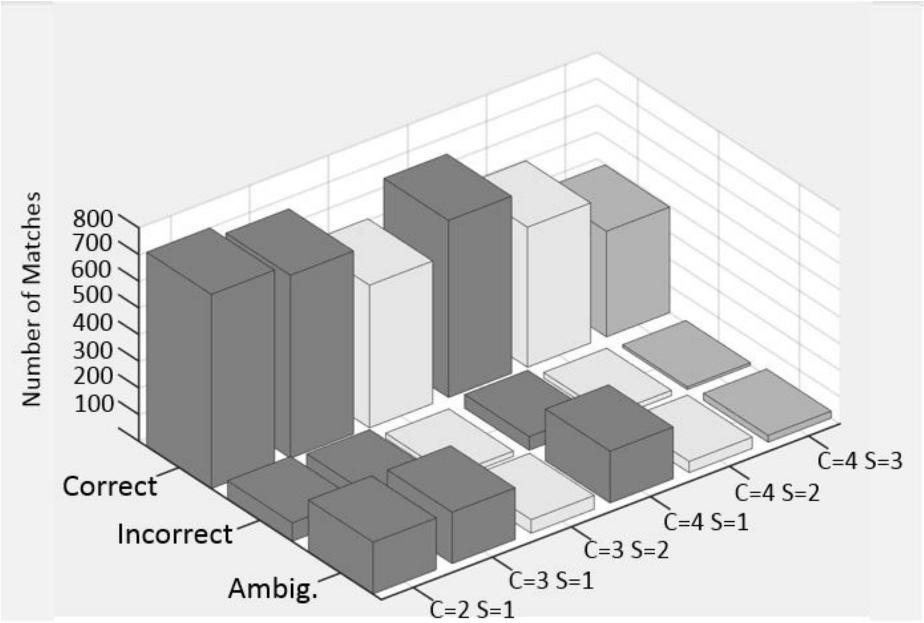


Figure 3.34: *Z* dataset showing correct, incorrect and ambiguous matches at various levels of error reduction

In figure 3.34, we see a dataset constructed from two substantial sets of Internet history. The C=2, C=3 and C=4 sets show a broadly similar number of ambiguous matches when the number of session are S=1, however the number of correct and ambiguous matches significantly falls as the *s-val* is increased, whereas the *c-val* has a much smaller effect.

This dataset had a notable number of incorrect matches, and we therefore see the *c-val* and the *s-val* have an effect reducing these as well as the ambiguous conditions. The implications of this are considered below in section 3.11.5.

3.11.4 THE *Y* DATASET

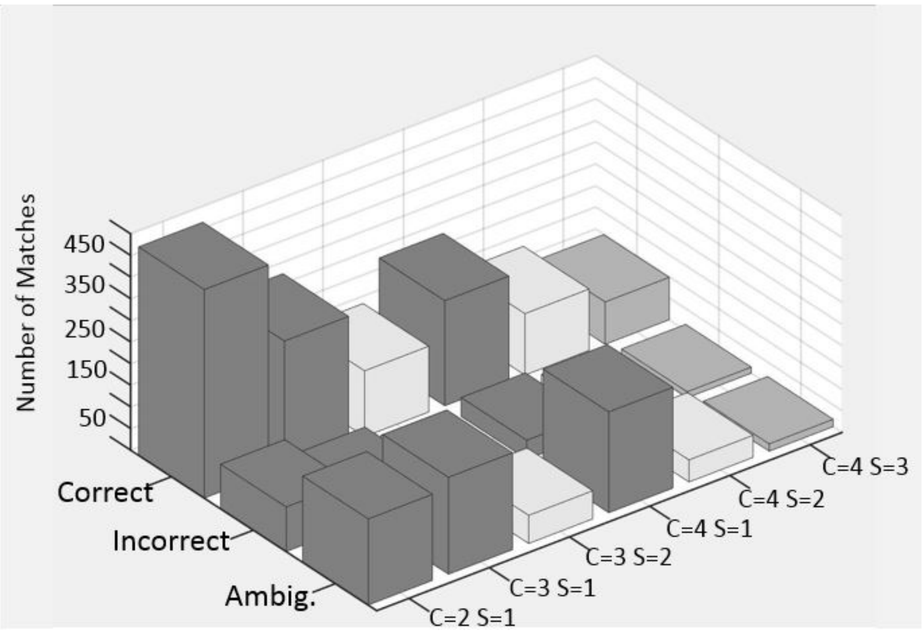


Figure 3.35: *Y* dataset showing correct, incorrect and ambiguous matches at various levels of error reduction

In figure 3.35, we see a dataset constructed from two substantial sets of Internet history. Adjusting the *c-val* with this dataset dramatically effects the availability of the number of the sessions for analysis when looking at the C=2 and C=3 columns, although much less so between C=3 and C=4. We indeed see that although less pronounced than with the *Z* dataset, by increasing the *c-val* the ambiguity problem increases. The implications of this are considered below in section 3.11.5.

3.11.5 THE CORRECT, INCORRECT AND AMBIGUOUS MATCHES

If we consider the percentage of components that overlap in these datasets, which is detailed in figure 3.20, we see that the *W* and *X* datasets have overlaps of 12.9% and 10.65 % respectively. Both of these sets of data have very low amounts of Incorrect matching and few Ambiguous matches, all of which was reduced by raising the *s-val* rather than the *c-val*.

The Y and Z datasets have substantial overlaps of 32.30% and 31.53 % respectively. Both of these datasets have a notable amount of Incorrect matching, and there is also a notable amount of Ambiguous matching, more so in the Y dataset. The amount of Ambiguous matching in the Z dataset is not substantially more than in the W dataset, but there is a difference between the amounts of wholly incorrect matches.

We see in both the Y and Z datasets the incorrect and ambiguous matches are reduced by controlling the *s-val* rather than the *c-val* and interestingly we also see a very pronounced effect reducing the overall correct matches when the *c-val* is adjusted within the Y dataset.

Where there is greater overlap in activity between the users on a device there appears that there is greater possibility of ambiguously mismatching sessions, but also of wholly incorrectly mismatching sessions, both of which conditions can be controlled by increasing the *s-val* rather than the *c-val* - i.e. the users may have shared interests but they are less likely to occur at the same time/within the same session as the other users of the machine.

We can therefore conclude that the manipulation of the *c-val*, beyond the initial filtering of non-repeating components C=1, is considerably less useful at controlling the ambiguous and incorrect matching conditions, and our subsequent research will focus primarily on the *s-val*.

3.12 EVALUATION

In this chapter, we have presented a number of key concepts:

- Fixed-length and Variable-length Session aggregates.
- Temporal Threshold for the session aggregates.
- Session-to-Session comparison approach using Jaccard Similarity coefficient.

- The *t-val* similarity threshold.
- The *c-val* minimum number of sessions that components must appear in.
- The *s-val* minimum number components that must appear in a session.

We have examined the above concepts by performing experiments and producing graphs:

- Plotted the number of sessions that are available depending upon the session aggregate method and the temporal threshold.
- Plotted the percentage of accurate sessions and the number of available sessions in relation to the temporal threshold and the Jaccard similarity *t-val*.
- Plotted all correct and incorrect Session-to-Session matches and the level of similarity they appear at.
- Plotted the number of correct, incorrect and potentially correct or incorrect (ambiguous) highest matches for various levels of the *c-val* and *s-val*.
- Introduced the 4 datasets which we will use throughout this thesis.

We see that by plotting the number of sessions against a hypothetical temporal threshold, we can estimate what the appropriate temporal threshold would be for the session aggregates and this has been shown to be accurate in our experiments where the ground truth was known. We see that the fixed-length session aggregate approach doesn't appear to be as effective on the whole as the variable-length approach when evaluating Internet history. We do consider that the fixed-length approach may work with other types of artefacts, file or operating system records for example, although that has been outside the scope of this work.

By plotting all of the session-to-session matches, as can be seen in figures 3.25 through 3.28, we see that there are many correct, but low similarity value matches plotted, but at the same time there are notable high similarity matches that are incorrect.

The *c-val* variable has some effect during the experiments, but beyond the initial filtering for C=1 conditions, where there is no repetition, it has considerably less effect than the *s-val*, which does have significant effect as it is raised. We noted that there were numerous instances of sessions that contained single components and, although at first consideration that might not be the most useful behaviour to analyse, these sessions could involve repeated access to that same component over an extended period of time. We therefore consider the *s-val* further in the next chapter as a potentially most useful way of filtering sessions if they contain uninteresting or non-indicative components.

Across the range of our experiments, we see that for a high similarity value for correct matches, we have a reduced amount of available sessions to examine. Therefore, the approach presented in chapter 4 is to propose and investigate techniques which identify or exclude as best as possible the components that lead to the ambiguous matching, and as such we can then rely upon lower levels of similarity *t-val* or less manipulation of the *s-val* dial to remove erroneous matching.

Chapter 4

ZERO-KNOWLEDGE INTERNET HISTORY SESSION FEATURE EXTRACTION

"Set patterns, incapable of adaptability, of pliability, only offer a better cage. Truth is outside of all patterns."

Bruce Lee - Tao of Jeet Kune Do, 1975

4.1 INTRODUCTION

The previous chapter described the Session-to-Session analysis approach and introduced the Jaccard similarity measure as it is used to compare two sessions to each other. The hypothesis is that if two sessions are compared and the resulting value is above a particular threshold value, the *t-val*, then they must be sufficiently self-similar to conclude that the sessions belong to the same user.

Ultimately, however, two very different users on some occasions will behave in exactly the same way because there will be some overlap in the websites they visit, and there will be, to some greater or lesser extent, overlap in how they use the Internet that will make the comparisons indistinguishable.

For example, a user that inputs a term into a popular search engine such as Google, looking for some definition or explanation of that term could find that the first-choice result presented by the search engine directs them to a reference site such as Wikipedia. Having sought and found a suitable answer, those two entries might be the total content of that online session. A search engine and well known reference website are not at all unique artefacts that could be profiled to one type of person or user of the Internet, yet this pattern of behaviour could occur as an exact match for any user with any other user.

The existence of sessions that contain only or substantially 'popular' sites or data that is likely to occur amongst all user on a system presents two significant problems:

Firstly, using pre-prepared lists of websites that are interesting or perhaps most significantly, that are known to be uninteresting is not desirable. The terms of this research project as outlined in chapter 1, suggest that using external information is undesirable as this could be said to come from a 'black box' source that is unavailable for inspection. Whilst this term does not preclude all types of external reference about what data may or may not be relevant to the investigation or likely to be the commonly overlapping websites, it does limit the use of certain techniques such as pre-established association rules or trained neural networks as these are built from data outside of the case and are not readily available for inspection.

Secondly and perhaps most importantly, a website being popular does not mean that it is insignificant to our analysis. I could have a preference to one particular search engine, whilst another person could have a preference to another (Google and Bing for example). These are both very popular tools and could be considered unhelpful in determining the uniqueness of the user of any sessions where they appear, if it was not for the fact that one person does have a preference towards a specific search engine and that preference excludes the use of the other search engine. Under that circumstance, the search engine is a unique identifying feature, regardless of the overall popularity of the site. Although the search engine example is perhaps an extreme case of user preference, the use of certain social media sites and news sites which could all be popular within the wider global context could very much distinguish between users within a household. Parents, for example, have access to a much more limited social media footprint compared to teenage children.

In this chapter, we present techniques for extracting from overlapping Internet history data between different users, with zero-knowledge about the nature of the websites, and segregating important elements that are unique to the users.

We present the following techniques in detail:

- Short-session and Long-session partitioning.
- Short-only, Long-only and Both session partitioning.
- Grouping based upon Relative Popularity

We discuss other types of grouping that could be interesting, such as known websites or known types of websites, temporal grouping and even spatial grouping when considering that mobile devices could behave differently depending upon where the device is.

4.2 SHORT SESSIONS AND LONG SESSIONS

As noted in chapter 3, our sessions are modelled on the idea that there is activity, an interval and then more activity. When the interval is sufficiently long and above a threshold value we would consider the new activity to be part of a different session. This model can however lead to very short, bursty periods of different activities, and also long periods of a single activity being accessed and re-accessed.

This kind of data can be grouped based upon the characteristic of the session length as the session can be classified into ‘short’ and ‘long’. What is ‘short’ and what is ‘long’ is a somewhat arbitrary definition and an area of potential further research but we note that the use of ‘short’ and ‘long’ is a simple and natural partitioning of the sessions. If the length of the whole session is equal or shorter than the threshold value that is being used to delimit the sessions, this is considered ‘short’, and ‘long’ if the overall length of the session is longer than the threshold. For example, we have selected a threshold of 900 seconds to delimit our Variable-length sessions: if the overall length of the session is less than or equal to 900 seconds it is short; if the total length is greater than 900 seconds it is long.

A short session need not however be insignificant. A user can access a website that is specific to them frequently throughout the day to look at social media updates, forum posts, sports scores, betting odds or any number of significant websites. Conversely, whilst a long session provides the possibility of more components appearing, it is also possible that a session could contain only overlapping components, such as a video streaming website, that is accessed repeatedly over an extended period.

We show here four datasets, the W, X, Y and Z datasets, the simplified versions of figures 3.32 to 3.35, showing only the *s-val* adjustment and the resulting Correct, Incorrect and Ambiguous matches for those datasets. Next, we show the same data broken down by Correct, Incorrect and Ambiguous matches for all of the Short sessions and Long sessions.

The overall trend that can be observed in these datasets is that the Long sessions shows the most consistency, such that *s-val* manipulation does not reduce the number of correct Long results as much as can be seen with the Short sessions, and the Long sessions have the least Incorrect and Ambiguous matches. We do note that the Long sessions are, at S=1, much fewer in number than the Short sessions however once the S=3 filter is used, the number of correct sessions is similar (generally there are more Long sessions, with the exception of the Z dataset).

4.2.1 THE *W* DATASET SHORT/LONG SESSIONS

We show here the simplified versions of figure 3.32 for the *W* dataset, showing only the *s-val* adjustment and the result when breaking the data into short (S) and long (L) sessions. For clarity, the figures are organised from S=3 to S=1 so the S=3 bars are not obscured.

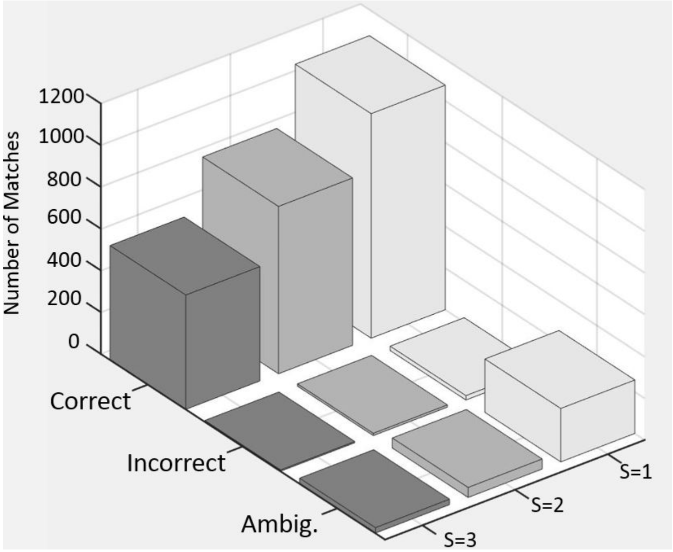


Figure 4.1: *W* Dataset showing Correct, Incorrect and Ambiguous highest matches for the data where the *s-val* has been adjusted

Figure 4.1 shows that adjustment of the *s-val* to S=3 removes the small number of wholly incorrect matches and substantially removes the ambiguous matches, at the cost of significantly reducing the overall number of sessions that can be analysed. This strongly indicates that although the two users within the dataset share some similar websites, they do not typically share two or more of them during the same sessions.

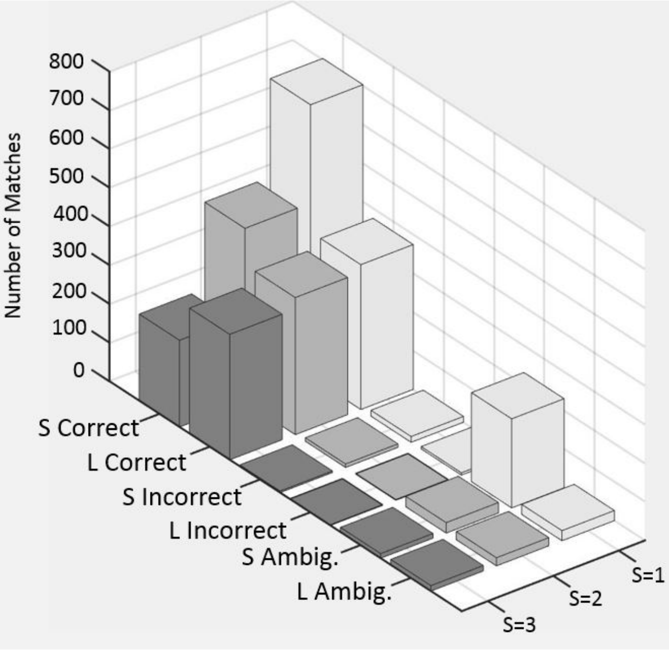


Figure 4.2: *W* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'short' and 'long' session partitioning method, where the *s-val* has been adjusted

Figure 4.2 is similar to figure 4.1 in that it shows that adjustment of the *s-val* removes the small number of wholly incorrect matches and substantially removes the ambiguous matches, at the cost of significantly reducing the overall number of sessions that can be analysed. We also see in this graph that the ambiguity primarily appears in the short sessions and there is little ambiguity or error in the long sessions. This further supports the discovery that it is the short-time accesses where only a single repetitive website that is accessed causes the majority of error or ambiguity.

4.2.2 THE *X* DATASET SHORT/LONG SESSIONS

We show here the simplified versions of figure 3.33 for the *X* dataset, showing only the *s-val* adjustment and the result when breaking the data into short (S) and long (L) sessions.

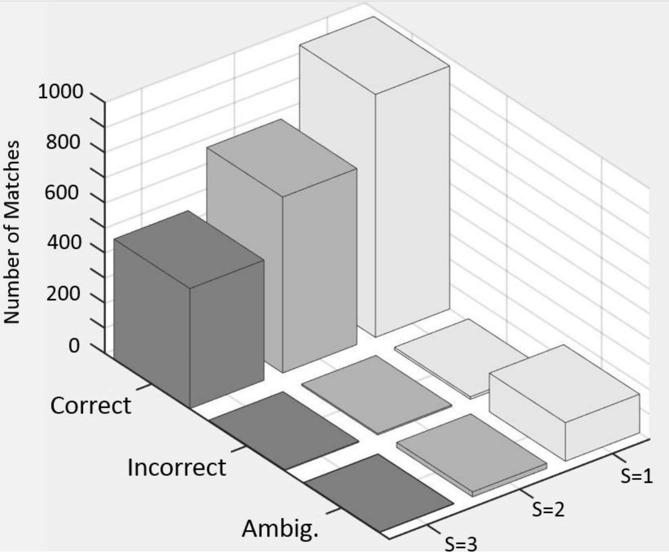


Figure 4.3: *X* Dataset showing Correct, Incorrect and Ambiguous highest matches for the data where the *s-val* has been adjusted

Figures 4.3 and 4.4 are similar to figures 4.1 and 4.2 and show the same findings, that the ambiguity appears in the short sessions and there is little ambiguity or error in the long sessions. This further supports the discovery that it is the short-time accesses where only a single repetitive website that is accessed causes the majority of error or ambiguity.

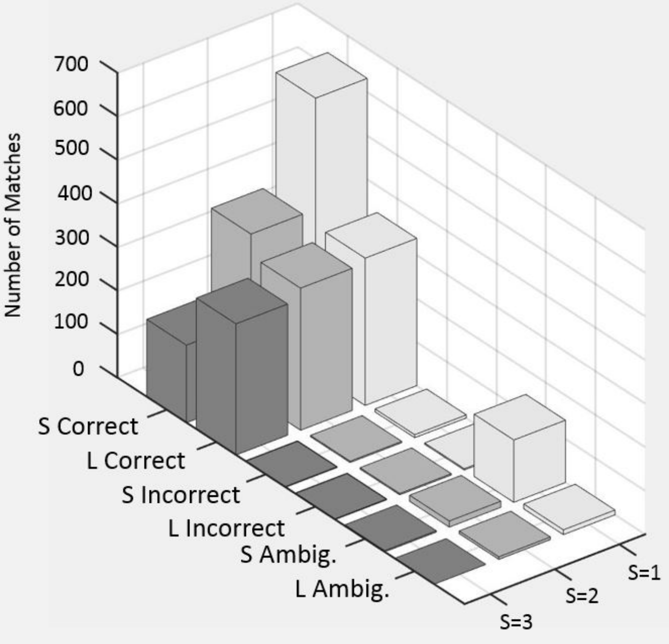


Figure 4.4: *X* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'short' and 'long' session partitioning method, where the *s-val* has been adjusted

4.2.3 THE *Y* DATASET SHORT/LONG SESSIONS

We show here the simplified versions of figure 3.35 for the *Y* dataset, showing only the *s-val* adjustment and the result when breaking the data into short (S) and long (L) sessions. For clarity, the figures are organised from S=3 to S=1 so the S=3 bars are not obscured.

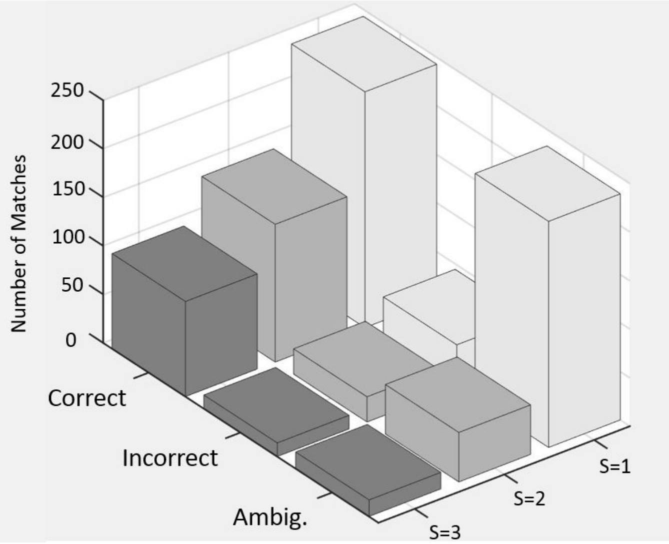


Figure 4.5: *Y* Dataset showing Correct, Incorrect and Ambiguous highest matches for the data where the *s-val* has been adjusted

Figure 4.5 shows a large proportion of ambiguous matches and a notable number of incorrect matches. Figure 4.6 also shows that the majority of the ambiguity appears in the short sessions. This dataset suggests that there is similarity either in the behaviour of the users or there is similarity in the websites associations between those websites as even an increase of the *s-val* to S=3 does not remove ambiguity or error.

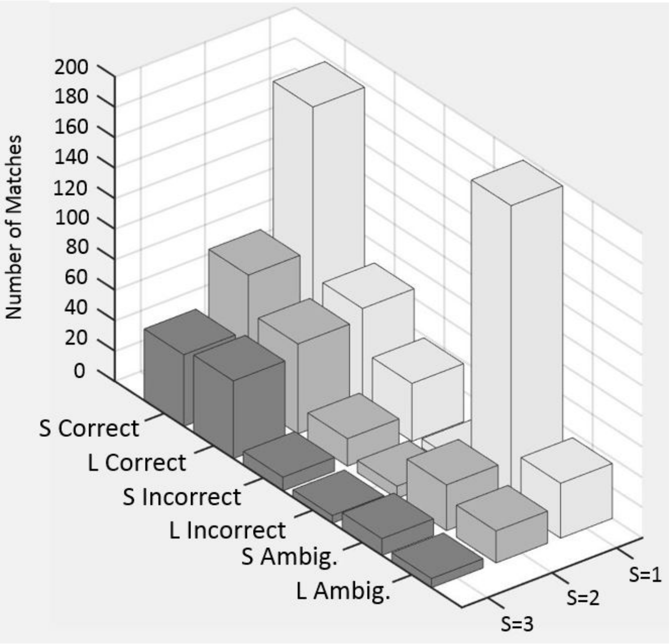


Figure 4.6: *Y* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'short' and 'long' session partitioning method, where the *s-val* has been adjusted

4.2.4 THE *Z* DATASET SHORT/LONG SESSIONS

We show here the simplified versions of figure 3.34 for the *Z* dataset, showing only the *s-val* adjustment and the result when breaking the data into short (S) and long (L) sessions. For clarity, the figures are organised from S=3 to S=1 so the S=3 bars are not obscured.

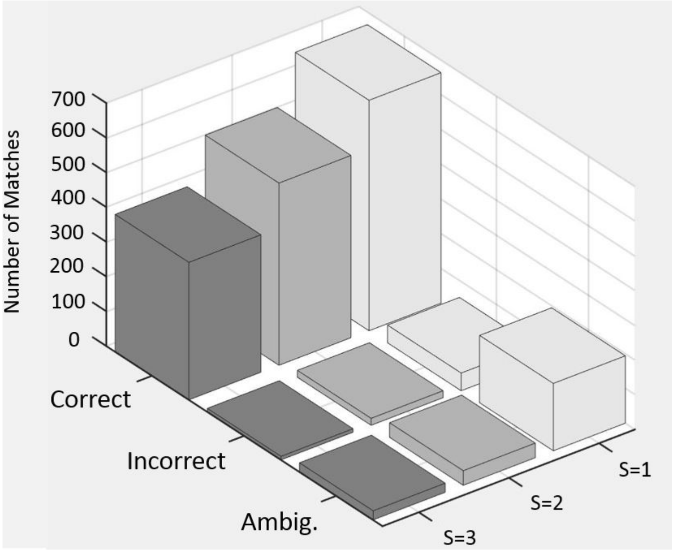


Figure 4.7: *Z* Dataset showing Correct, Incorrect and Ambiguous highest matches for the data where the *s-val* has been adjusted

Figures 4.7 and 4.8 are again similar to figures 4.1 and 4.2, with a similar performance in when the data is broken down into the short and long sessions. The same conclusions that this data strongly indicates that although the two users within the dataset share some similar websites, they do not share them typically two or more times during the same sessions.

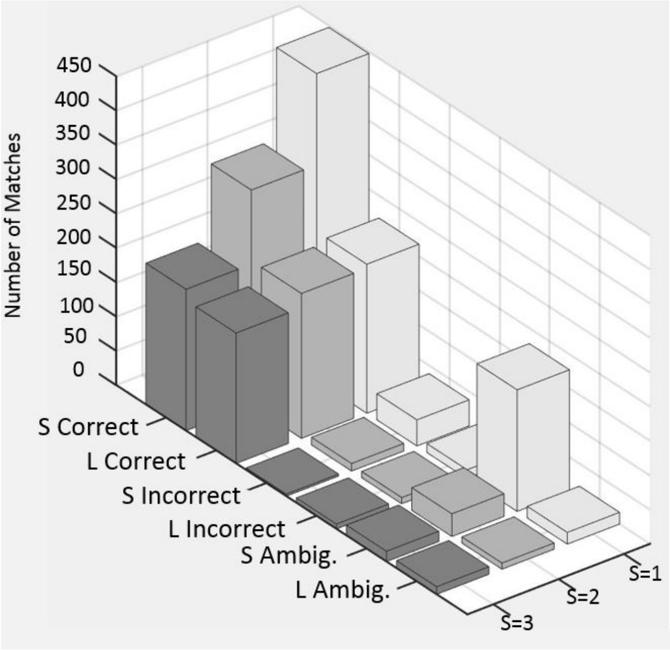


Figure 4.8: *Z* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'short' and 'long' session partitioning method, where the *s-val* has been adjusted

4.2.5 EVALUATION OF THE SHORT/LONG SESSIONS

This approach is simple, computationally cheap, easy to explain and with manipulation of the *s-val* it is very effective at removing incorrect and ambiguous matches. Indeed by selectively using an *s-val* of S=2 or S=3 on the Short sessions and not manipulating the Long sessions at all (I.e. S=1) we can gain the most possible Long sessions, whilst at the same time removing the high probability of ambiguous Short sessions. The selective use of different *s-val*s is explored more in chapter 5.

4.3 SHORT-ONLY AND LONG-ONLY COMPONENTS

There appears to be a value to the Short and Long session partitioning of the datasets, but a further consideration is the behaviour in the Short sessions sufficiently different to the behaviour in the Long sessions, such that we can consider the length of the session a feature?

We can perform a conditional grouping where the components that appear only within the Short sessions are placed in one group, the components that appear only in Long sessions in another group and the else condition contains the components that appear in both categories. These are referred to as the 'Short-only', the 'Long-only' and the 'Both' components.

Figures 4.9 and 4.10 illustrate examples of this for the W and Y datasets and it can be seen that the graphs become quite unwieldy when evaluating the differences between these groups.

Although there is a large number of correct matches with components that appear in both session, the Short-only and Long-only components do infact perform reasonably well. The Short-only components are notably sensitive to the change in *s-val*, but interestingly with the W data sets the Short-only and Long-only components show very low error and ambiguity. This performance is not replicated with the Y set, however the adjustment of the *s-val* such that S=2 does reduce error and ambiguity.

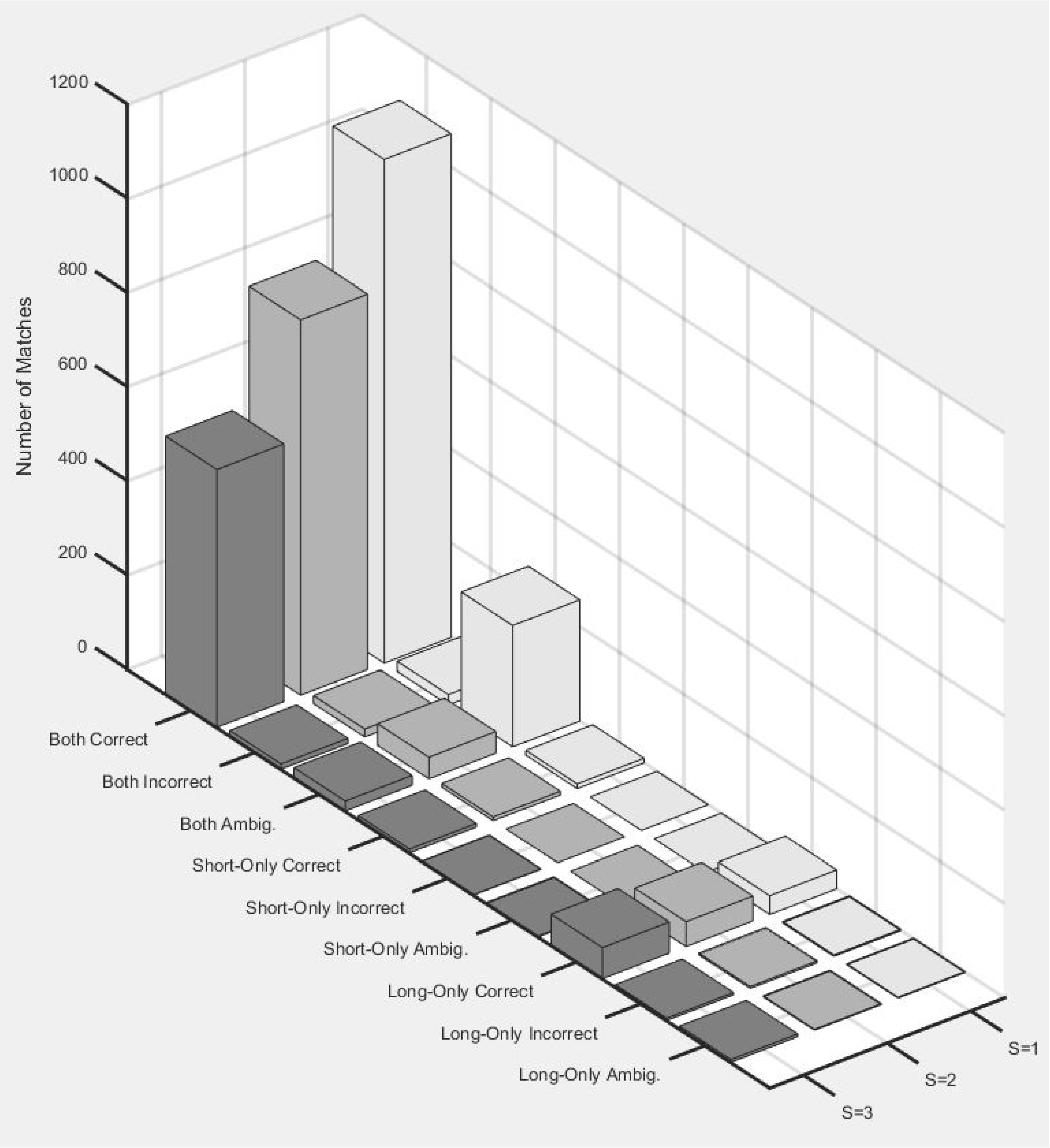


Figure 4.9: *W* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'Short-only', 'Long-only' and 'Both' session partitioning method, where the *s-val* has been adjusted

Ultimately this does suggest that individuals do perform different activities during their Short access times, although it is more reliably detected in their longer access sessions, which does stand to reason as longer periods of time suggest more opportunities for individuality to be expressed.

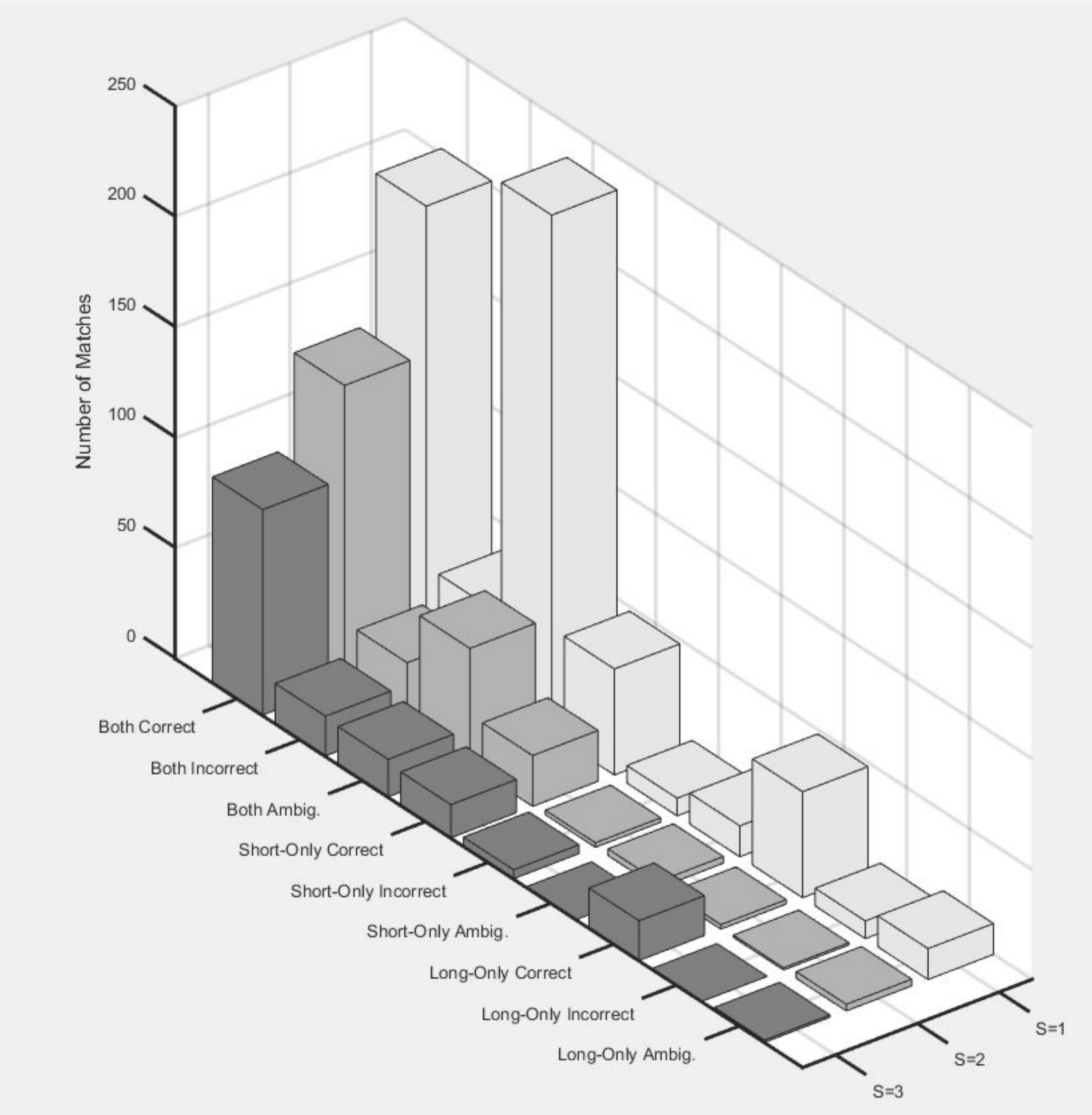


Figure 4.10: *Y* Dataset showing Correct, Incorrect and Ambiguous highest matches for the 'Short-only', 'Long-only' and 'Both' session partitioning method, where the *s-val* has been adjusted

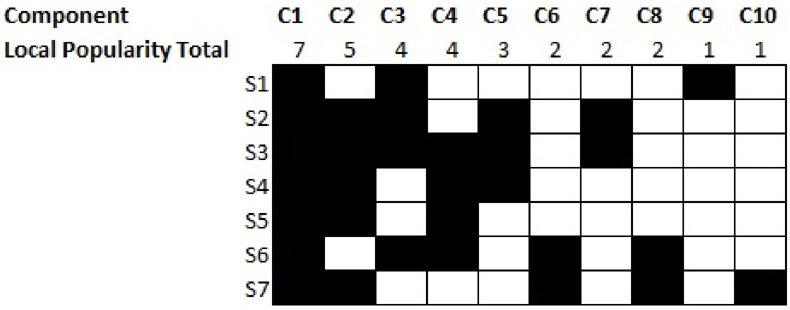
4.4 IDENTIFYING COMPONENTS BASED UPON AN EXTERNAL POPULARITY REFERENCE

Although we noted in chapter 1 that analysis which relies upon external information may become susceptible to an argument that it is invalid because it is a ‘black box’, this is not necessarily the case if the information that is used is freely available for inspection. The information need not be ‘free’ in the sense of it being completely open to the public, but freely available and understandable to other analysts performing examination on the data, which is especially important in an adversarial legal system.

We propose in this section a grouping that uses reference data from an external source, but the information is a simple metric of the popularity of the component. As noted in section 4.1, the popularity of a website itself is not necessarily a useful, or unhelpful, feature and consequently we describe how the relative popularity becomes important.

4.4.1 LOCAL POPULARITY

By breaking up the Internet history records into sessions we can see by totalling the number of sessions that certain website hosts or components are regularly visited. Figure 4.11 shows ten websites (or Components) listed as C1 to C10 and seven periods of time or sessions listed as S1 to S7. If a website appears, even once, during the session then the box is filled in black (a binary component). For each component, the total number of sessions that it appears in is shown at the top of the table as the Local Popularity Total. We see that in this case C1 appears in every session, whereas C9 and C10 appear only in single sessions.



Component	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Local Popularity Total	7	5	4	4	3	2	2	2	1	1
S1	■	□	■	□	□	□	□	□	■	□
S2	■	■	■	□	■	□	■	□	□	□
S3	■	■	■	■	■	□	■	□	□	□
S4	■	■	□	■	■	□	□	□	□	□
S5	■	■	□	■	□	□	□	□	□	□
S6	■	□	■	■	□	■	□	■	□	□
S7	■	■	□	□	□	■	□	■	□	■

Figure 4.11: Table of seven example sessions (S1 to S7) comprising ten components (C1 to C10), with the Component Totals referred to as 'Local Popularity' Totals

The regularity of access to a website may not be the result of a user's behaviour, for example a particular site appears in every session because it is the home page of the browser that appears each time the application is started. The presence of such a pattern is an artefact of the system's behaviour as opposed to a user's behaviour and any kind of automatic session-to-session analysis should be able to disregard such artefacts.

Where there are regular visits to websites, that do not appear because of automated behaviour, there is behavioural significance as a user has through some method of selection visited and revisited the site.

We therefore define:

Local Popularity (LP) the number of sessions that the component appears in on the machine or history that is being analysed. It may be desirable to remove from the analysis components that appear a single time and do not appear in multiple sessions (i.e. *c-val* =1).

We further define:

Local Popularity Ranking (LP Rank) the ranking of the Local Popularity of the components, from 1^{st} rank for the component with the highest LP, to the n^{th} for the components with the lowest LP. It should be noted that many components will have the same value LP and consequently the average of the tied rankings is used in these cases.

4.4.2 GLOBAL POPULARITY

As we noted above, each component has a Local Popularity value representing how frequently that website has been visited over a number of sessions, but this itself gives us no suggestion to the relative value of the component. We therefore define:

> **Global Popularity (GP)** - a measure of popularity for the component that is external to the data gathered during the analysis.

Global Popularity measures provide some level of impact assessment, which could be link-based algorithms that identify how well referenced sites are by other sites, such as PageRank [63], HITS [26], CLEVER [27] or the impact assessment could be based upon the analysis of the volume of web traffic, such as Alexa Internet [68].

As part of our experiments we are using a single GP metric from the Alexa Internet Traffic Rank, which provides a global ranking metric for a substantial quantity of websites. Therefore, we define:

> **Global Popularity Ranking (GP Rank)** The Global Popularity value is listed from 1^{st} rank for the component with the highest GP, to the n^{th} for the components with the lowest GP.

The Alexa ranking used in our experiments was provided in GP rank ordering, with the exception of sites that were particularly niche where no data had been gathered for them. Consequently, the last GP Rank was always a tied average of the niche/unranked websites.

4.4.3 COMPARING LOCAL POPULARITY AGAINST GLOBAL POPULARITY

Without any prior knowledge about the users, their interests or the types of websites they like to visit we can compare the locally popular (LP Rank) components to external global popularity (GP Rank) metrics.

We can compare LP rank to GP rank, which allows us to determine if the result is Low or High difference between the Local and Global popularity and four basic conditions can be inferred from this about the components:

1. **High difference: Low GP, High LP** This is the potentially idiosyncratic web sites 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 particular popular service, such as a user not having a significant social media footprint, but occasionally follows 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 scenario where multiple users have access to the same device, one user may have preference to the use of one social media site 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 condition is infrequent viewing of a fairly niche website. From our experience to date, this tends to make up a sizable bulk of a subject's Internet history.

We can see from the above four conditions that high local popularity is always significant, principally because session-to-session analysis is an analysis of repetitive behaviour, and the more repetition of behaviour the better. Condition 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.

In figure 4.12 we see the four conditions illustrated. Some difference in the rankings is always present and consequently the selection of the X threshold value that determines high difference and low difference conditions is critical.

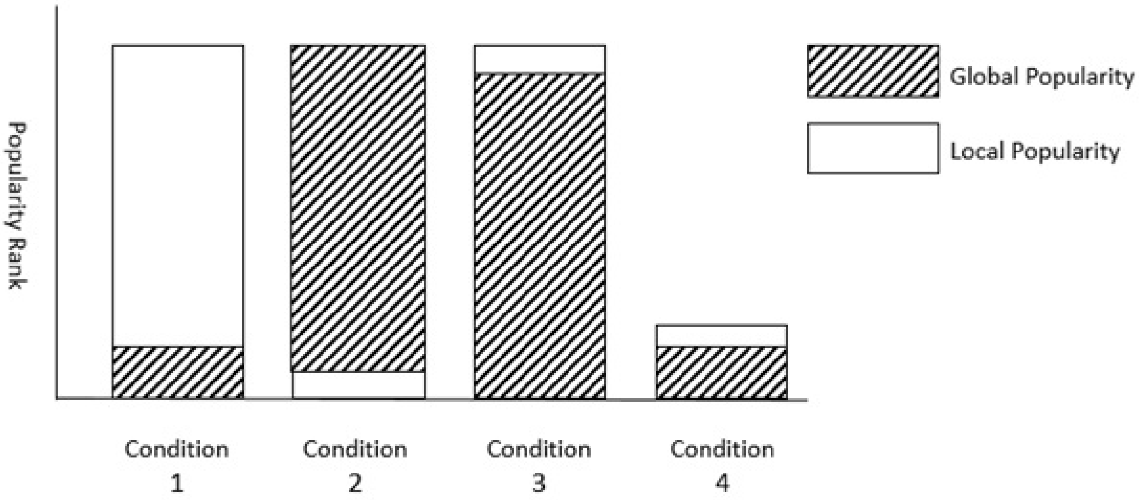


Figure 4.12: Illustration of the possible Relative Difference conditions

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).

Although not ‘idiosyncratic’, condition 3 components can be significant. 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.

We show here in figures 4.13 to 4.20 the graphs of the difference between the LP rank and the GP rank for the components in the datasets. We have shown the individual datasets and the combined two-user datasets. For example, figure 4.13 shows the D1 dataset and the the S1 dataset, whereas figure 4.14 shows the result of combining these two sets as the W dataset.

By combing the individual datasets to form our two-user datasets we increase the number of components which means there are now a larger range range of Local Population components to be rank ordered against a larger range of Global Popularity components to be rank ordered and this has an effect of making the combined plots to be wider. This can be seen in figures 4.14, 4.16, 4.18 and 4.20.

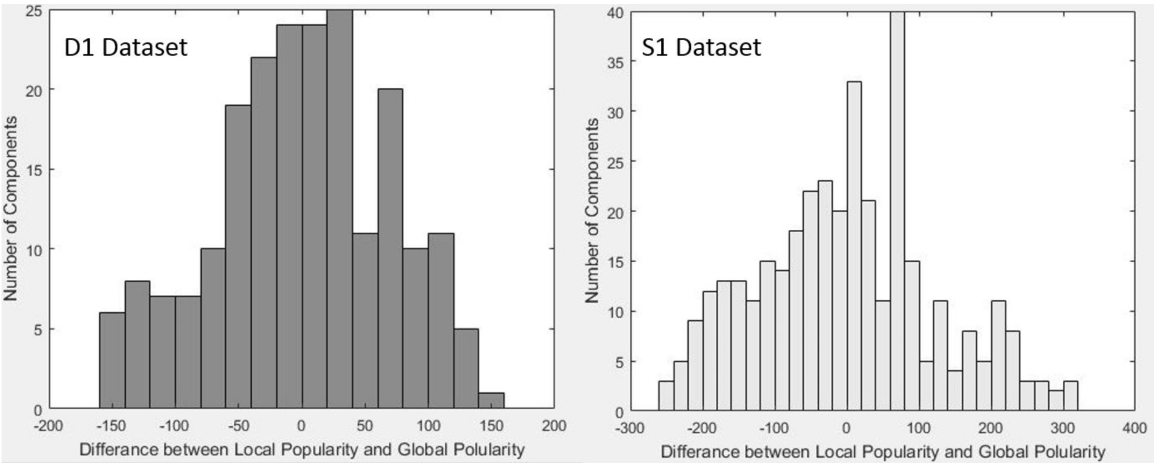


Figure 4.13: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the D1 (left) and S1 (right) datasets

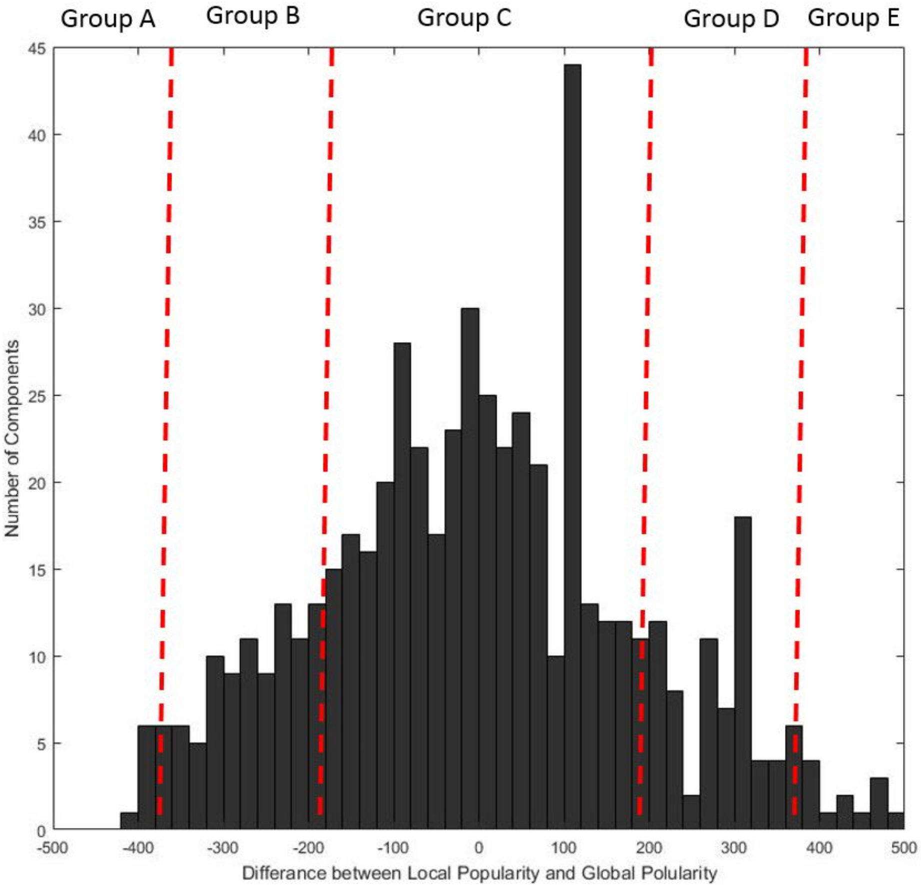


Figure 4.14: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the *W* dataset. We also show here the Relative Popularity grouping scheme

The S2 dataset (shown in figure 4.15) has a much smaller number of sessions, representing a ‘minority user’ and consequently this has a smaller flattening effect in figure 4.16.

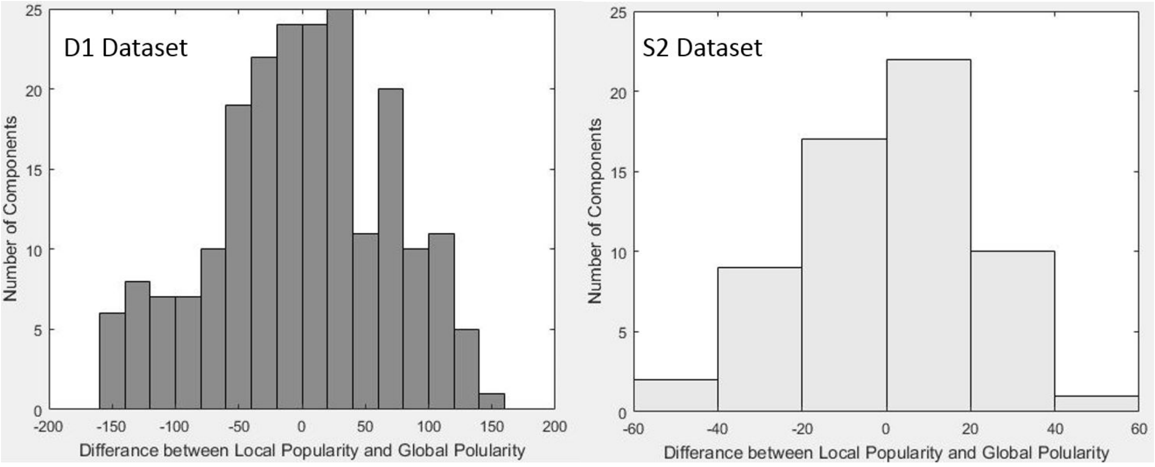


Figure 4.15: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the D1 (left) and S2 (right) datasets

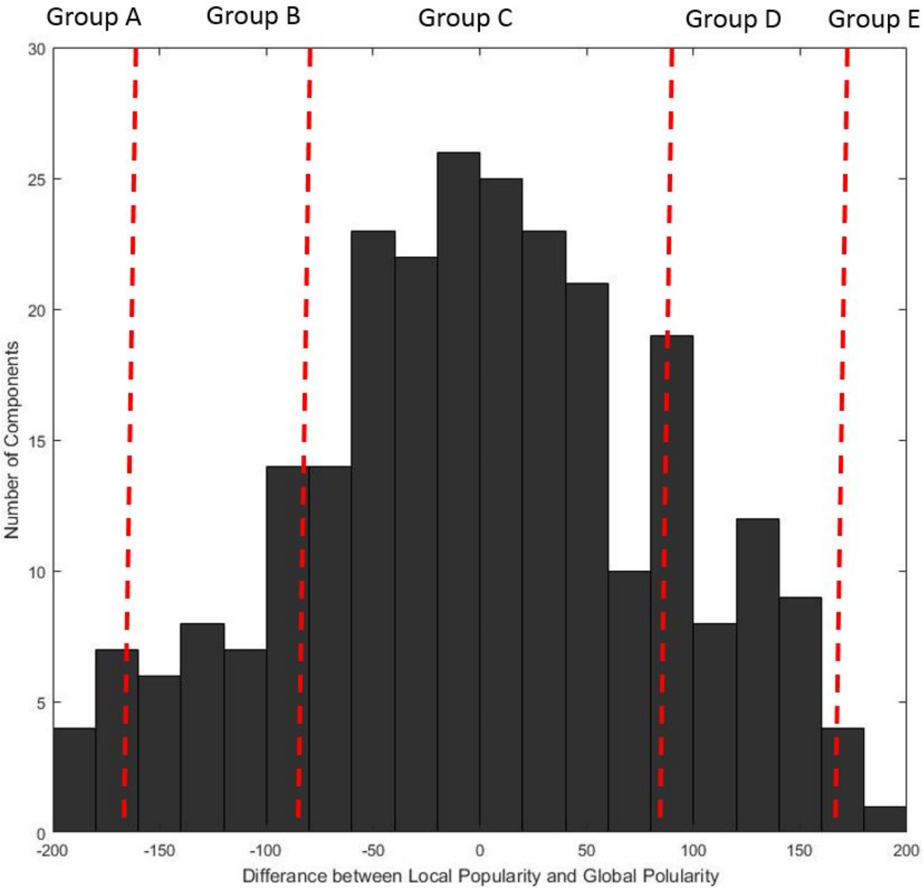


Figure 4.16: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the X dataset. We also show here the Relative Popularity grouping scheme

The Y set (figure 4.18) and Z dataset (figure 4.20), which we can see from the tables in section 3.9 that the A1 and M1 datasets (figure 4.17) shared 32.3% of the components in their history, and the datasets that comprise the Z set (figure 4.19) also overlap by 31.53%. Therefore, as we can see that as the Z dataset is considerably wider and flatter, we can conclude the flattening effect is not accounted for by overlap, rather by the large number of components and how they differ from the Global Popularity zero point.

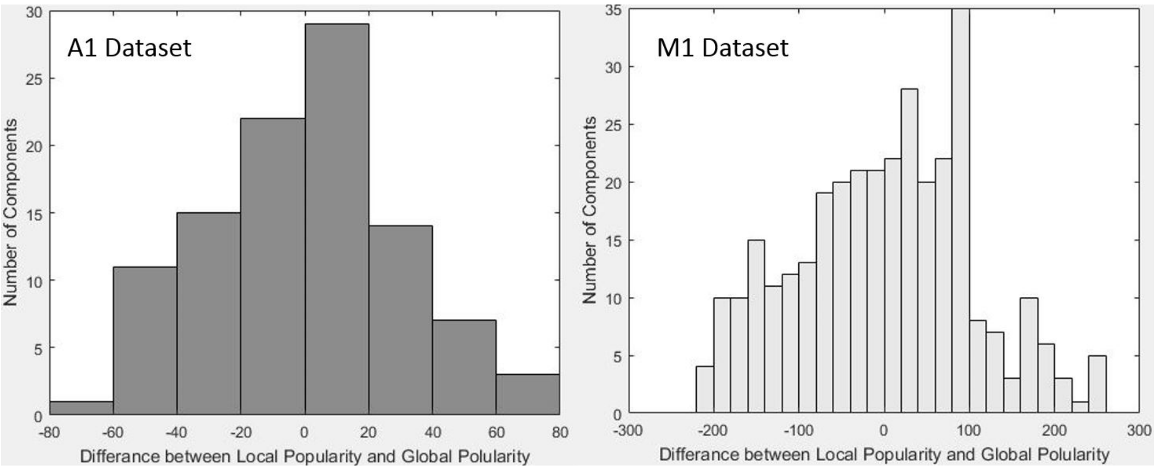


Figure 4.17: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the A1 (left) and M1 (right) datasets

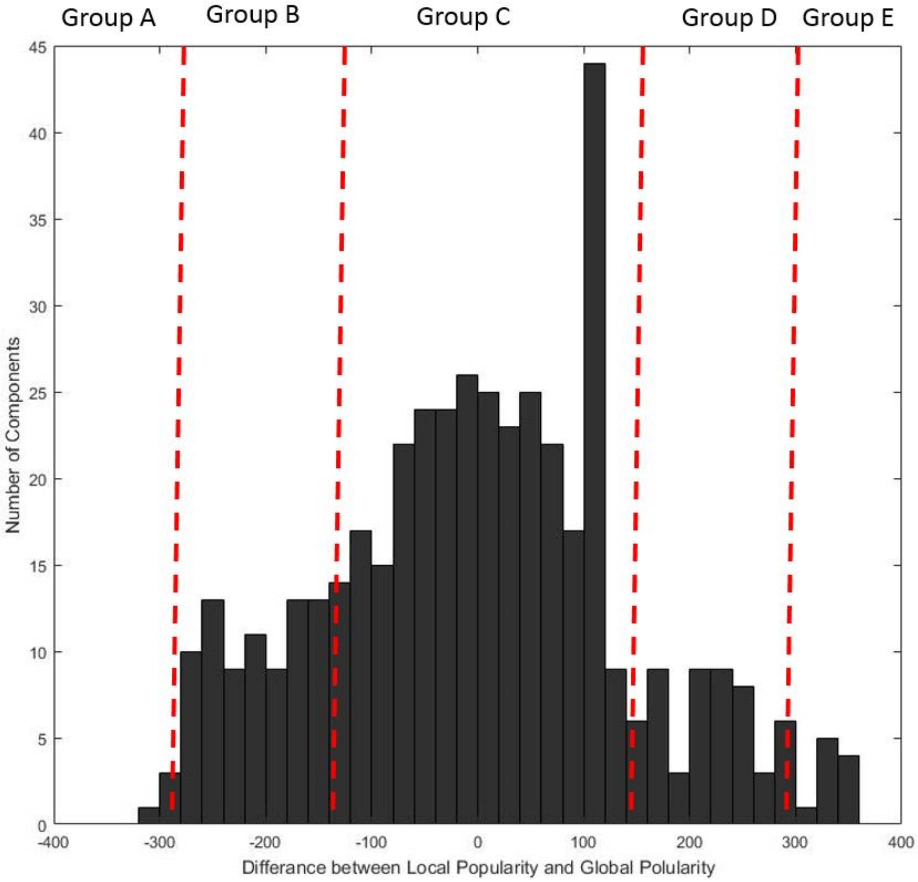


Figure 4.18: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the Y dataset. We also show here the Relative Popularity grouping scheme

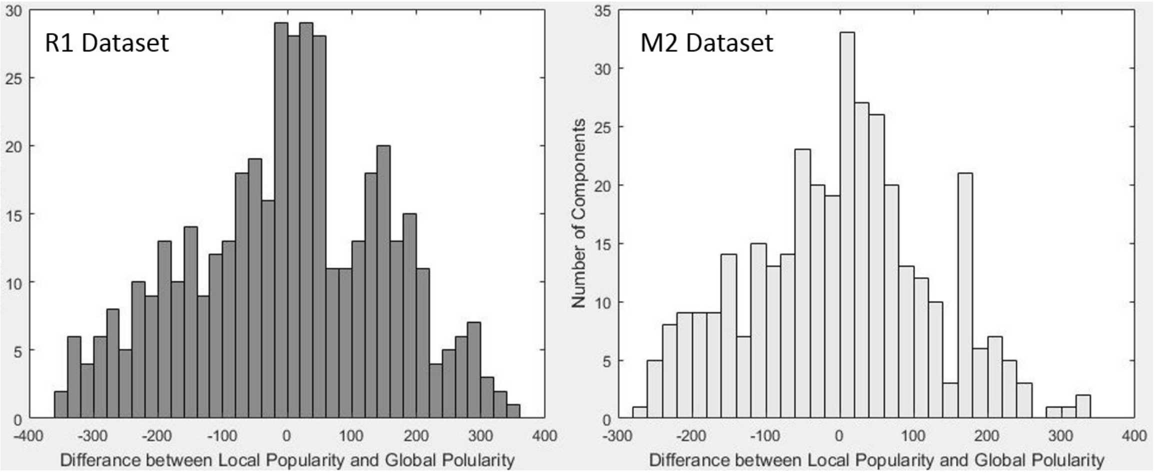


Figure 4.19: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the R1 (left) and M2 (right) datasets

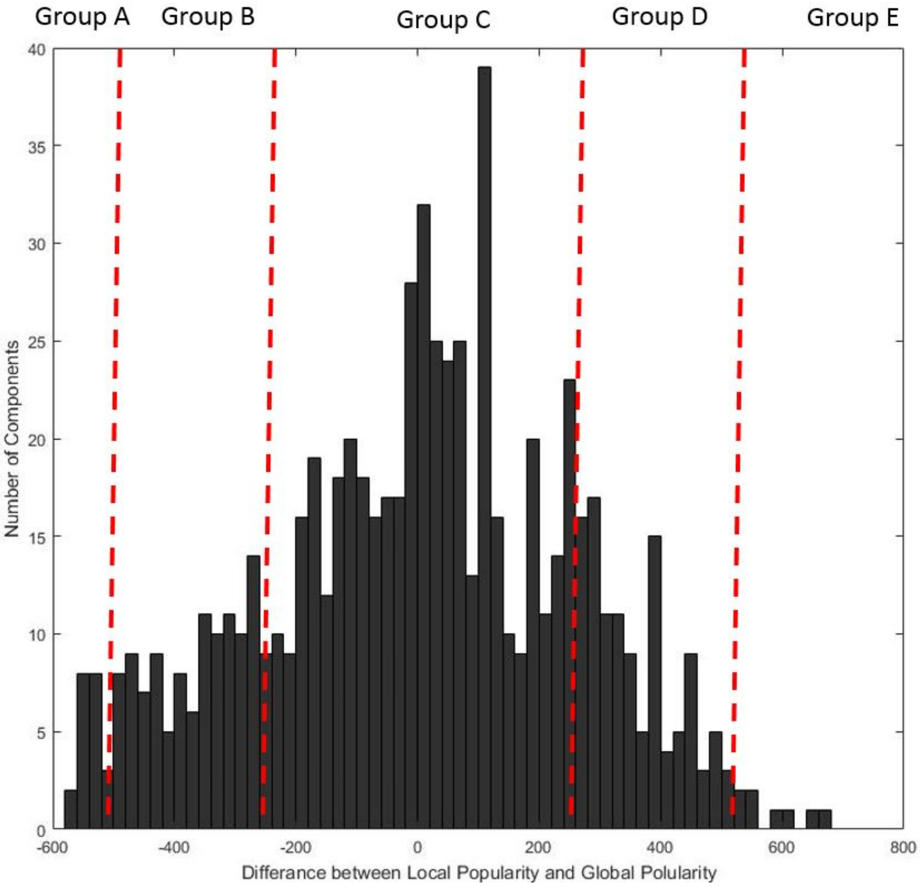


Figure 4.20: The number of components and their Relative Difference from zero for the LP Rank Order and GP Rank Order, plotted for the Z dataset. We also show here the Relative Popularity grouping scheme

4.4.4 GROUPING RELATIVE POPULARITY

We can illustrate relative popularity by plotting the difference between the LP and GP ranks on a graph, such as seen in figure 4.21, and then group based upon some threshold value. We have performed our experiments using standard deviation (which is shown in section 3.9.4) to group the Relative Popularity data. Figure 4.21 shows that 'condition 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.

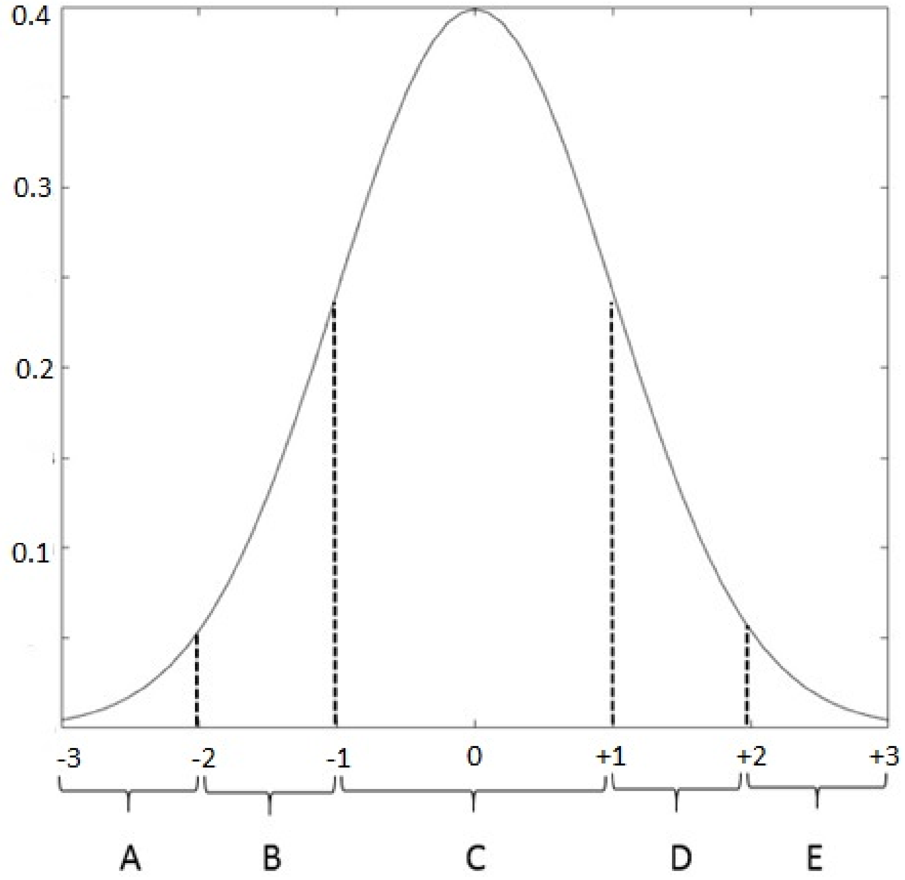
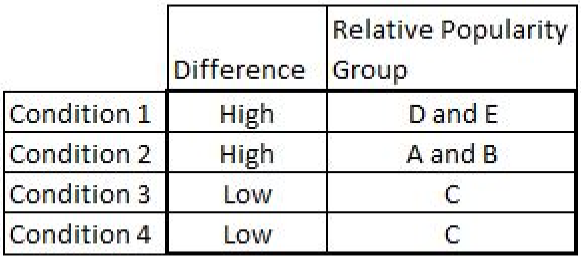


Figure 4.21: An illustration of the standard distribution and how Relative Popularity groups A through E could be plotted on the X axis as a deviation from 0 difference between the Global Popularity Rank and the Local Popularity Rank

A summary of the relative difference conditions and the groups they have been categorised into is shown in figure 4.22.



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

Figure 4.22: Relative Popularity Groups

Figures 4.23 and 4.24 illustrates the distribution curve plotted over the four sets of data that show the number of artefacts (Y Axis) that have a relative difference between the LP and the GP (X axis).

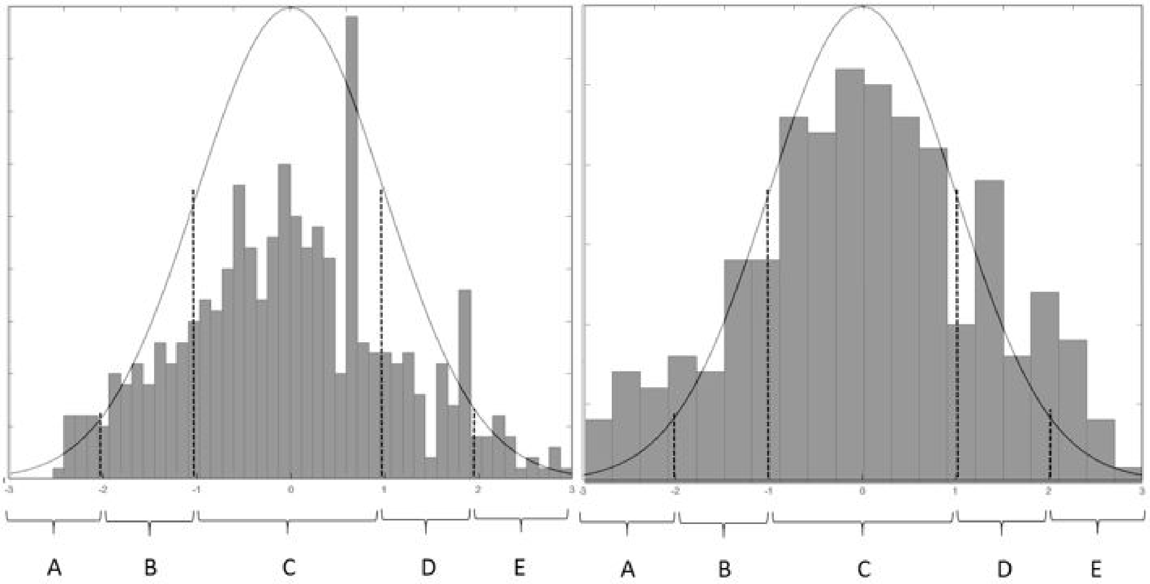


Figure 4.23: The *W* Dataset (left) and the *X* Dataset (right) both plotted against the Standard Distribution curve to illustrate the relative similarity with a normal curve

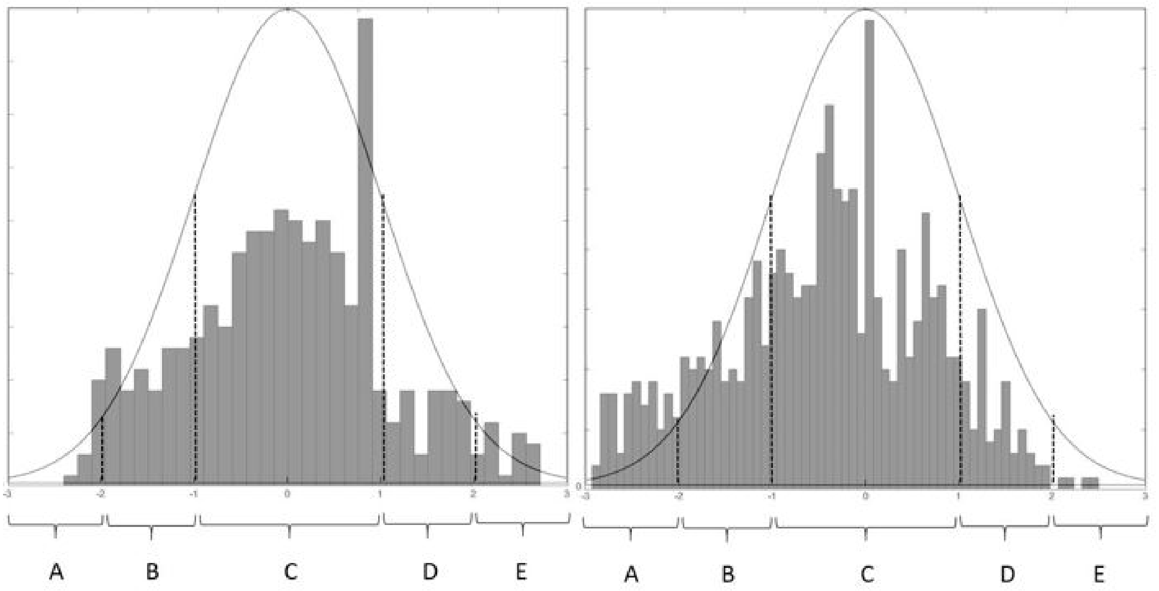


Figure 4.24: The Y Dataset (left) and the Z Dataset (right) both plotted against the Standard Distribution curve to illustrate the relative similarity with a normal curve

4.4.5 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' [113], contain many regionalised versions of websites have a global reach or presence, 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 the example of The University of Greenwich in London, UK with ranking data from Alexa Internet:

> ‘gre.ac.uk’ The University of Greenwich
> 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 UoG’s 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? 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 when they should have appeared in C. Any correlation between a high local (UK) regional ranking along with a high global (predominantly English speaking) 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.

4.4.6 HISTORIC POPULARITY AS A CONSIDERATION

As we have shown when referring to Triage in chapter 1, a law enforcement investigation can typically expect to see a lag between the time that a device was seized and the time that it was analysed. This lag could be hours in the most urgent of cases, but most realistically this can be a substantial period of time in the order of many months.

Using a global ranking system that is rapidly changing from day-to-day, hour-to-hour can lead to some variation in the data with sites intentionally jockeying for position of higher popularity or unintentionally changing position due to the ebbs and flows of Internet traffic. Figure 4.25 below shows the variation in the popularity of the University of Greenwich over a 6-month period in 2016.

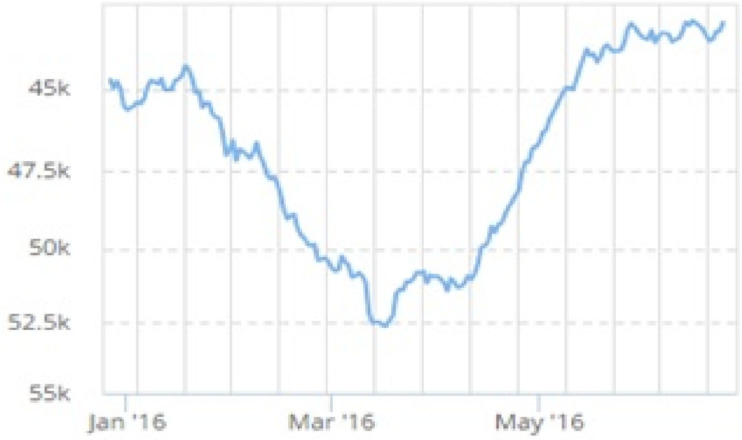


Figure 4.25: Showing the global popularity position (Y axis) of the 'gre.ac.uk' website in the first half of 2016, from Alexa Internet

In addition to the lag between seizure and analysis there may an extensive quantity of Internet history artefacts present on the device going back many years. Change of a few thousand places is highly unlikely at all to affect a particularly niche website that languishes several million places down the global popularity ranking but even a few hundred position changes may significantly affect a globally popular website.

The popularity of a website can change significantly over a number of years and when dealing with Internet history artefacts recovered from a computer it is perfectly conceivable that artefacts from over a period of years can be present.

When comparing sessions to other sessions a historical context and an understanding that an extended pattern of life, or examination of the normality of the behaviour over time should acknowledges that websites will fall in and out of favour as time passes.

Ultimately there are a number of research question relating to the temporal change in site popularity and impact. How significant or prevalent is the widespread change of popularity ratings? Is it subject to the type of website, such as social media, search engine, ecommerce? Should the GP to LP difference be recomputed for different points in the timeline, or should some overall average be taken and a single popularity metric be applied to the whole analysis? These questions remain an area of ongoing research.

4.5 OTHER APPROACHES

Grouping must be meaningful and have an emphasis on behaviour. We consider other possible forms of grouping components that are based upon the type of component, if it is a component known to be indicative of behaviour, or part of a spatial grouping.

4.5.1 GROUP BY TYPE

Grouping by type is primarily membership-based, where components are placed by checking against a large corpora of websites, which have been categorised by the general 'theme' of the website. Examples of this could be 'commerce', 'news', 'social media', 'entertainment' etc. but categorising websites, by a simple theme is non-trivial and the subject of substantial further research.

Normally we suggest that each component appear in only a single group, although we also note that each session may because of multiple components, belong to different groups. Groups by type we suggest may benefit from components appearing in multiple type groups.

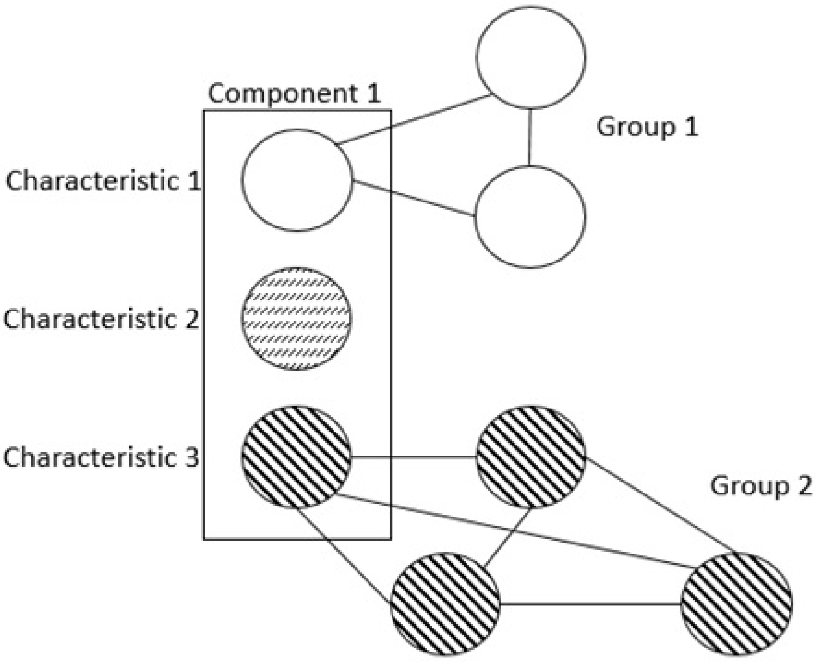


Figure 4.26: Multiple Characteristics belonging to a single component

Figure 4.26 illustrates a single component with multiple characteristics, and as such that component may belong to multiple groups.

The principal problem with this approach is that the type classification scheme is difficult to construct for every possible website, it may be possible for thousands of the most popular sites but the niche sites which are likely to be indicative of the individual would have to be classified on demand. There is also a temporal component if a site changes ownership from a historic point when it was viewed in the Internet history and the point when the analyst compiles the type data, at a later date.

There are additional problems with this approach when trying to classify a user's behaviour where there are a variety of services, or sufficiently different services that it is important to not include the component in all the groups. For example, a picture sharing website that contains non-notable photographs and pornographic pictures. It is possible to classify that session as potentially pornographic, but could lead to substantial errors and incorrect associations unless it was known exactly which pictures were being viewed.

4.5.2 GROUP BY KNOWN

Much like the grouping by Type approach, this approach does require some prior knowledge for the membership grouping, although this does not necessarily have to be from a third party. This approach is indeed a more specific form of the group by type in that specific individual sites are classified as:

- Known sites
- Known and Significant
- Unknown

Rather than a single 'Known' category, there could be multiple Known groups (Known Group 1, Known Group 2 etc.), which could be constructed from regularly associated websites. If site X and site Y appear together in Z percentage of sessions across all cases, then that could be reduced to a Known Group.

Known and Significant could include 'red flag' sites that immediately identify sessions of being of particular note such as containing websites or keywords that are specific to the circumstances of the case or generally known sites that contain objectionable material such as illegal pornography.

4.5.3 TEMPORAL GROUPING

Grouping by time characteristics allows the grouping by time of day, day of week, day of the month, combinations thereof and so on.

The following is an example of a grouping scheme which illustrates a a 24-hour period, but rather than being equally divided into fixed time chunks, it is heuristically fitted to a modern urban life-cycle where there is a presumption of the period before lunch, after lunch and before the end of the working day, the evening, the 'staying up late' and 'middle of the night/early morning':

- Early (0300 to 0700)
- Morning (0700 to 1200)
- Day (1200 to 1800)
- Evening (1800 to 2200)
- Late (2200 to 0300)

It may be desirable to also consider adjacent groups:

- Late and Early (2200 to 0700)
- Early and Morning (0300 to 1200)
- Morning and Day (0700 to 1800)
- Day and Evening (1200 to 2200)
- Evening and Late (1800 to 0300)

It may be desirable to adjust the groups based upon information within the case about the working environment of the individuals that have access to the Internet history.

This kind of grouping scheme may suit conditional approach where there are morning-only components and so on, but the number of possible grouping conditions can quickly become large especially if attempting to enumerate all of the possible options (early, early and morning, early and morning and day) etc.

4.5.4 SPATIAL GROUPING

Although beyond the scope of this thesis which is focused principally on a single desktop or laptop-style computer at a single location, there is also the consideration with laptop computers, or indeed any other kinds of mobile devices, that the device can be in different physical and different network locations.

A device can be connected to a situated wireless Wi-Fi network, such as home, work, school, coffee shop, hotel etc. A device could be connected to a limited-resource, but always-on service such as cellular 2G/3G/4G etc. How the person behaves and the sites they visit could be entirely based upon the type of network they are connected to at any time.

For example, a user may not use high-capacity media streaming whilst on a cellular connection, whilst that might be much more common whilst on unlimited hotel Wi-Fi. People may have a greater tendency to visit certain sites whilst ‘on the go’, using cellular networks, than they would whilst situated, for example reading travel, timetable information etc.

A “stay point” as described in Ye et al. [104] could also dictate the type of sites that are visited: At a work location there would be lesser chance of ‘entertainment’ sites; a public place such as a coffee shop would likely have fewer/no accesses to objectionable material such as pornography; at university there would be a higher chance for educational/subject related material; across all platforms and locations there is a likelihood of a background hum of social media. The size of the stay point is an interesting consideration: too small a location and you are profiling the behaviour at that specific coffee shop; too large a location and you will be profiling the behaviour of ‘home’, ‘work’ and ‘other’, which might be a desirable classification but ostensibly seems simplistic at this point in our research.

It is also possible to consider the multiple characteristic model of a component such as illustrated in figure 4.26, with regard to a spatial model, where a location can be both physical, logically connected to a Wi-Fi, or cellular network.

4.6 EVALUATION OF EFFECTIVE FEATURE EXTRACTION

4.6.1 SESSION LENGTH AS A FEATURE

Comparing short and long sessions as two separate membership groups is computationally quick, there is almost no additional work required as the ‘Short’ or ‘Long’ characteristic can be determined as the sessions are computed, and there is no additional computation required as the same number of comparisons are performed. Performing this analysis will have the effect of extracting the reliable sessions, i.e. the long sessions rather than the short sessions.

This approach does require a threshold value to determine if a session is ‘Short’ or ‘Long’ and we have for simplicity used the same value as is used to separate the sessions, however it is conceivable that a different threshold could be used. Other, more complex schemes can be used for this type of Membership Grouping, but we suggest that trying to extend this approach to ‘medium’, ‘long’, ‘very long’ etc. is ultimately not very productive.

We can use the ‘Short’ and ‘Long’ sessions to extract the components that appear only in ‘short-only’ in ‘long-only’ and the else condition of appearing in ‘both’ where we put the components that appear in either category. This approach has the principal advantage that no external look-up or reference source is required. It is entirely possible in realistic law enforcement scenarios when dealing with sensitive information such as websites that distribute illegal and indecent material, that the sending of the details of those sites to a 3rd party is not desired. If the investigation is in an environment where the websites are particularly niche or from a region that is not well covered by an external reference source, then the conditional grouping may be more useful than using an external membership-based reference.

4.6.2 THE GROUPS

All data within the Internet history has the capacity to be individual, even the commonly occurring 'globally popular' data as we outlined at the beginning of this chapter, where for example two users could have a distinct preference to one search engine to the exclusion of using another. However, the practical reality is that there will be some level of overlap between users. The amount of overlap one could typically expect and the general 'types of users' that can be identified from their Internet history does remain an interesting area for future work on behavioural profiling.

We do not want to create too many groups as we could miss important associations between data across different groups. However not all groups are created equal. Within the Relative Popularity difference method, we can see that there is potentially significant information across the entire set of data but the most useful components for identifying idiosyncratic behaviour without having ambiguous matches is in the 'category 1', high difference data, shown as the D and E groups (or even D and E taken together as a single group). In the Short and Long conditional method the Short-only and Long-only components appear to contain interesting information. Indeed, you might expect the Long-only data to be more significant as the user has greater amount of time to behave uniquely, but they also have greater amount of time to access components which overlap as well, whereas a person accessing a website to 'just check in', such as their social media, email or some news forum then those Short-only bursts of activity have a possibility to be unique.

We have highlighted some interesting considerations such as the temporal and regional differences that may skew the analysis, although we note that these characteristics are an open area of research in regions and for websites that do not use the English language.

4.6.3 DEGREES OF CONFIDENCE WHEN MATCHING MULTIPLE GROUPS

In situations where there are one or more groups of components in a session that match with one or more groups in another session. For example, sessions X and Y both match the Short-only groups and the Both groups, however sessions X and Z only match the Both groups. In this case, we would have a higher degree of confidence that X and Y belong to the same user not only because the Short-only components match and they are less likely to Incorrectly or Ambiguously match another user, but also because there are 2 groups matching. As long as some form of 'dialing' (with the *s-val* for example) has been used on the Both group, then we would have two tests of confidence for the sessions that this match is authentic.

In another example, Sessions X and Y could both have components in the Short-only groups and in the Both groups, however rather than matching two tests of confidence, if a stringent filtering scheme was used to test the Both group, for example S=3 and T=0.9, and this was not reached then the session could still be considered matched, but only at one test of confidence for the Short-only group.

With the Relative Popularity Difference technique above we proposed an example with 5 groups (A to E) and consequently there are 5 possible degrees of confidence for every session to session match, although realistically we find a group of A and B, a group of C and a group of D and/or E (i.e. 3 or 4 groups) more useful. With the Short and Long approach there are also 3 groups, but the Short-only and Long-only groups are mutually exclusive (the session is either short, or it is long) and therefore the maximum number of degrees of confidence are 2.

4.6.4 OTHER METHODS

GROUP BY TYPE AND KNOWN

Grouping by Type and Known categories are standard techniques in File System Analysis [14], where there are limited types of files, or large datasets of known operating system files such as the National Software Reference Library [88], due to the large number of possible components when dealing with websites this approach might not be the most appropriate at this time. The approach however remains valid if context analysis was used in the File System environment and that is an area of further research.

TEMPORAL GROUPING

For Temporal grouping the advantages are that at its most basic level it is not based upon any external knowledge, although the time groups can be adjusted based upon 'culture', such is what are the norms are in the society, for example 9 to 5 working or later 'evening' periods. The time groups could be based upon case-specific details such as events that happen during lunch breaks, or events that happen after other members of the household are alleged to have gone to bed. However, a large number of groups can be produced, especially when considering that it may be desirable to include adjacent groups.

SPATIAL GROUPING

Although the Spatial grouping approach does not require an external reference source, it does require additional information that remains outside the scope of this thesis, such as the network type, access point location, cellular tower information, or device location.

When analysing historic data, there is a concern that the locations of access points, cellular or wireless networks may have changed [49] and there are similarities with Spatial grouping to Type grouping.

We consider spatial grouping to be pattern-based rather than behavioural-based analysis and we briefly describe in the next chapter how spatial patterns-of-life could be analysed and tested.

Chapter 5

GRAPHICAL REPRESENTATION AND USE OF SESSION-TO-SESSION ANALYSIS

"Deep in the human unconscious is a pervasive need for a logical universe that makes sense. But the real universe is always one step beyond logic."

Frank Herbert - Dune, 1965

5.1 INTRODUCTION

In Chapter 3 we showed how the Session-to-Session matches between two different users could be done using a similarity coefficient, such as Jaccard's. We showed the Session-to-Session comparisons in our figures 3.25 to 3.28 and it is notable that at the high-end of the Jaccard similarity there are incorrect comparisons, and at the low-end there are extensive numbers of correct comparisons.

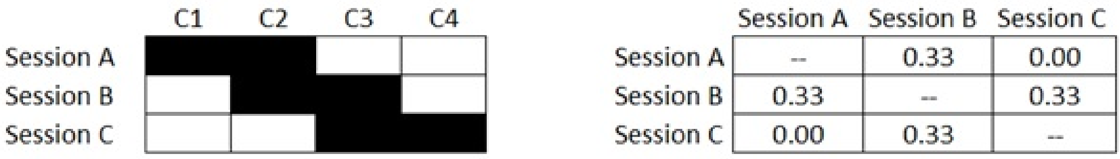
This led to the development of approaches in Chapter 4 that attempted to group sessions such that they contained a higher likelihood of similarity, such that the differences between the multiple users' sessions would stand out.

In this chapter, we show how all of the Session-to-Session comparisons between the sessions can be plotted onto a graph/network to visually represent all the periods of time as interconnected nodes and we investigate the concept we identified in chapter 3, that as we insist on using increasingly stringent variables, the availability of the sessions is reduced, and we will look at this in the context of an investigator attempting to show the interconnectedness of two or more sessions.

5.2 GRAPHICAL REPRESENTATIONS OF SESSION-TO-SESSION COMPARISONS

5.2.1 SESSION-TO-SESSION COMPARISONS

As we have noted, there are more matches than just the highest match. Some of these matches may be from the correct user, but some, especially at the lower-end of the similarity coefficient, may be incorrect matches. If trying to show that Session A is associated with Session B, and Session B is associated with Session C then the transitive association between A and C seems reasonable, except for the fact that it may be completely unrelated and belong to different users.



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

	Session A	Session B	Session C
Session A	--	0.33	0.00
Session B	0.33	--	0.33
Session C	0.00	0.33	--

Figure 5.1: The Jaccard distance between three sessions

Figure 5.1 shows that the Jaccard similarities between Sessions A and B, and Sessions B and C are 0.33, as both pairs share 1 of the 3 components, but the similarity between Sessions A and C is 0.0, as these sessions share none of the components.

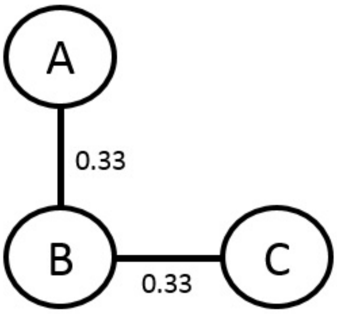


Figure 5.2: Three sessions displayed graphically

We can visually represent Session-to-session comparison as nodes connected by undirected edges, as we can see the above example of Sessions A, B and C in figure 5.2. In this case, we do not draw the 0.0 value relationship between Sessions A and C.

5.2.2 DIRECT AND INDIRECT RELATIONSHIPS

It may be desirable to include all edges, but highlight edges that fall below the matching conditions, the *t-val* as broken edges, as we can see in figure 5.3. This has the advantage of explicitly showing that there is no direct relationship, no matter how small, between Sessions A and C, but practically it is only useful for very small networks.

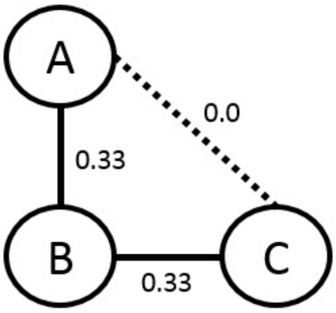


Figure 5.3: All edges in the graph shown, but non-matching edges are displayed as broken lines

There are different notations that could be used, such as a broken line for a direct connection above zero, but below the *t-val.* This notation of showing the 'nearly, but not quite' relationships has an advantage that an investigator trying to determine if there is a relationship between two sessions or events can see if there is truly no relationship between the events, or if there is a relationship but it is a low confidence one. Similarly, there is potentially a value in emphasising the very high matching comparisons so it may be desirable to make those connections appear thicker.

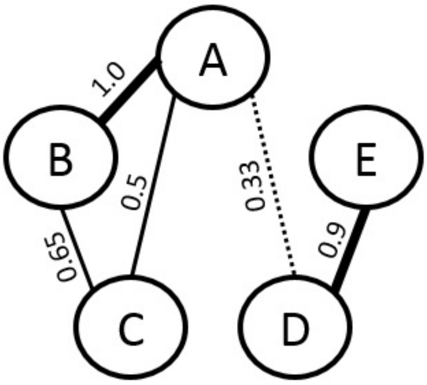


Figure 5.4: A graph showing the relationship between five sessions, using different thickness edges to indicate strength of similarity

Figure 5.4 shows an example where there are five sessions displayed as nodes, the session-to-session Jaccard values are displayed on the edges and we see that between A and D there is some form of association but the value 0.33 is displayed as a broken line and is presumably below the matching criteria, indicating that there is a direct relationship, but it is a low confidence one. The edges between A and B, and D and E are both shown as thick lines to indicate the high Jaccard similarity values.

The disadvantage of increasingly complex notation is that with the kind of graphs we see realistically coming out of this analysis, such as can be seen in section 5.3, where there are hundreds of nodes and tens of thousands of edges, this quickly becomes impractical.

A simplified notation can be seen in figure 5.5 where all of the undirected edges are a uniform thickness and no weighted value is provided, as the assumption is that as long as the matching criteria is above the *t-val*, there is sufficiently value to include it on the graph.

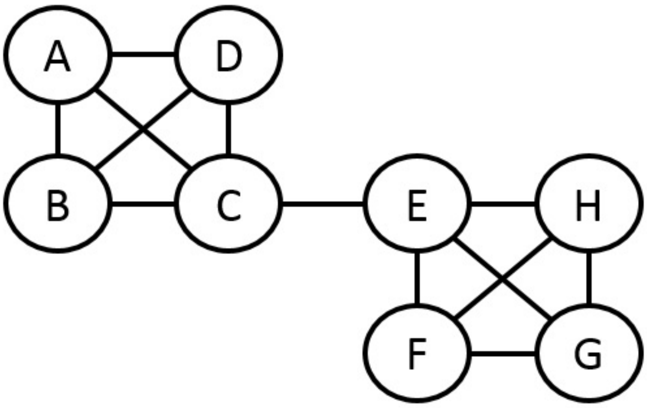


Figure 5.5: Simplified Notation showing two Cliques bridged by a single edge

Figure 5.5 illustrates that there can be interconnected clusters within the graph. There may not be an exact match (a *t-val* of 1.0) amongst A, B, C or D, but there is sufficient similarity, a core set of components that those sessions must share, which suggests that the ABCD clique constitutes some type of collective behaviour and the EFGH clique a different behaviour.

If all of the indirect connections below the *t-val*, but greater than 0.0 were included on even a relatively small graph such as shown in figure 5.5 this would less useful, and consequently we note this technique but have avoided using it.

5.2.3 SESSION-TO-SESSION GROUPS

Sessions are made up of one or more components. We have shown in the previous chapter that it is possible, indeed advantageous, to group these components based upon characteristics that will minimise the ambiguity of the user at the time of the session. What this does however mean is that there may be different groups in each session, and each of those groups have their own independent relationship with other members of the group.

We have shown that to reason A is related to B, B is related to C, therefore A is related to C can be problematic (figures 5.2 and 5.3) but this becomes even more problematic when considering the relationship between different groups.

In figure 5.6 we introduce a different notation where the session name is followed by a numeric value to indicate the group which that data belongs. We can see that there is a relationship between A and B, and C and D within the first group. There is also a relationship between B and C in the second group. Does this mean that there is a direct relationship between A and D?

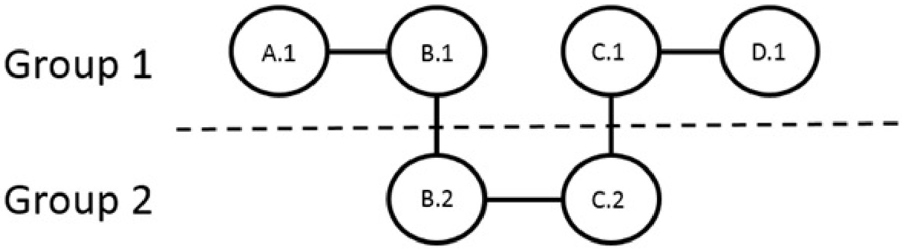


Figure 5.6: Session relationships within different groups

If we were to remove the grouping technique from the data there would likely be a direct relationship between A to D, so transitioning between groups should not alone disqualify the relationship. In figure 5.7 we see that in group 1, A is connected in cluster of sessions where it appears strongly connected to B and C, and sessions B and C also appear in group 2, strongly connected to F. A transitive relationship between A, BC and F is present and even though there is no direct connection between A and F, it may be possible for an investigator to draw and association between those sessions.

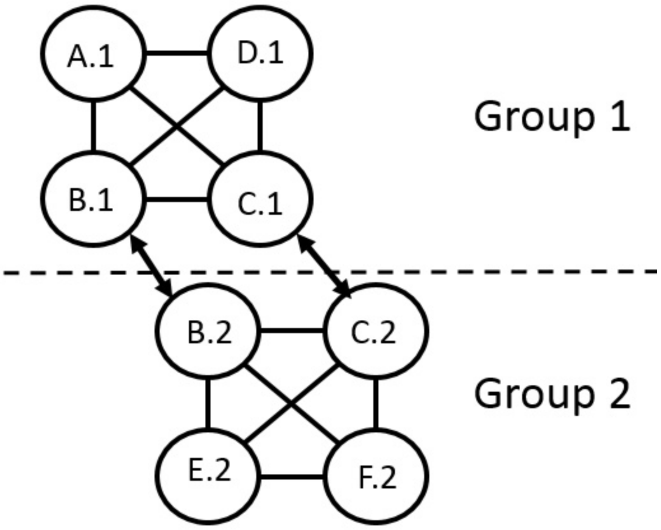


Figure 5.7: Relationship between session clusters across groups

Ultimately the different group ‘levels’ are created during the feature extraction stage (chapter 4) so as to provide assurance that the sessions belong to the same user. If there is a higher degree of confidence that Group 1 will produce idiosyncratic sessions than Group 2 will, then drawing an indirect association across groups should be considered with due consideration to that lower confidence.

5.3 MEASURING THE ACCURACY OF GROUPING

5.3.1 COMMUNITY DETECTION

Community detection is an established area of graph theory. 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.

The objective within this chapter of the thesis is to demonstrate the application of community detection. We have not exhaustively tested the available community detection algorithms and that remains an area of future work. We have essentially classified the types of community detection into weighted and non-weighted options.

A weighted approach could be to use Hierarchical Clustering, which has the advantage that the Jaccard similarity coefficient weight of the edges allow the nodes to be clustered due to strength of similarity. This is illustrated in [42].

We have primarily used the non-weighted Louvain Method of Modularity Detection implemented within the Gephi software [119] based upon the algorithm presented in Blondel et al. [9] for all 4 of our datasets and we also include in Appendix 1 the use of the weighted algorithm for the Y dataset. The difference between weighted and non-weighted does not seem particularly significant in this application. Modularity measures the density of edges inside a community and the density of the edges outside the community, the technique used in this project also allows a 'Resolution' setting to be used when determining the modularity and we present results in the appendix for modularity '1.0' and '0.1', which were selected as they represent an upper value for the calculation and a reasonable bottom end of the spectrum (discussed later in section 5.6).

5.3.2 MEASUREMENT CRITERIA

We can create a network for each of our data groupings, apply community detection to the network and then measure a number of characteristics to determine the accuracy and correctness of that grouping, and the variables (the *s-val* and *t-val*). The characteristics we can measure can be illustrated diagrammatically in figure 5.8:

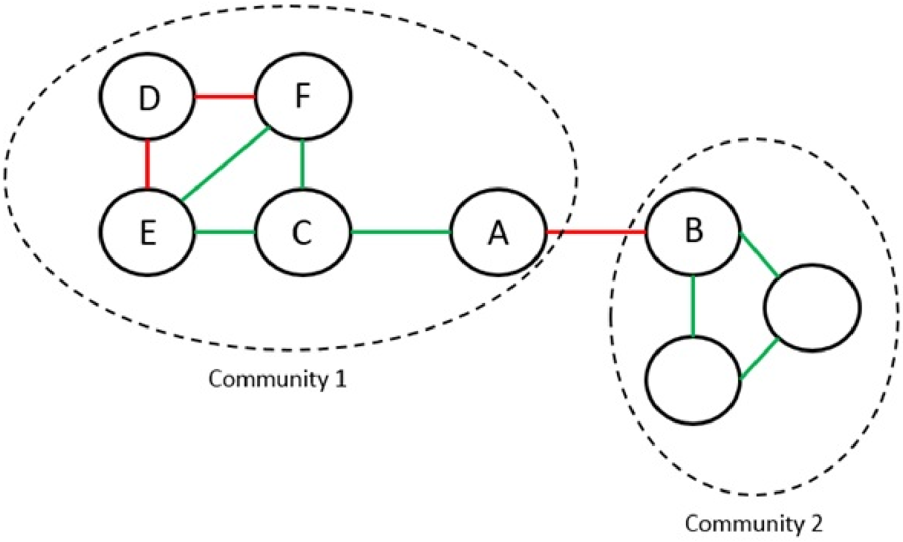


Figure 5.8: Two communities of nodes with correct matches shown with green edges and incorrect matches shown with red edges

We see in figure 5.8 that there are two communities that have been detected. The edge between A and B goes cross 2 communities, because there is a connection the implication is that community 1 and community 2 were created by the same user. If we know the ground truth that those communities belong to different users edge AB is considered a community-to-community error.

The edges DF and DE have been denoted in red also with the same implication, i.e. that node/session D was created by the same user as the other sessions within the community. If the ground truth is that edges DE and DF are incorrect we can refer to this as Intra-community error.

In this case we have one good community, Community 2 that contains no Intra-community error, and Community 1 a bad community as it contains at least one edge of intra-community error.

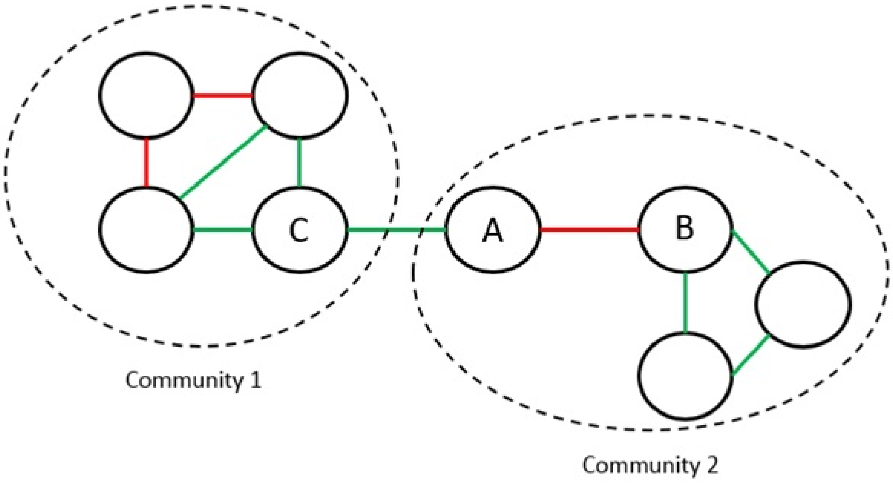


Figure 5.9: The same communities as shown in figure 5.8 but with an alternative community detection

Figure 5.9 however shows that a different community grouping where edge AB has been classified belonging to Community 2, rather than Community 1. Now edge AB is an intra-community error, whereas edge AC is a correct community-to-community edge.

5.3.3 EXPERIMENTAL RESULTS FOR DATASETS

As noted in chapter 1, "beyond reasonable doubt" (BRD) is not a well-defined standard of proof, but we will consider a goal for BRD precision to be over 91%. That can be 91% if dealing solely with Intra-community edges, or 91% total accuracy across all edges, both the intra-community and community-to-community edges. This leads to two questions which we can examine:

- Are there *s-val*, *t-val* and grouping schemes that reliably produce BRD results? If so, what kind of 'availability' of analysis does this produce?
- Can we, with a desired level of availability, predict the BRD result for groups of data?

The full set of results from these experiments have been included in Appendix 1. We demonstrate the results of our analysis on the following criteria:

- Nodes - The number of sessions in the graph.
- Good Com - The number of communities that contain no edges that are incorrect.
- Bad Com - The number of communities that contain one or more edges that are incorrect.
- Total Correct % - The percentage of edges within the graph that have nodes that correctly match to the same user.
- Total Incorrect % - The percentage of edges within the graph that have nodes that incorrectly as belonging to the same user.

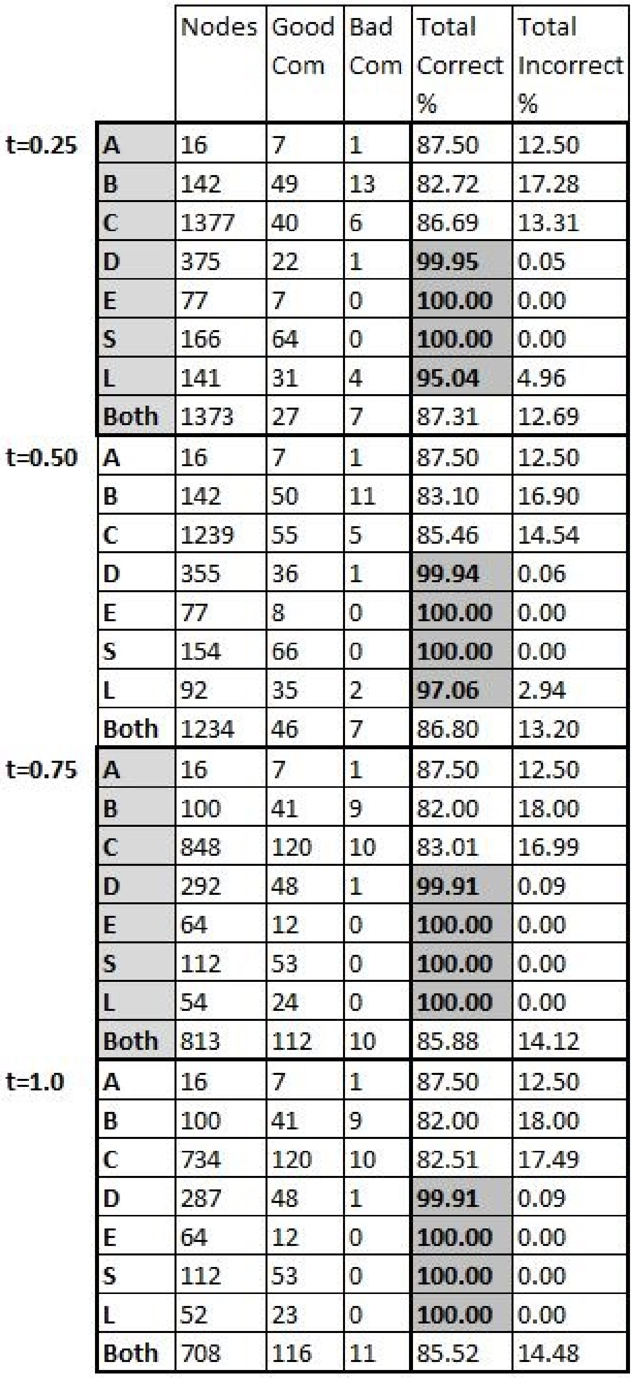
We can assess the 'availability' of our analysis as the quantity of nodes/sessions and the number of communities, compared against the 'precision' which is the number of correctly matching edges versus the incorrect matches. We can consider the availability and the precision for the different grouping schemes, and we can consider the results with respect to different *s-val* and *t-val* levels.

For simplicity, we have performed our experiments at four *t-val* levels: 0.25, 0.5, 0.75 and 1.0. These *t-val* levels do correspond with observably interesting spikes in the figures 3.25 to 3.28 and as such we believe that these are useful and representative points to assess the results.

We have performed our experiments at three *s-val* levels: S=1, S=2 and S=3. The setting S=1 is where a session group must contain at least 1 component and S=2 is a minimum of two components and so on.

W DATASET, S=1

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.1.

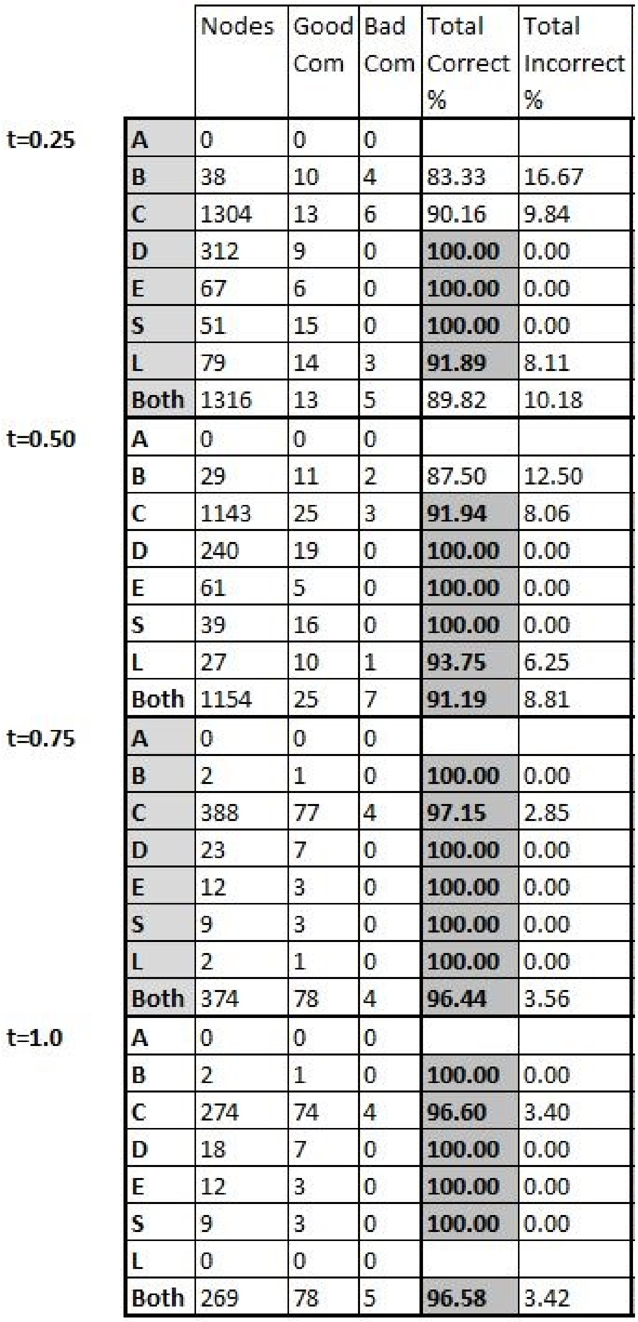


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	16	7	1	87.50	12.50
	B	142	49	13	82.72	17.28
	C	1377	40	6	86.69	13.31
	D	375	22	1	**99.95**	0.05
	E	77	7	0	**100.00**	0.00
	S	166	64	0	**100.00**	0.00
	L	141	31	4	**95.04**	4.96
	Both	1373	27	7	87.31	12.69
t=0.50	**A**	16	7	1	87.50	12.50
	B	142	50	11	83.10	16.90
	C	1239	55	5	85.46	14.54
	D	355	36	1	**99.94**	0.06
	E	77	8	0	**100.00**	0.00
	S	154	66	0	**100.00**	0.00
	L	92	35	2	**97.06**	2.94
	Both	1234	46	7	86.80	13.20
t=0.75	**A**	16	7	1	87.50	12.50
	B	100	41	9	82.00	18.00
	C	848	120	10	83.01	16.99
	D	292	48	1	**99.91**	0.09
	E	64	12	0	**100.00**	0.00
	S	112	53	0	**100.00**	0.00
	L	54	24	0	**100.00**	0.00
	Both	813	112	10	85.88	14.12
t=1.0	**A**	16	7	1	87.50	12.50
	B	100	41	9	82.00	18.00
	C	734	120	10	82.51	17.49
	D	287	48	1	**99.91**	0.09
	E	64	12	0	**100.00**	0.00
	S	112	53	0	**100.00**	0.00
	L	52	23	0	**100.00**	0.00
	Both	708	116	11	85.52	14.48

Figure 5.10: *W* dataset grouped by *s-val* S=1

W DATASET, S=2

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.2.

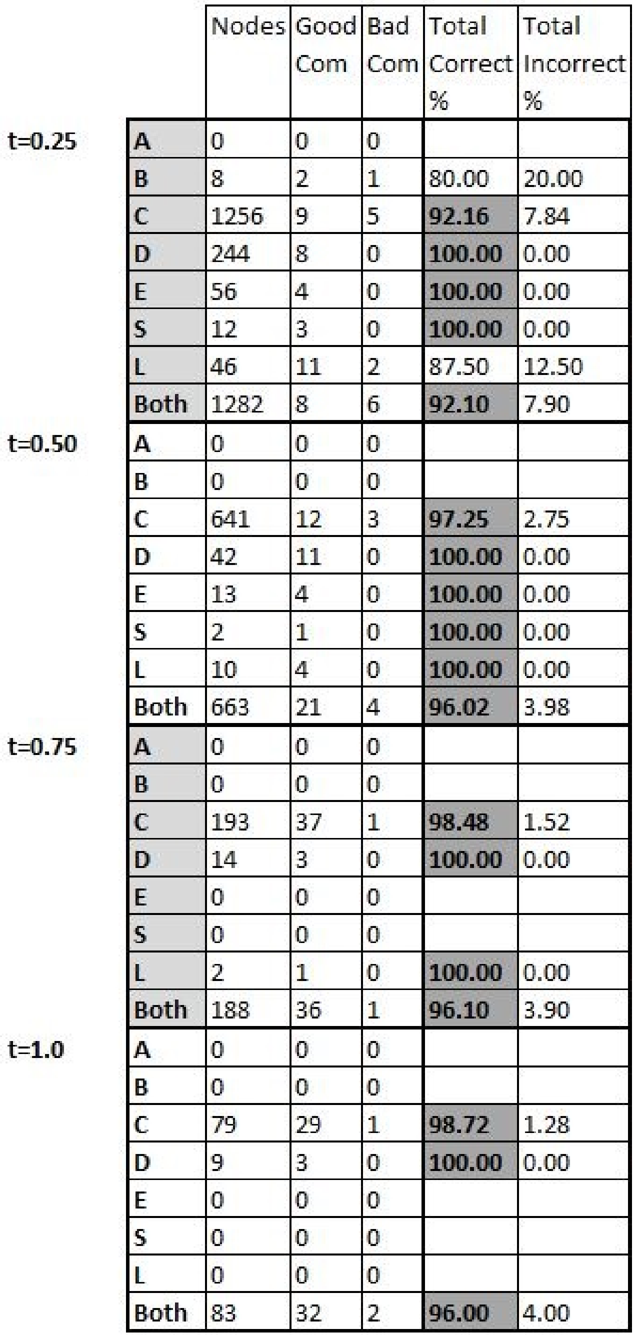


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	0	0	0		
	B	38	10	4	83.33	16.67
	C	1304	13	6	90.16	9.84
	D	312	9	0	**100.00**	0.00
	E	67	6	0	**100.00**	0.00
	S	51	15	0	**100.00**	0.00
	L	79	14	3	**91.89**	8.11
	Both	1316	13	5	89.82	10.18
t=0.50	**A**	0	0	0		
	B	29	11	2	87.50	12.50
	C	1143	25	3	**91.94**	8.06
	D	240	19	0	**100.00**	0.00
	E	61	5	0	**100.00**	0.00
	S	39	16	0	**100.00**	0.00
	L	27	10	1	**93.75**	6.25
	Both	1154	25	7	**91.19**	8.81
t=0.75	**A**	0	0	0		
	B	2	1	0	**100.00**	0.00
	C	388	77	4	**97.15**	2.85
	D	23	7	0	**100.00**	0.00
	E	12	3	0	**100.00**	0.00
	S	9	3	0	**100.00**	0.00
	L	2	1	0	**100.00**	0.00
	Both	374	78	4	**96.44**	3.56
t=1.0	**A**	0	0	0		
	B	2	1	0	**100.00**	0.00
	C	274	74	4	**96.60**	3.40
	D	18	7	0	**100.00**	0.00
	E	12	3	0	**100.00**	0.00
	S	9	3	0	**100.00**	0.00
	L	0	0	0		
	Both	269	78	5	**96.58**	3.42

Figure 5.11: *W* dataset grouped by *s-val* S=2

W DATASET, S=3

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.3.



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	A	0	0	0		
	B	8	2	1	80.00	20.00
	C	1256	9	5	92.16	7.84
	D	244	8	0	100.00	0.00
	E	56	4	0	100.00	0.00
	S	12	3	0	100.00	0.00
	L	46	11	2	87.50	12.50
	Both	1282	8	6	92.10	7.90
t=0.50	A	0	0	0		
	B	0	0	0		
	C	641	12	3	97.25	2.75
	D	42	11	0	100.00	0.00
	E	13	4	0	100.00	0.00
	S	2	1	0	100.00	0.00
	L	10	4	0	100.00	0.00
	Both	663	21	4	96.02	3.98
t=0.75	A	0	0	0		
	B	0	0	0		
	C	193	37	1	98.48	1.52
	D	14	3	0	100.00	0.00
	E	0	0	0		
	S	0	0	0		
	L	2	1	0	100.00	0.00
	Both	188	36	1	96.10	3.90
t=1.0	A	0	0	0		
	B	0	0	0		
	C	79	29	1	98.72	1.28
	D	9	3	0	100.00	0.00
	E	0	0	0		
	S	0	0	0		
	L	0	0	0		
	Both	83	32	2	96.00	4.00

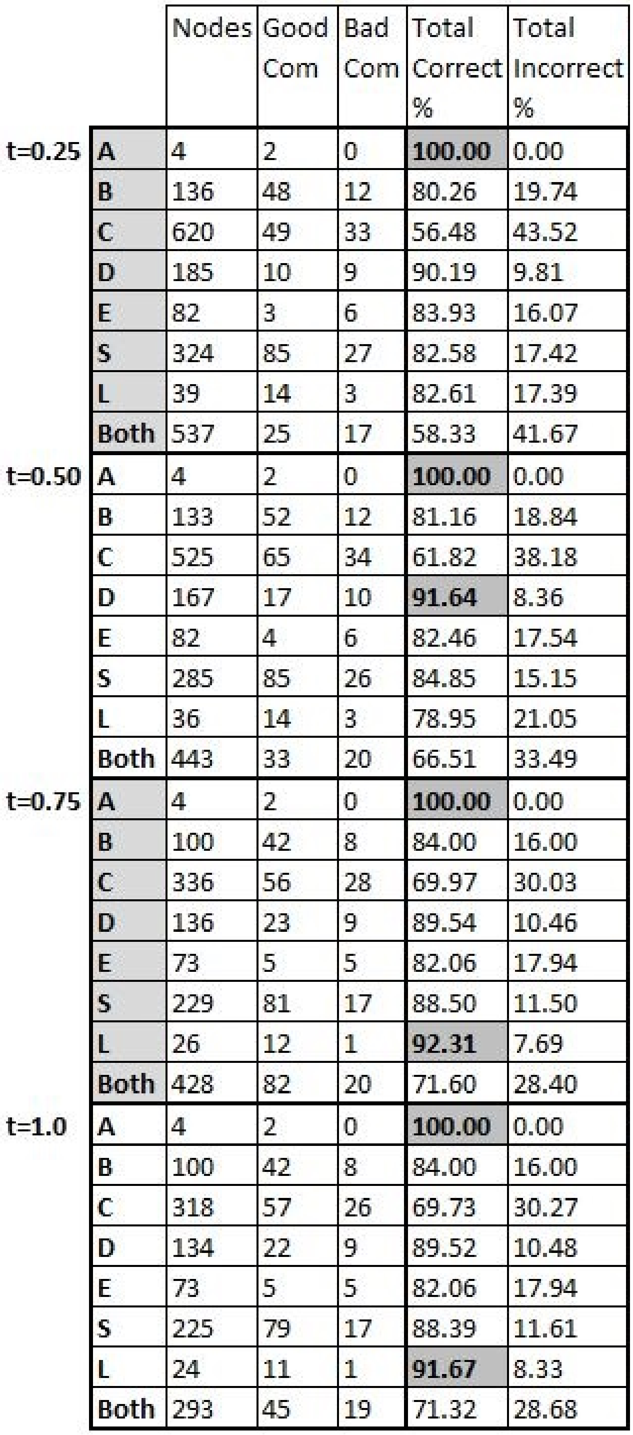
Figure 5.12: *W* dataset grouped by *s-val* S=3

COMMENTS ON THE *W* DATASET

There is a high degree of accuracy with this set for the D, E, S-only and L-only groups at a *t-val* of 0.25 and an *s-val* of S=1. An *s-val* of S=2 with a *t-val* of 0.5 improves the accuracy of C and Both to BRD.

Y DATASET, S=1

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.10.

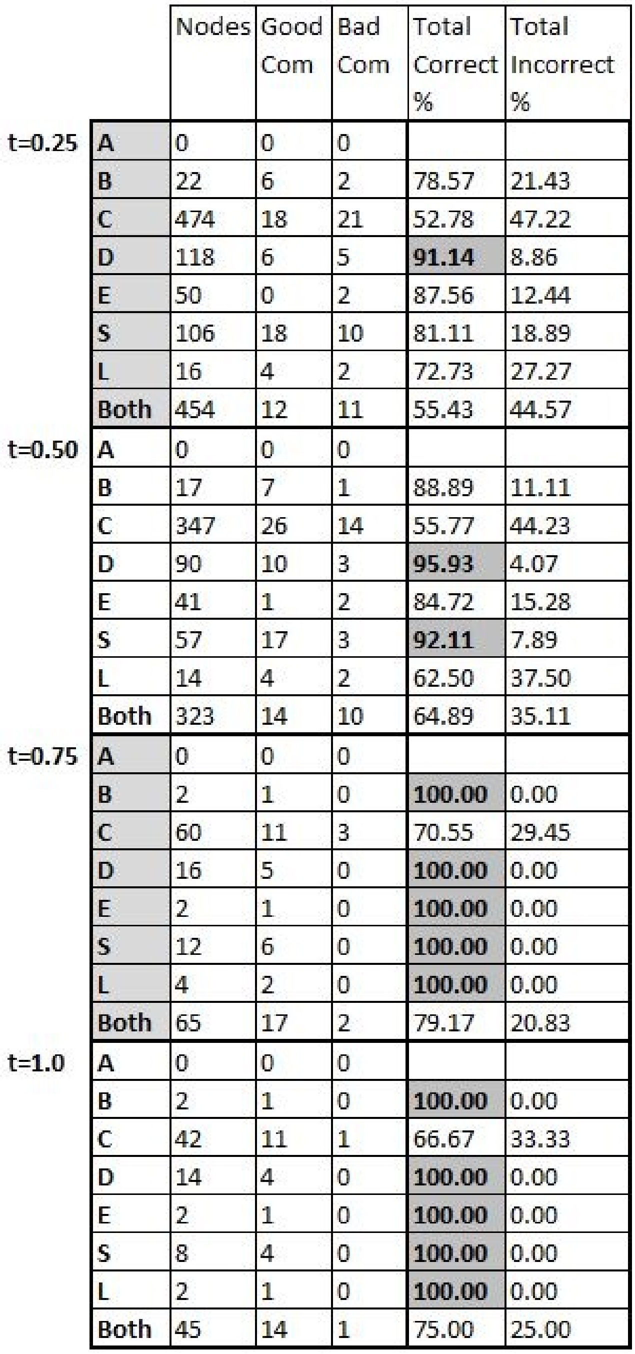


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	4	2	0	**100.00**	0.00
	B	136	48	12	80.26	19.74
	C	620	49	33	56.48	43.52
	D	185	10	9	90.19	9.81
	E	82	3	6	83.93	16.07
	S	324	85	27	82.58	17.42
	L	39	14	3	82.61	17.39
	Both	537	25	17	58.33	41.67
t=0.50	**A**	4	2	0	**100.00**	0.00
	B	133	52	12	81.16	18.84
	C	525	65	34	61.82	38.18
	D	167	17	10	**91.64**	8.36
	E	82	4	6	82.46	17.54
	S	285	85	26	84.85	15.15
	L	36	14	3	78.95	21.05
	Both	443	33	20	66.51	33.49
t=0.75	**A**	4	2	0	**100.00**	0.00
	B	100	42	8	84.00	16.00
	C	336	56	28	69.97	30.03
	D	136	23	9	89.54	10.46
	E	73	5	5	82.06	17.94
	S	229	81	17	88.50	11.50
	L	26	12	1	**92.31**	7.69
	Both	428	82	20	71.60	28.40
t=1.0	**A**	4	2	0	**100.00**	0.00
	B	100	42	8	84.00	16.00
	C	318	57	26	69.73	30.27
	D	134	22	9	89.52	10.48
	E	73	5	5	82.06	17.94
	S	225	79	17	88.39	11.61
	L	24	11	1	**91.67**	8.33
	Both	293	45	19	71.32	28.68

Figure 5.13: *Y* dataset grouped by *s-val* S=1

Y DATASET, S=2

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.11.

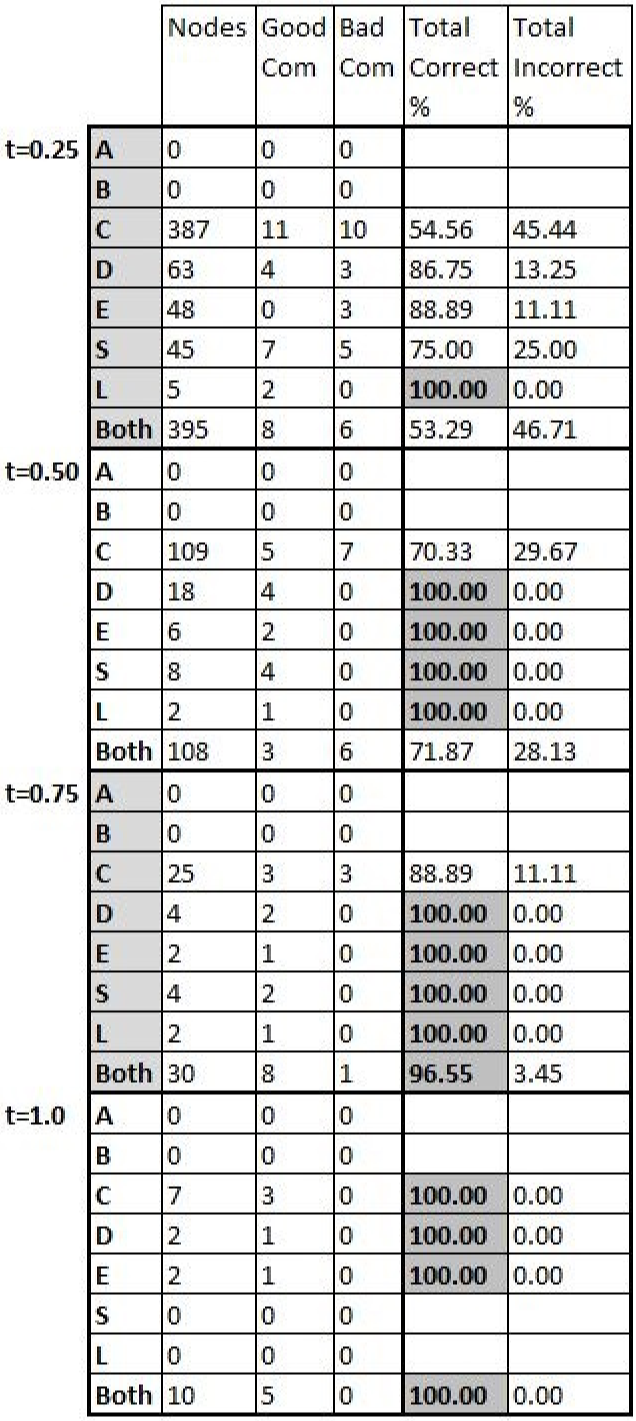


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	0	0	0		
	B	22	6	2	78.57	21.43
	C	474	18	21	52.78	47.22
	D	118	6	5	**91.14**	8.86
	E	50	0	2	87.56	12.44
	S	106	18	10	81.11	18.89
	L	16	4	2	72.73	27.27
	Both	454	12	11	55.43	44.57
t=0.50	A	0	0	0		
	B	17	7	1	88.89	11.11
	C	347	26	14	55.77	44.23
	D	90	10	3	**95.93**	4.07
	E	41	1	2	84.72	15.28
	S	57	17	3	**92.11**	7.89
	L	14	4	2	62.50	37.50
	Both	323	14	10	64.89	35.11
t=0.75	**A**	0	0	0		
	B	2	1	0	**100.00**	0.00
	C	60	11	3	70.55	29.45
	D	16	5	0	**100.00**	0.00
	E	2	1	0	**100.00**	0.00
	S	12	6	0	**100.00**	0.00
	L	4	2	0	**100.00**	0.00
	Both	65	17	2	79.17	20.83
t=1.0	A	0	0	0		
	B	2	1	0	**100.00**	0.00
	C	42	11	1	66.67	33.33
	D	14	4	0	**100.00**	0.00
	E	2	1	0	**100.00**	0.00
	S	8	4	0	**100.00**	0.00
	L	2	1	0	**100.00**	0.00
	Both	45	14	1	75.00	25.00

Figure 5.14: *Y* dataset grouped by *s-val* S=2

Y DATASET, S=3

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.12.



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	A	0	0	0		
	B	0	0	0		
	C	387	11	10	54.56	45.44
	D	63	4	3	86.75	13.25
	E	48	0	3	88.89	11.11
	S	45	7	5	75.00	25.00
	L	5	2	0	100.00	0.00
	Both	395	8	6	53.29	46.71
t=0.50	A	0	0	0		
	B	0	0	0		
	C	109	5	7	70.33	29.67
	D	18	4	0	100.00	0.00
	E	6	2	0	100.00	0.00
	S	8	4	0	100.00	0.00
	L	2	1	0	100.00	0.00
	Both	108	3	6	71.87	28.13
t=0.75	A	0	0	0		
	B	0	0	0		
	C	25	3	3	88.89	11.11
	D	4	2	0	100.00	0.00
	E	2	1	0	100.00	0.00
	S	4	2	0	100.00	0.00
	L	2	1	0	100.00	0.00
	Both	30	8	1	96.55	3.45
t=1.0	A	0	0	0		
	B	0	0	0		
	C	7	3	0	100.00	0.00
	D	2	1	0	100.00	0.00
	E	2	1	0	100.00	0.00
	S	0	0	0		
	L	0	0	0		
	Both	10	5	0	100.00	0.00

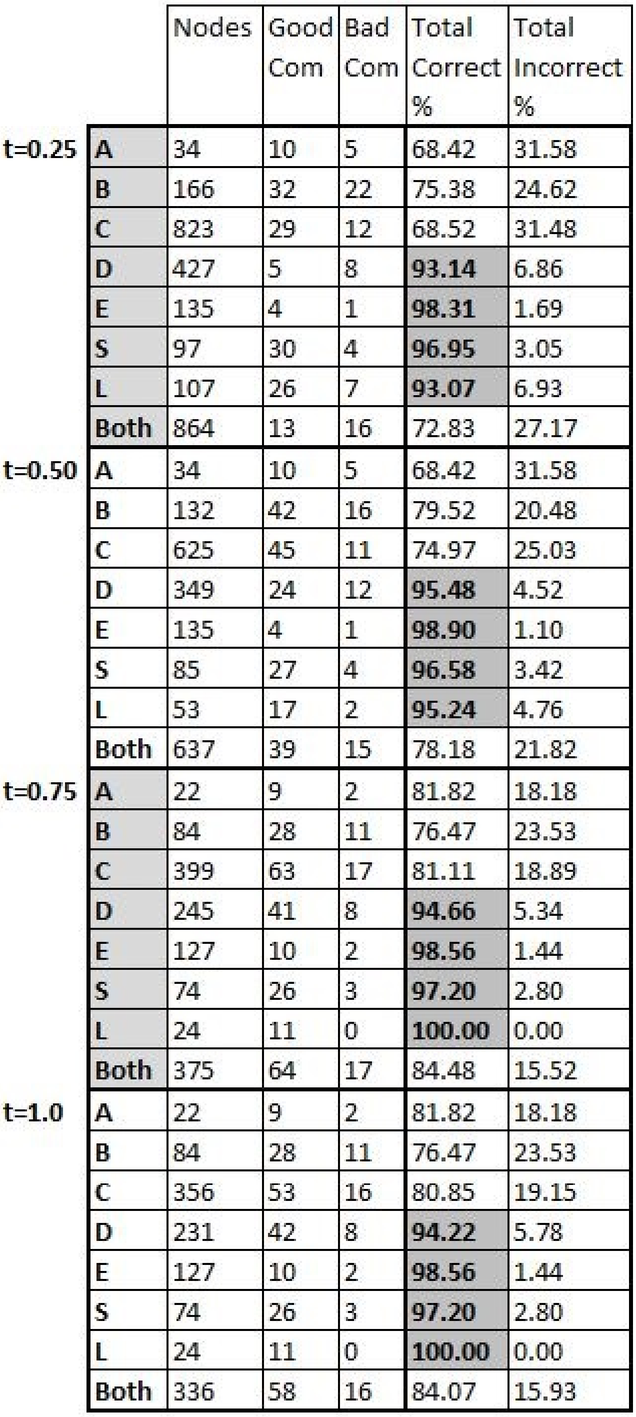
Figure 5.15: *Y* dataset grouped by *s-val* S=2

COMMENTS ON THE *Y* DATASET

The D, E, S-only and L-only groups perform consistently well at a *t-val* of 0.75 and an *s-val* of S=2, and skirts the BRD values for S=1, suggesting there may be some overlap between the two users in this dataset. The C and the Both groups do not pass the BRD level until the *s-val* and *t-val* has been adjusted so high that the number of nodes and communities is reduced to a small quantity.

Z **DATASET, S=1**

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.7.

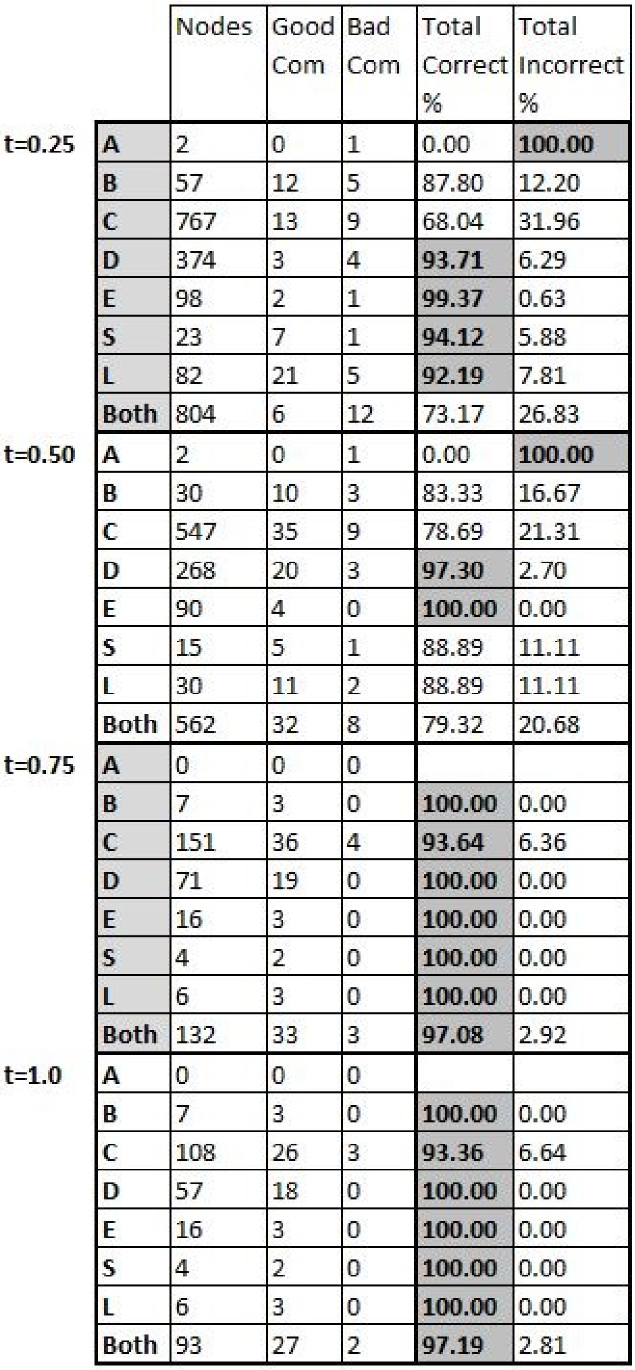


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	34	10	5	68.42	31.58
	B	166	32	22	75.38	24.62
	C	823	29	12	68.52	31.48
	D	427	5	8	**93.14**	6.86
	E	135	4	1	**98.31**	1.69
	S	97	30	4	**96.95**	3.05
	L	107	26	7	**93.07**	6.93
	Both	864	13	16	72.83	27.17
t=0.50	**A**	34	10	5	68.42	31.58
	B	132	42	16	79.52	20.48
	C	625	45	11	74.97	25.03
	D	349	24	12	**95.48**	4.52
	E	135	4	1	**98.90**	1.10
	S	85	27	4	**96.58**	3.42
	L	53	17	2	**95.24**	4.76
	Both	637	39	15	78.18	21.82
t=0.75	**A**	22	9	2	81.82	18.18
	B	84	28	11	76.47	23.53
	C	399	63	17	81.11	18.89
	D	245	41	8	**94.66**	5.34
	E	127	10	2	**98.56**	1.44
	S	74	26	3	**97.20**	2.80
	L	24	11	0	**100.00**	0.00
	Both	375	64	17	84.48	15.52
t=1.0	**A**	22	9	2	81.82	18.18
	B	84	28	11	76.47	23.53
	C	356	53	16	80.85	19.15
	D	231	42	8	**94.22**	5.78
	E	127	10	2	**98.56**	1.44
	S	74	26	3	**97.20**	2.80
	L	24	11	0	**100.00**	0.00
	Both	336	58	16	84.07	15.93

Figure 5.16: *Z* dataset grouped by *s-val* S=1

Z DATASET, S=2

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.8.

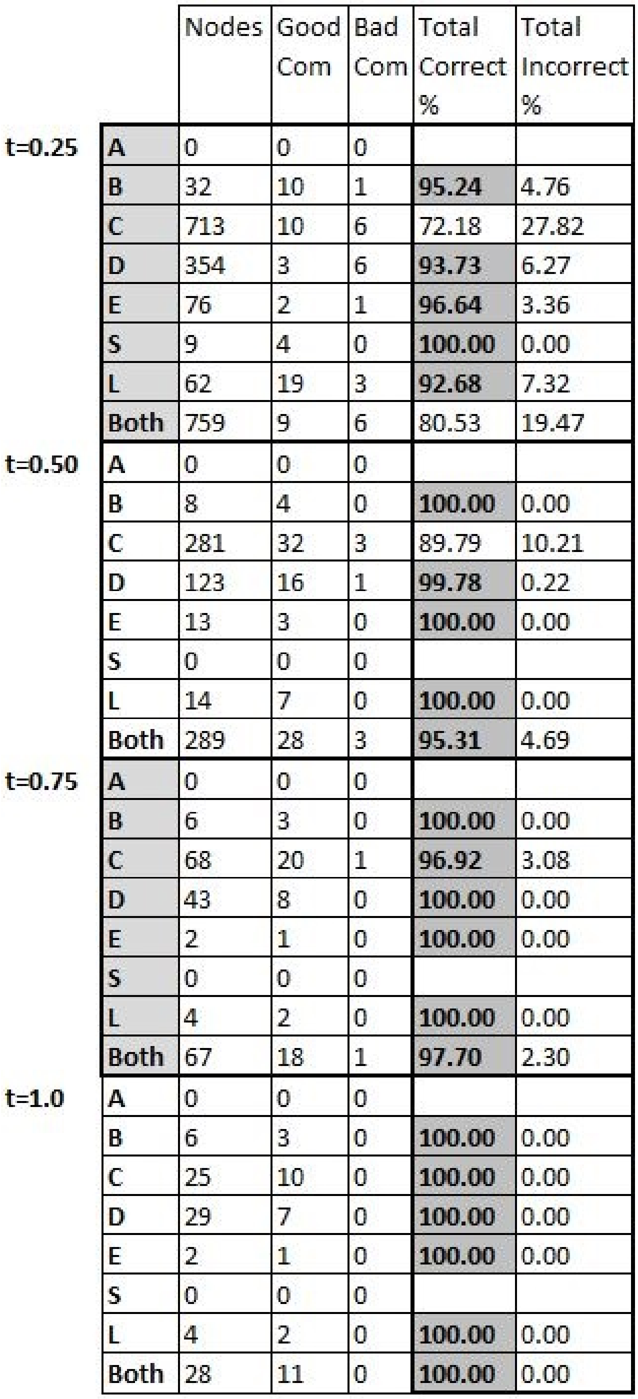


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	2	0	1	0.00	**100.00**
	B	57	12	5	87.80	12.20
	C	767	13	9	68.04	31.96
	D	374	3	4	**93.71**	6.29
	E	98	2	1	**99.37**	0.63
	S	23	7	1	**94.12**	5.88
	L	82	21	5	**92.19**	7.81
	Both	804	6	12	73.17	26.83
t=0.50	**A**	2	0	1	0.00	**100.00**
	B	30	10	3	83.33	16.67
	C	547	35	9	78.69	21.31
	D	268	20	3	**97.30**	2.70
	E	90	4	0	**100.00**	0.00
	S	15	5	1	88.89	11.11
	L	30	11	2	88.89	11.11
	Both	562	32	8	79.32	20.68
t=0.75	**A**	0	0	0		
	B	7	3	0	**100.00**	0.00
	C	151	36	4	**93.64**	6.36
	D	71	19	0	**100.00**	0.00
	E	16	3	0	**100.00**	0.00
	S	4	2	0	**100.00**	0.00
	L	6	3	0	**100.00**	0.00
	Both	132	33	3	**97.08**	2.92
t=1.0	**A**	0	0	0		
	B	7	3	0	**100.00**	0.00
	C	108	26	3	**93.36**	6.64
	D	57	18	0	**100.00**	0.00
	E	16	3	0	**100.00**	0.00
	S	4	2	0	**100.00**	0.00
	L	6	3	0	**100.00**	0.00
	Both	93	27	2	**97.19**	2.81

Figure 5.17: *Z* dataset grouped by *s-val* S=2

***Z* DATASET, S=3**

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.9.



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	0	0	0		
	B	32	10	1	**95.24**	4.76
	C	713	10	6	72.18	27.82
	D	354	3	6	**93.73**	6.27
	E	76	2	1	**96.64**	3.36
	S	9	4	0	**100.00**	0.00
	L	62	19	3	**92.68**	7.32
	Both	759	9	6	80.53	19.47
t=0.50	**A**	0	0	0		
	B	8	4	0	**100.00**	0.00
	C	281	32	3	89.79	10.21
	D	123	16	1	**99.78**	0.22
	E	13	3	0	**100.00**	0.00
	S	0	0	0		
	L	14	7	0	**100.00**	0.00
	Both	289	28	3	**95.31**	4.69
t=0.75	**A**	0	0	0		
	B	6	3	0	**100.00**	0.00
	C	68	20	1	**96.92**	3.08
	D	43	8	0	**100.00**	0.00
	E	2	1	0	**100.00**	0.00
	S	0	0	0		
	L	4	2	0	**100.00**	0.00
	Both	67	18	1	**97.70**	2.30
t=1.0	**A**	0	0	0		
	B	6	3	0	**100.00**	0.00
	C	25	10	0	**100.00**	0.00
	D	29	7	0	**100.00**	0.00
	E	2	1	0	**100.00**	0.00
	S	0	0	0		
	L	4	2	0	**100.00**	0.00
	Both	28	11	0	**100.00**	0.00

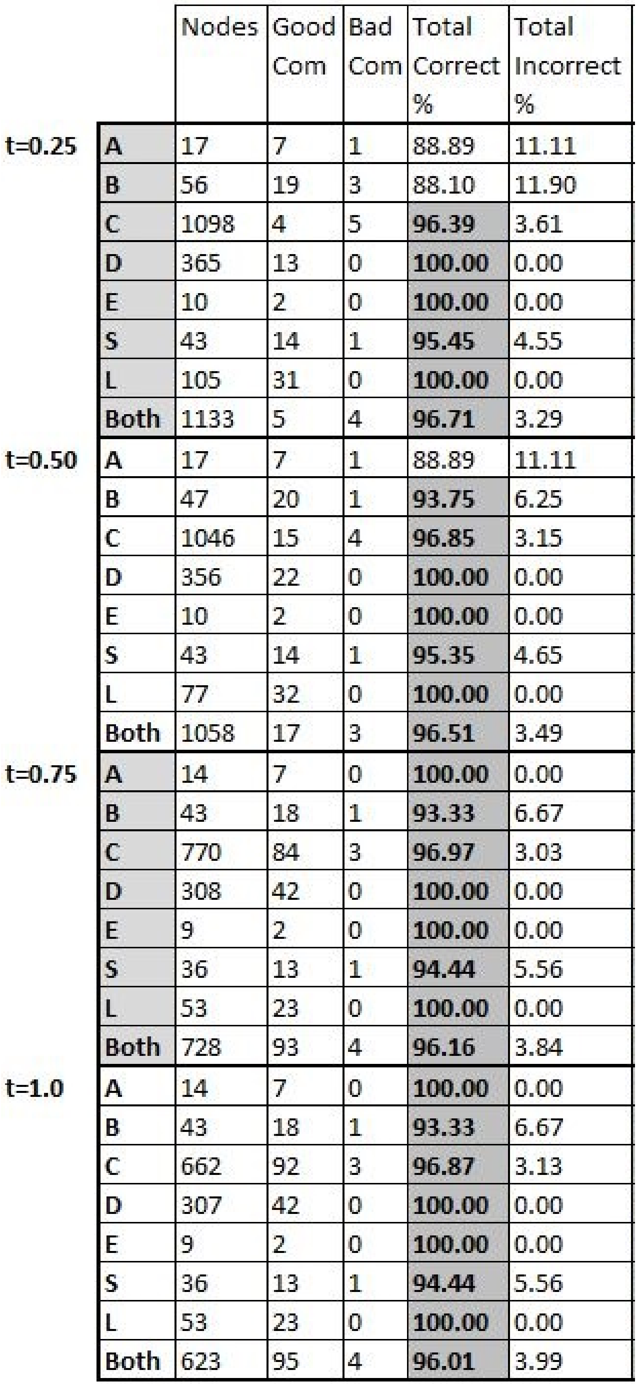
Figure 5.18: *Z* dataset grouped by *s-val* S=3

COMMENTS ON THE *Z* DATASET

There is a high degree of accuracy with this set for the D, E, S-only and L-only groups at a *t-val* of 0.25 and an *s-val* of S=1. An *s-val* of S=2 with a *t-val* of 0.75 improves the accuracy of C and Both to BRD.

X DATASET, S=1

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.4.

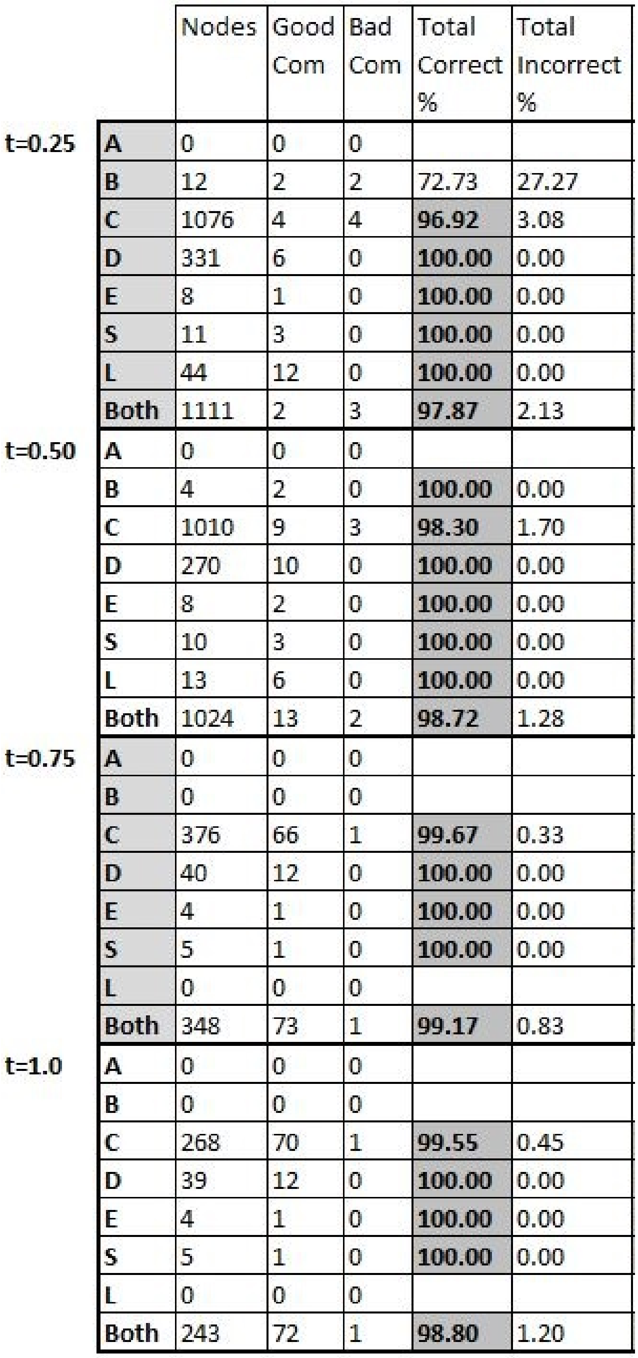


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	**A**	17	7	1	88.89	11.11
	B	56	19	3	88.10	11.90
	C	1098	4	5	**96.39**	3.61
	D	365	13	0	**100.00**	0.00
	E	10	2	0	**100.00**	0.00
	S	43	14	1	**95.45**	4.55
	L	105	31	0	**100.00**	0.00
	Both	1133	5	4	**96.71**	3.29
t=0.50	**A**	17	7	1	88.89	11.11
	B	47	20	1	**93.75**	6.25
	C	1046	15	4	**96.85**	3.15
	D	356	22	0	**100.00**	0.00
	E	10	2	0	**100.00**	0.00
	S	43	14	1	**95.35**	4.65
	L	77	32	0	**100.00**	0.00
	Both	1058	17	3	**96.51**	3.49
t=0.75	**A**	14	7	0	**100.00**	0.00
	B	43	18	1	**93.33**	6.67
	C	770	84	3	**96.97**	3.03
	D	308	42	0	**100.00**	0.00
	E	9	2	0	**100.00**	0.00
	S	36	13	1	**94.44**	5.56
	L	53	23	0	**100.00**	0.00
	Both	728	93	4	**96.16**	3.84
t=1.0	**A**	14	7	0	**100.00**	0.00
	B	43	18	1	**93.33**	6.67
	C	662	92	3	**96.87**	3.13
	D	307	42	0	**100.00**	0.00
	E	9	2	0	**100.00**	0.00
	S	36	13	1	**94.44**	5.56
	L	53	23	0	**100.00**	0.00
	Both	623	95	4	**96.01**	3.99

Figure 5.19: *X* dataset grouped by *s-val* S=1

X DATASET, S=2

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.5.

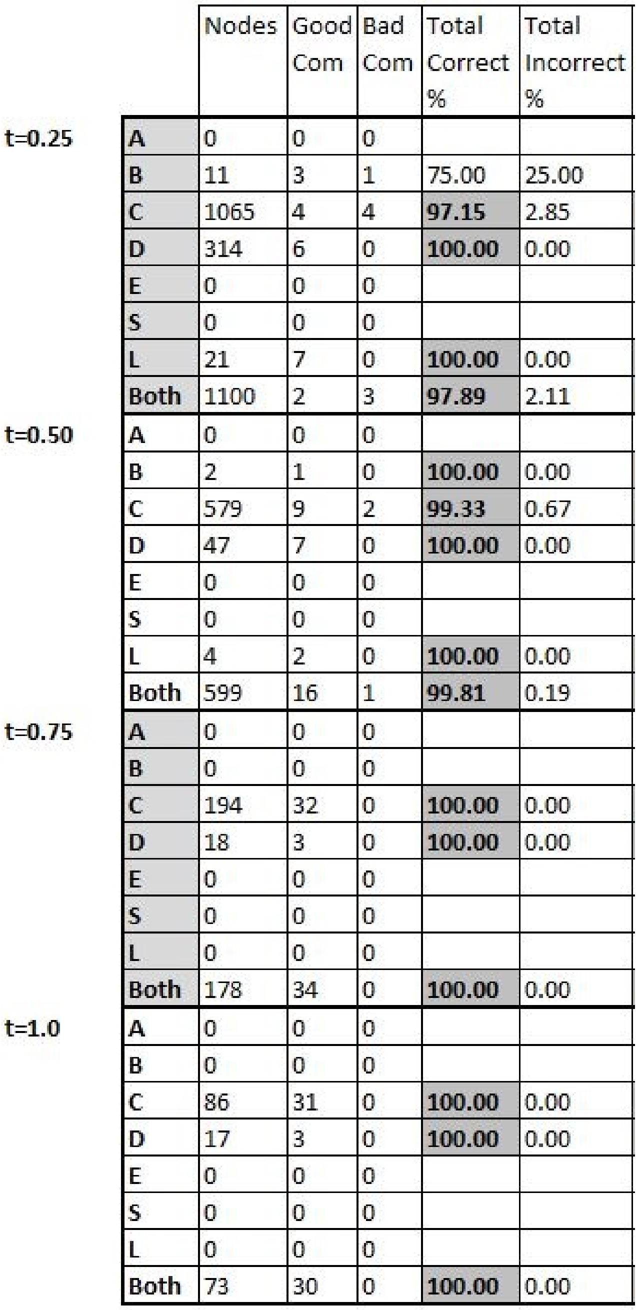


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	A	0	0	0		
	B	12	2	2	72.73	27.27
	C	1076	4	4	96.92	3.08
	D	331	6	0	100.00	0.00
	E	8	1	0	100.00	0.00
	S	11	3	0	100.00	0.00
	L	44	12	0	100.00	0.00
	Both	1111	2	3	97.87	2.13
t=0.50	A	0	0	0		
	B	4	2	0	100.00	0.00
	C	1010	9	3	98.30	1.70
	D	270	10	0	100.00	0.00
	E	8	2	0	100.00	0.00
	S	10	3	0	100.00	0.00
	L	13	6	0	100.00	0.00
	Both	1024	13	2	98.72	1.28
t=0.75	A	0	0	0		
	B	0	0	0		
	C	376	66	1	99.67	0.33
	D	40	12	0	100.00	0.00
	E	4	1	0	100.00	0.00
	S	5	1	0	100.00	0.00
	L	0	0	0		
	Both	348	73	1	99.17	0.83
t=1.0	A	0	0	0		
	B	0	0	0		
	C	268	70	1	99.55	0.45
	D	39	12	0	100.00	0.00
	E	4	1	0	100.00	0.00
	S	5	1	0	100.00	0.00
	L	0	0	0		
	Both	243	72	1	98.80	1.20

Figure 5.20: *X* dataset grouped by *s-val* S=2

X DATASET, S=3

The results showing the total percentage of correct session-to-session edges at various *t-val*s. The full results for this dataset can be found in appendix A.6.



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %
t=0.25	A	0	0	0		
	B	11	3	1	75.00	25.00
	C	1065	4	4	**97.15**	2.85
	D	314	6	0	**100.00**	0.00
	E	0	0	0		
	S	0	0	0		
	L	21	7	0	**100.00**	0.00
	Both	1100	2	3	**97.89**	2.11
t=0.50	A	0	0	0		
	B	2	1	0	**100.00**	0.00
	C	579	9	2	**99.33**	0.67
	D	47	7	0	**100.00**	0.00
	E	0	0	0		
	S	0	0	0		
	L	4	2	0	**100.00**	0.00
	Both	599	16	1	**99.81**	0.19
t=0.75	A	0	0	0		
	B	0	0	0		
	C	194	32	0	**100.00**	0.00
	D	18	3	0	**100.00**	0.00
	E	0	0	0		
	S	0	0	0		
	L	0	0	0		
	Both	178	34	0	**100.00**	0.00
t=1.0	A	0	0	0		
	B	0	0	0		
	C	86	31	0	**100.00**	0.00
	D	17	3	0	**100.00**	0.00
	E	0	0	0		
	S	0	0	0		
	L	0	0	0		
	Both	73	30	0	**100.00**	0.00

Figure 5.21: *X* dataset grouped by *s-val* S=3

COMMENTS ON THE *X* DATASET

The dataset is similar to the *W* and as such there is a high degree of accuracy with this set for the D, E, S-only and L-only groups at a *t-val* of 0.25 and an *s-val* of S=1. In this set the C and Both groups are also at the BRD level with a *t-val* of 0.25 and an *s-val* of S=1.

5.3.4 OVERALL COMMENTS ON ALL FOUR DATASETS

There is not a single clear setting of the *t-val* and *s-val* for particular groupings that is the ‘best’.

For the *W*, *X* and the *Z* datasets the D, E, S-only and L-only groupings all perform well at an *s-val* of S=1. The *t-val* of 0.25 produces a BRD of 91% accuracy across the whole dataset, with an increase to accuracy as the *t-val* is raised to 0.5, with a corresponding drop in the number of nodes/sessions on the graph.

With the *W*, *X* and the *Z* sets, the C and Both groups perform generally well when the *s-val* is S=2 and with higher values of the *t-val* such as 0.75.

The *Y* group has an overlap between the two users, not that dissimilar to *Z* (i.e. the low 30%), but the performance of this dataset requires higher *s-val* and *t-val* to achieve BRD results across all of the grouping schemes.

Groups A and B generally perform fairly poorly (below the BRD), or when the *s-val* and *t-val* are raised to the point where they do pass the 91% mark, the number of nodes and communities that remain are very low. These groups have therefore been excluded from the following graphs.

5.3.5 PLOTTING THE ACCURACY OF THE GROUPING METHODS

As we know the ground truth of the results, shown in tables Figure 5.10 to 5.21, we can plot each of the *t-val*s as a sequence of *s-vals* and use a line of best fit to estimate the shape of the performance. In figure 5.22 we can see that if there is a known set of sessions that we want to have on our graph, in this example we want to display 1000 nodes and we want to produce results above the reasonable doubt level of 91%, then we see that the S=1 line is not going to produce the accuracy we demand. The S=2 and S=3 line will produce the correct number of nodes if for the S=2 line the *t-val* is between 0.5 and 0.75 (but closer to 0.5). Whereas for the S=3 line the *t-val* should be set between 0.25 and 0.5 (but closer to 0.25).

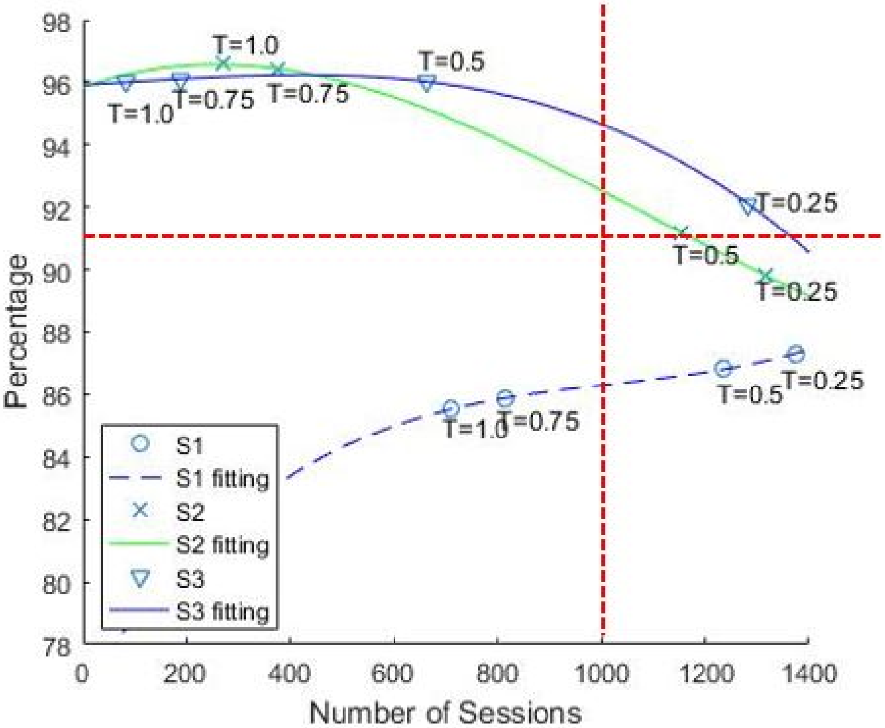


Figure 5.22: The *W* dataset showing the 'Both' group. The percentage of correctly matching edges (Y-Axis) compared against the number of session (X-axis), for different *s-val* and *t-val* settings

We illustrate this approach used for the *W* dataset in figures 5.23 to 5.28. Ultimately the *W* dataset performs well overall with large numbers of nodes available for analysis

The W dataset showing percentage of correct edges (Y-Axis) versus number of sessions (X-axis):

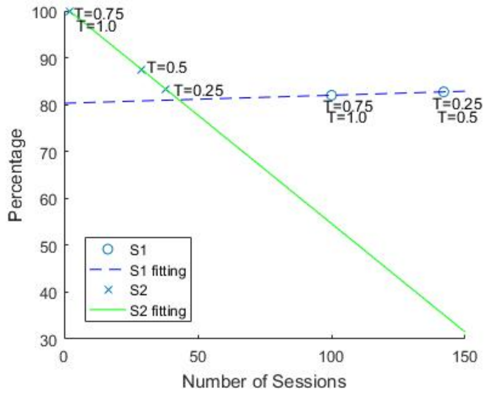


Figure 5.23: ‘B’ group

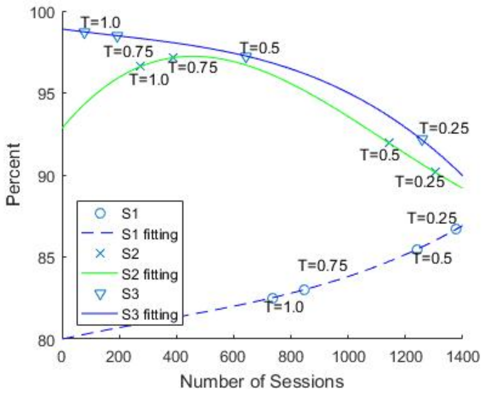


Figure 5.24: ‘C’ group

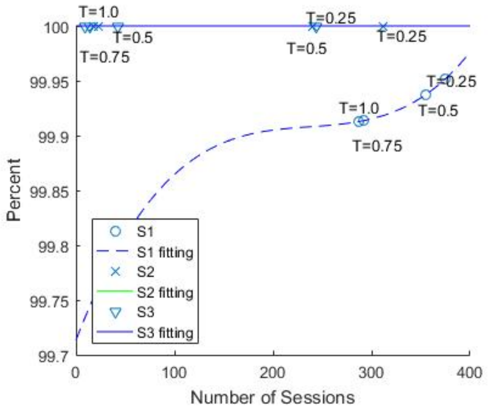


Figure 5.25: ‘D’ group

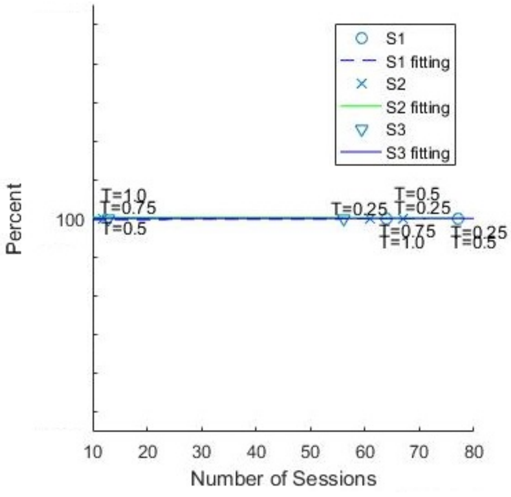


Figure 5.26: ‘E’ group

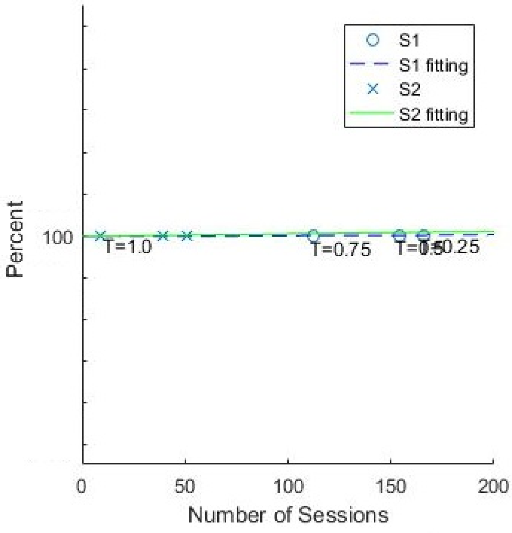


Figure 5.27: ‘S-only’ group

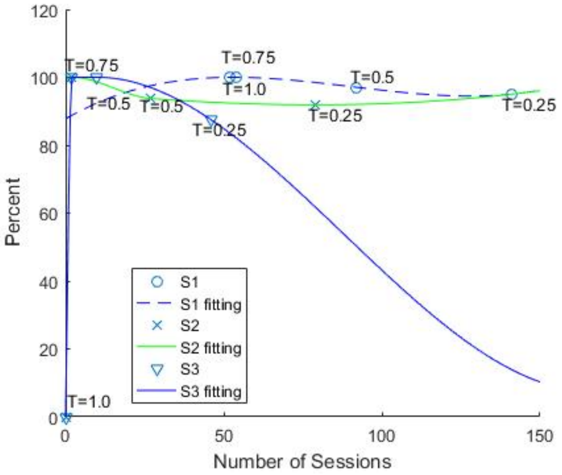


Figure 5.28: ‘L-only’ group

(X axis) corresponding with high degrees of accurate edges between those nodes (Y axis). It can be seen in tables 5.10 to 5.21 that not all of the datasets perform this well, so we present a comparative plot for the different grouping types C, D, S, L and Both groupings below.

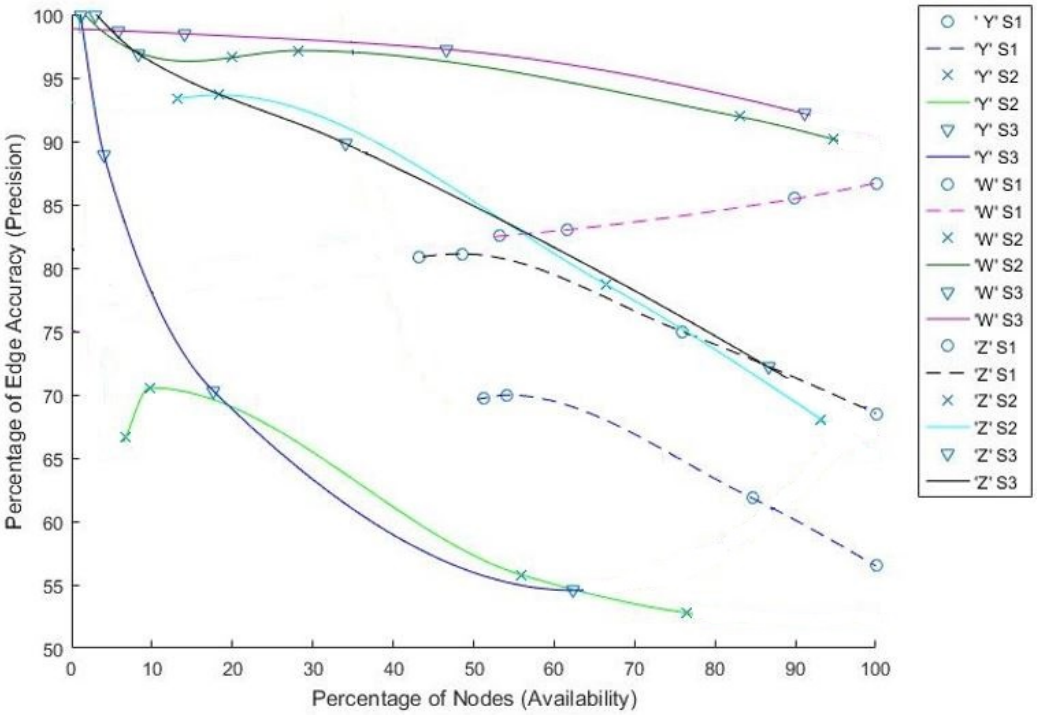


Figure 5.29: The C Grouping for the *W*, *Y* and *Z* datasets with the lower *t-val* T=0.25 plots to the right, increasing in value as sequence moves to the left

Figure 5.29 shows the C grouping for the *W*, *Y* and *Z* datasets (we have omitted the *X* dataset as it is highly accurate across all groupings). We see that the S=3 plots all reach 100% correct and the S=2 plots almost all reach 100% correct (the *Y* dataset does not get above the BRD level). We note however that the gradient on the lines of best fit is sufficiently steep that there are few nodes available at the BRD level.

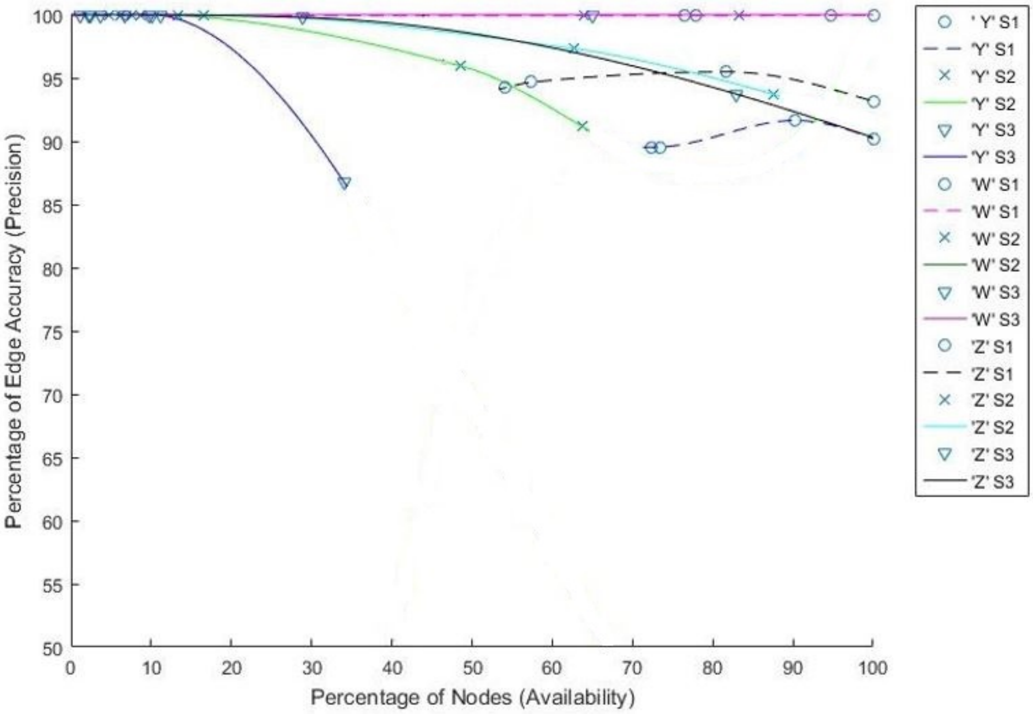


Figure 5.30: The D Grouping for the W, Y and Z datasets with the lower *t-val* T=0.25 plots to the right, increasing in value as sequence moves to the left

Figure 5.30 shows the D grouping for the W, Y and Z datasets (we have omitted the X dataset as noted). Although the performance is good in this grouping scheme, an interesting feature of S=1 is that there is an initial rise in accuracy between T=0.25 and T=0.5 but then we can see a falling off of accuracy.

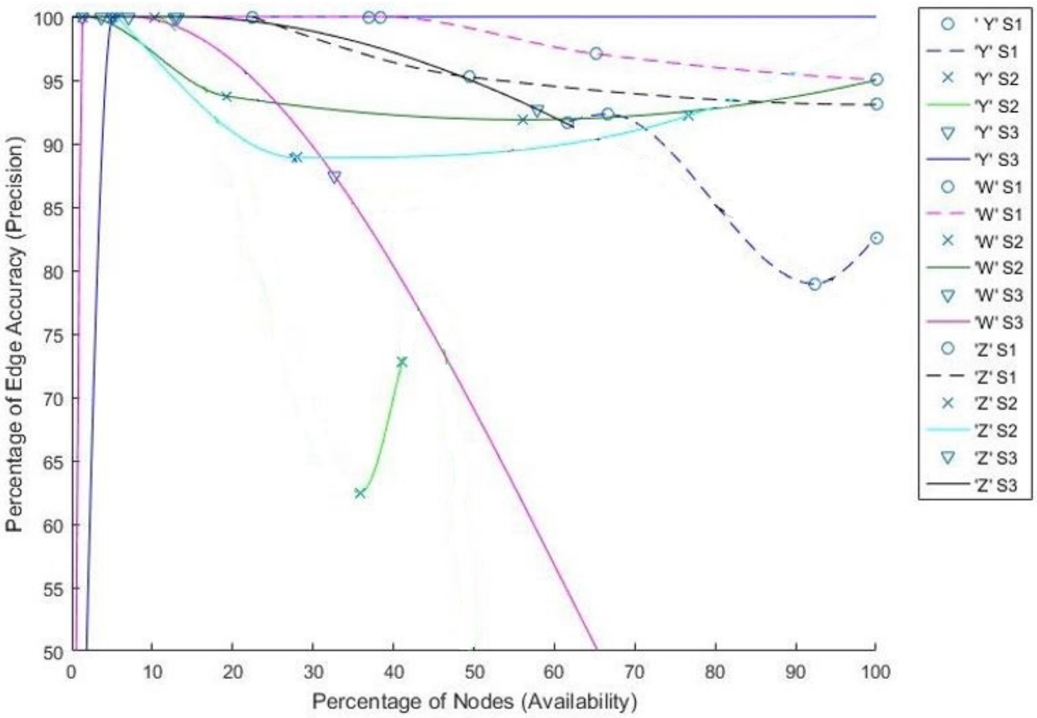


Figure 5.31: The L Grouping for the *W*, *Y* and *Z* datasets with the lower *t-val* T=0.25 plots to the right, increasing in value as sequence moves to the left

Figure 5.31 shows the L grouping for the *W*, *Y* and *Z* datasets (we have omitted the *X* dataset as noted). We see that S=1 performs well, with the exception of the *Y* dataset, which still ends up reaching BRD with a sufficiently high *t-val*. A notable difference with this graph is that plots of of S=3 and T=1.0 can lead to results of 0%.

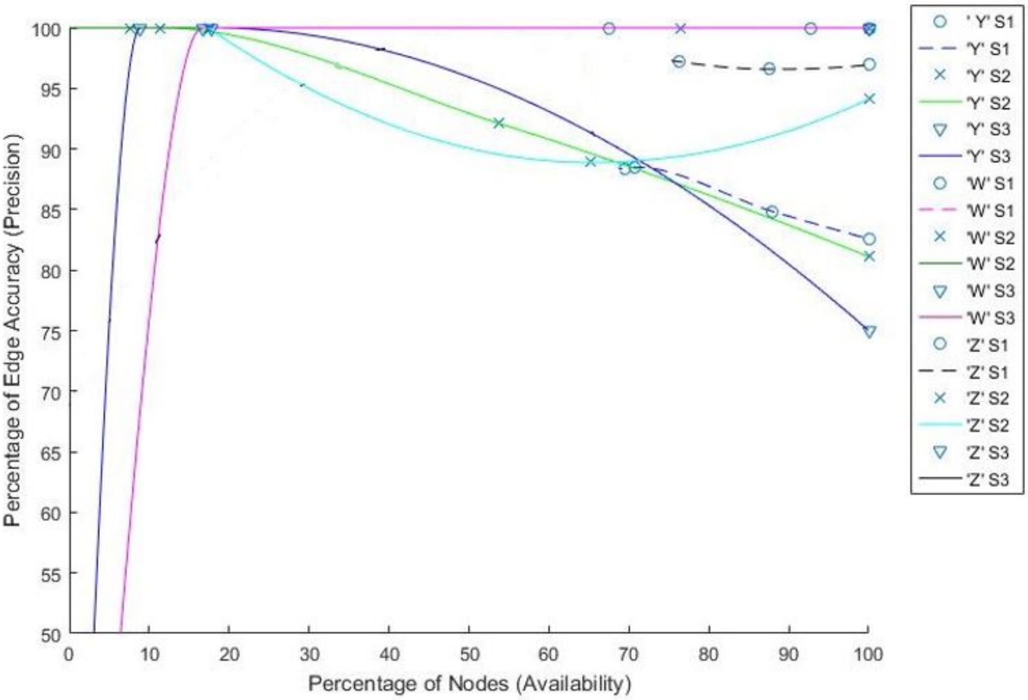


Figure 5.32: The S Grouping for the W, Y and Z datasets with the lower *t-val* T=0.25 plots to the right, increasing in value as sequence moves to the left

Figure 5.32 shows the S grouping for the W, Y and Z datasets (we have omitted the X dataset as noted). Similar to the L grouping there is a fall to 0% for the high *s-val* and *t-val* combinations.

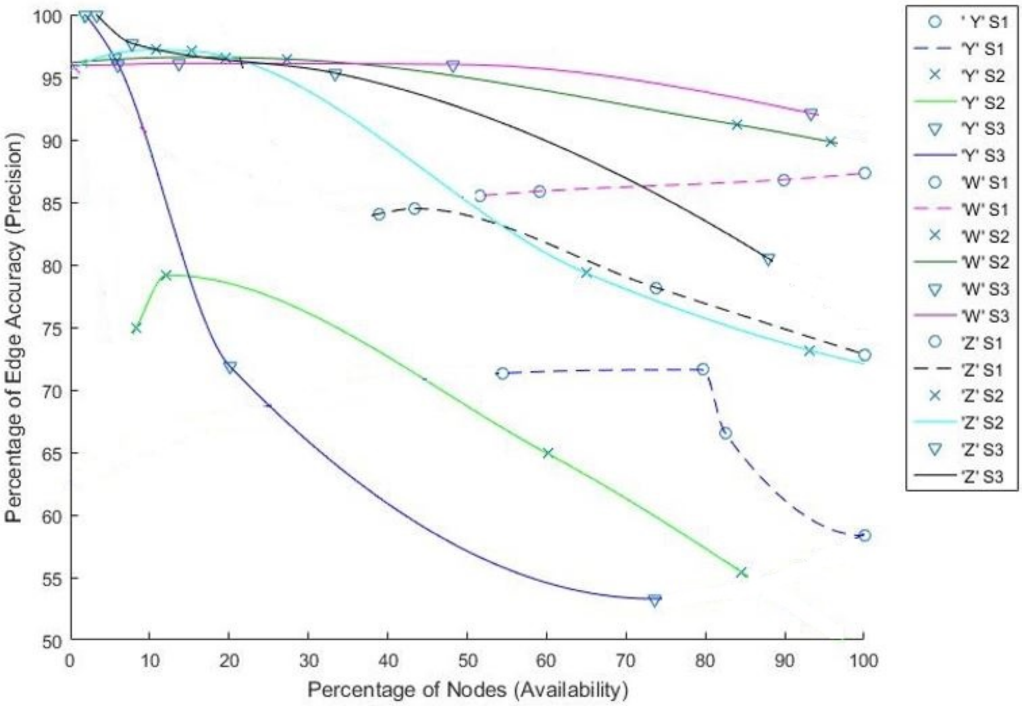


Figure 5.33: The Both Grouping for the *W*, *Y* and *Z* datasets with the lower *t-val* T=0.25 plots to the right, increasing in value as sequence moves to the left

Figure 5.33 shows the Both grouping for the *W*, *Y* and *Z* datasets (we have omitted the *X* dataset as noted). Much like the C grouping the S=1 sequence perform poorly and the S=2 and S=3 sequences ultimately reach above BRD levels, but they do so with a small percentage of the available nodes.

OVERALL COMMENTS

There is not one clear shape that can be observed with these 3 datasets. Four plots for each series is a small amount of data to provide a good best fit, however we see in figures 3.25 to 3.28 that there are relatively few points beyond 0.5 where we could take measurements, potentially 0.66 would provide an additional point, but this is not going to substantially change the shape of the graphs as they are generally in steady growth or fall.

We see that the results can be 100% accurate, with high *s-val* and *t-val* combinations, but this only produces a couple of percentages of the nodes that were initially available with the C and Both grouping. Given that the C and Both grouping do start with hundreds to a thousand nodes in some datasets, this can however still be a substantial amount of history.

There is an interesting shape that can be observed on the D grouping graphs, where one series of data ends, such as at the S=1 T=1.0 point, there is a gap in the number of available nodes before where the S=2 T=0.25 point begins. It shows that there is within the D grouping a interesting difference in the 'class' of S=1 grouping and the S=2 grouping etc. This is illustrated later in the figures 5.58 to 5.69, where we see certain clusters of sessions are available for the S=3, S=2, S=1 groupings, and whilst the session in the S=3 grouping are available in the S=1 grouping, the reverse is not true and there is a stark difference in the availability of the groupings.

Ultimately what these graphs show is that for any grouping and setting of the *t-val* and *s-val*, if we have sufficiently disentangled the Internet activity of the users in the dataset we will have a high accuracy and large number of sessions available. The underlying issue is therefore how similar the data is, and how similar the users are. If we have two individuals that have the same interests then our ability to differentiate between them is based on their behaviour. For example, User 1 like Site A and Site B and always goes to Site B during the same session as Site A, whereas User 2 also like Site A and Site B, but does not visit them during the same sessions. In this case manipulating the *s-val* could accurately discriminate between the two users (S=2 would find User 1's visiting both sites), but it is the behaviour of the users and presumably the content of those sites that dictates how effective our method is. We will discuss in the conclusions that the way forward at predicting and automating the selection of the correct *t-val* and *s-val* to get the best accuracy and number of nodes would be to have a concept of the closeness of similarity of the users.

5.4 GRAPHING THE RESULTS OF THE GROUPING AT THE 'BEYOND REASONABLE DOUBT' LEVEL

The nodes/sessions and the edges, the Jaccard similarity coefficients above the designated *t-val*, can be plotted onto a graph. We show here coloured graphs, where the colours of the nodes indicate automatically detected community groupings. The edges have been coloured green for correct matches and red for incorrect matches between nodes.

We have provided the D, E, S-only, L-only, C and Both graphs, for *s-val* and *t-val* that correctly match a BRD for the whole set of data. We have not included for simplicity the A and B graphs as these were either missing (in many cases the A group) or very low numbers (such as the B group).

5.4.1 *W* DATASET GRAPHS

Figure 5.34 shows the session-to-session graph for the *W* dataset using the D grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a 'Beyond Reasonable Doubt (BRD) result of 99.95%.

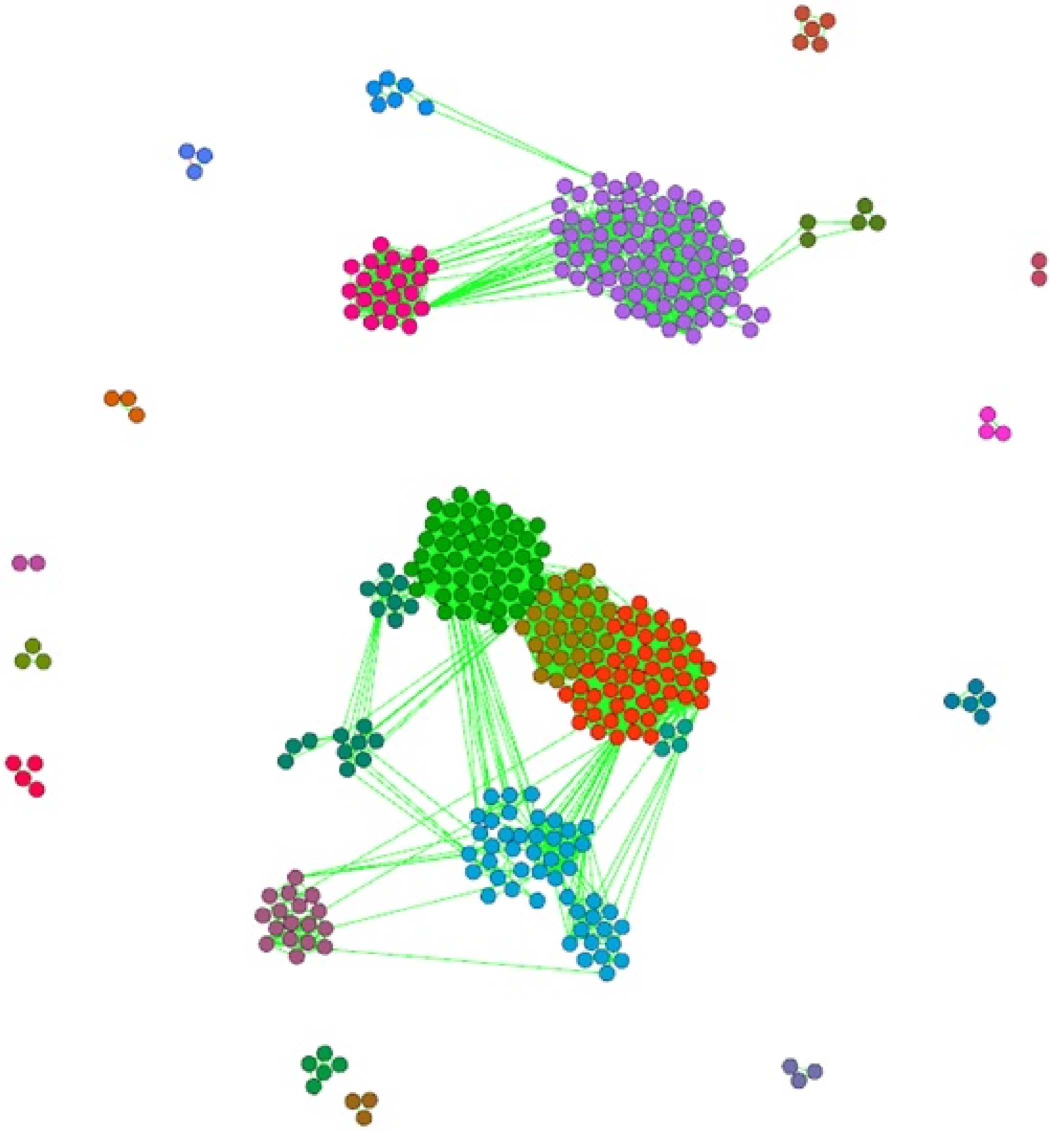


Figure 5.34: *W* dataset, D Groups, t=0.25, S=1

Figure 5.35 shows the session-to-session graph for the W dataset using the E grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

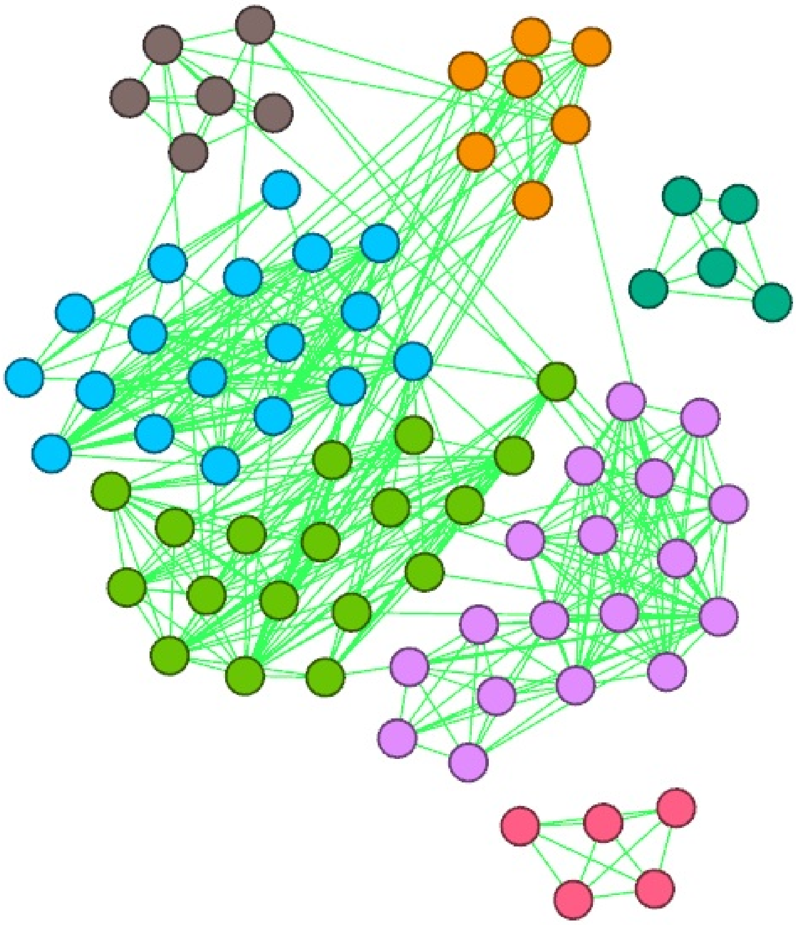


Figure 5.35: W dataset, E Group, t=0.25, S=1

Figure 5.36 shows the session-to-session graph for the *W* dataset using the S-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

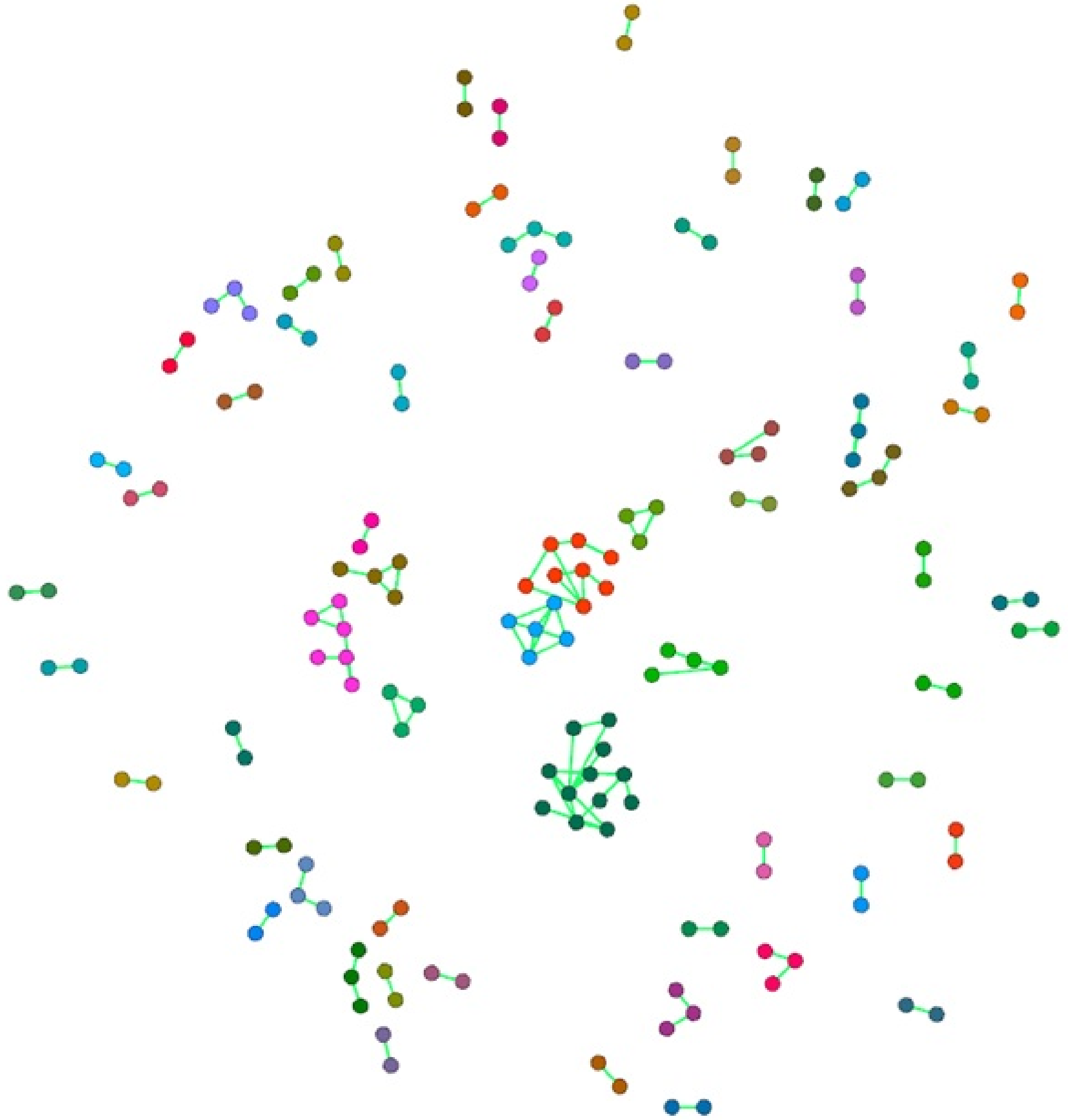


Figure 5.36: *W* dataset, S-only Group, t=0.25, S=1

Figure 5.37 shows the session-to-session graph for the *W* dataset using the L-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 95.04%.

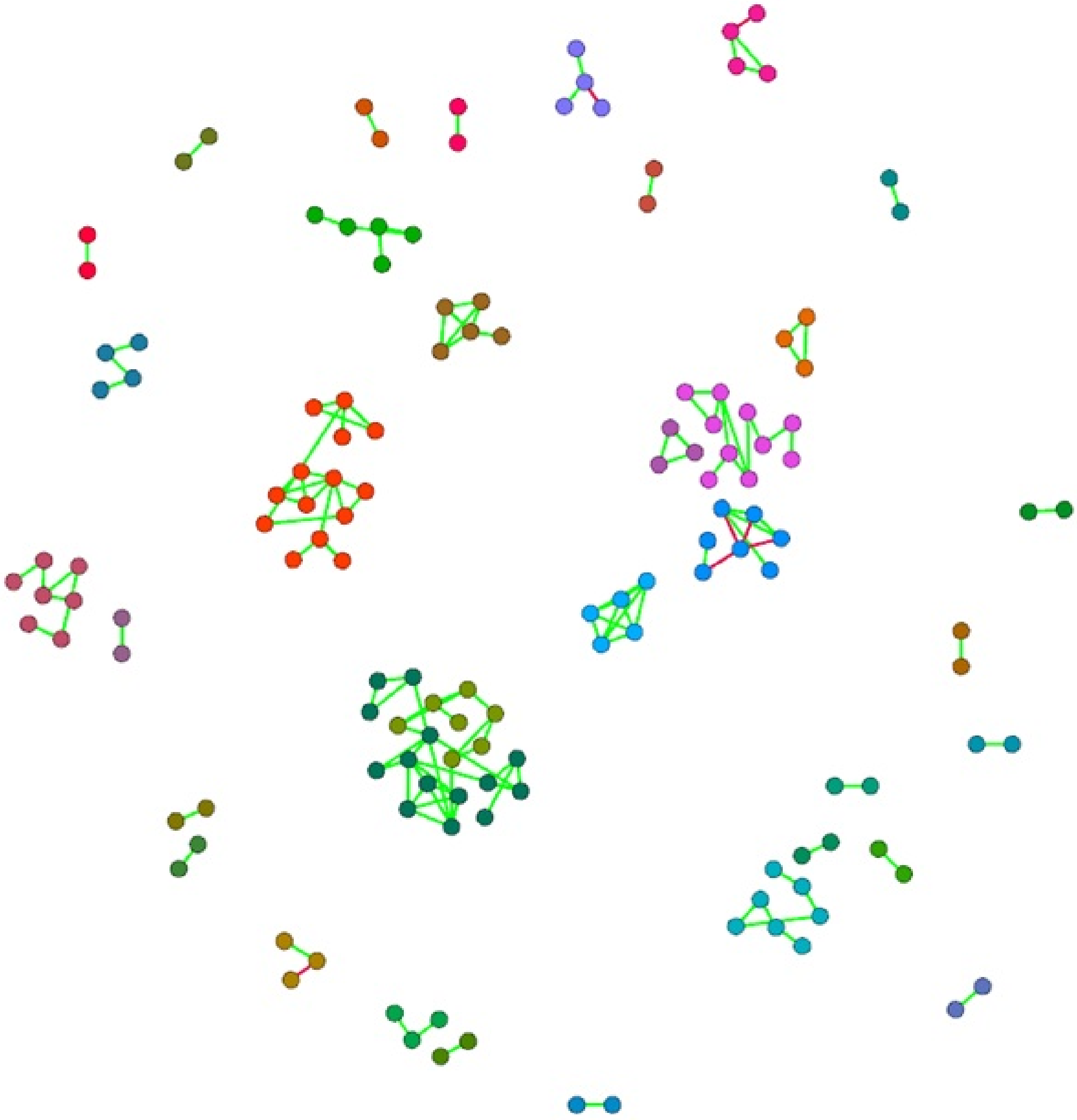


Figure 5.37: *W* dataset, L-only Group, t=0.25, S=1

Figure 5.38 shows the session-to-session graph for the *W* dataset using the C grouping. This graph has a *t-val* of T=0.5 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 91.94%.

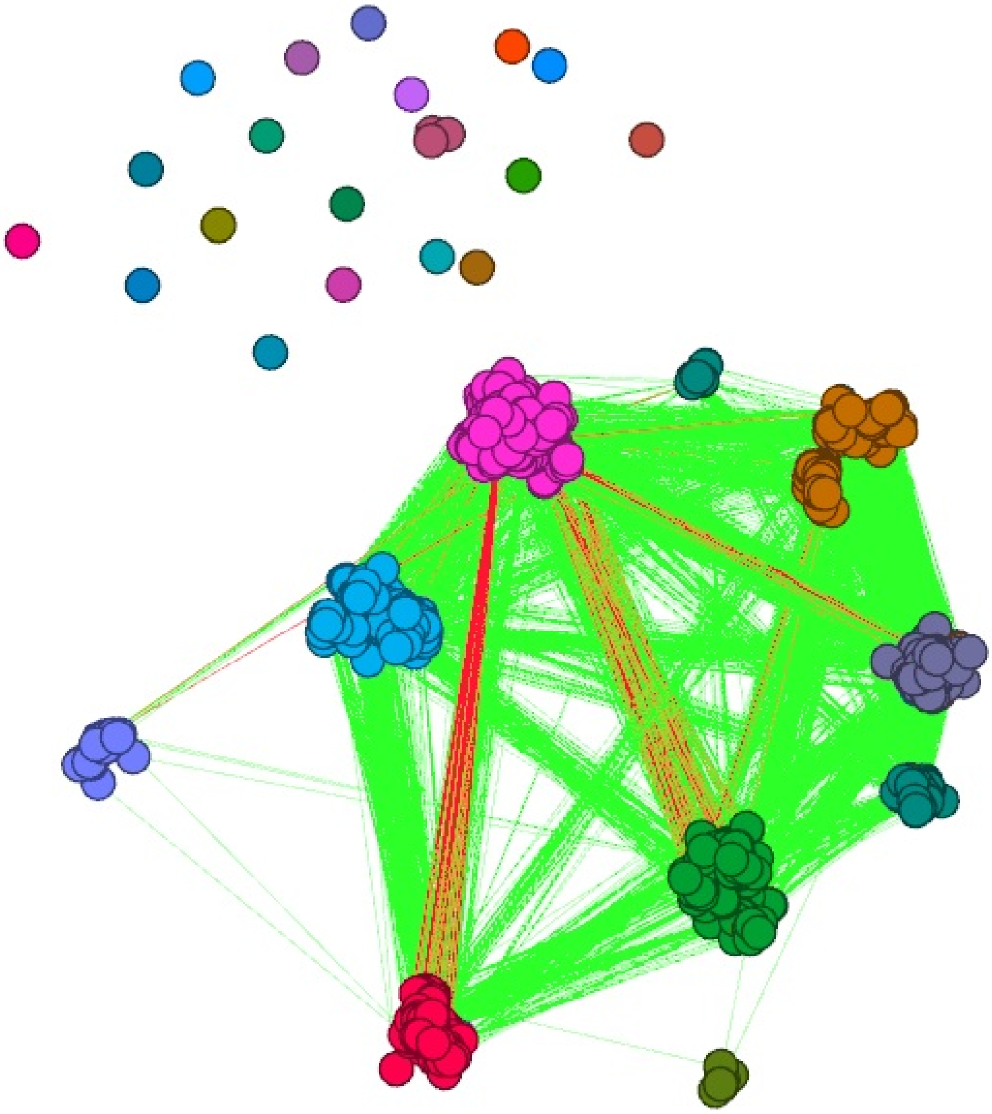


Figure 5.38: *W* dataset, C Group, t=0.5, S=2 n.b. this graph is extremely large (1143 sessions) and as such the communities have not been expanded

Figure 5.39 shows the session-to-session graph for the *W* dataset using the Both grouping. This graph has a *t-val* of T=0.5 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 91.19%.

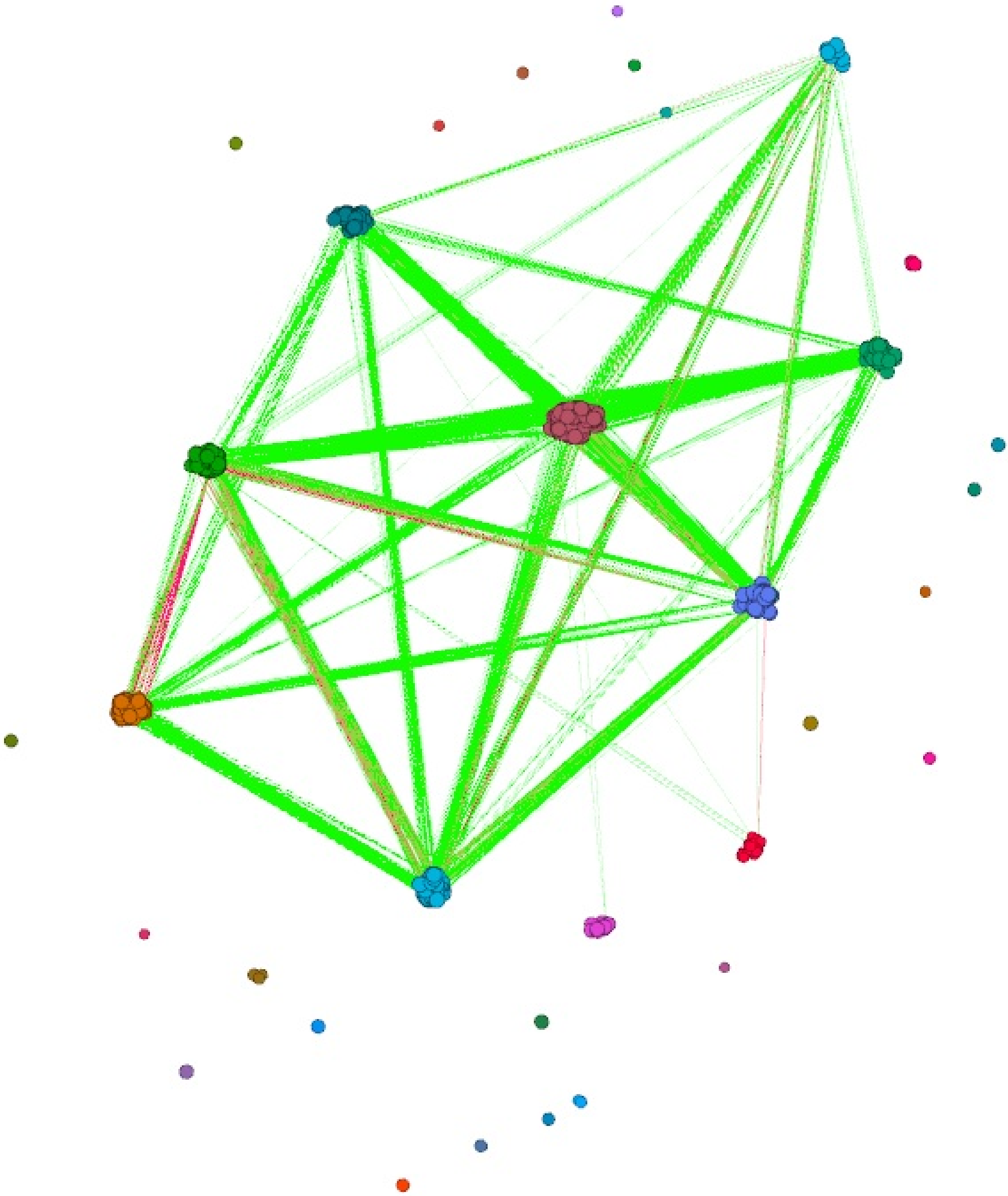


Figure 5.39: *W* dataset, Both Group, t=0.5, S=2 n.b. this graph is extremely large (1154 sessions) and as such the communities have not been expanded

5.4.2 *Y* DATASET GRAPHS

Figure 5.40 shows the session-to-session graph for the *Y* dataset using the D grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a ‘Beyond Reasonable Doubt (BRD) result of 91.14%.

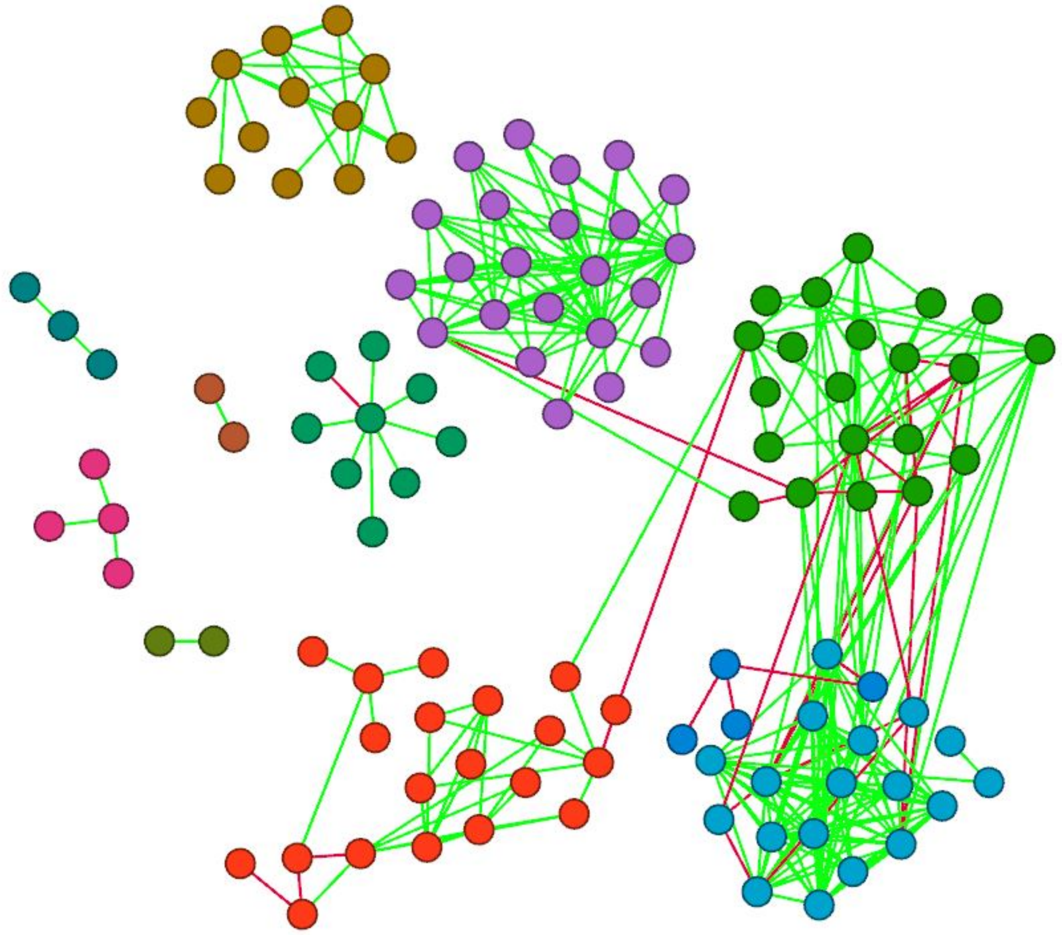


Figure 5.40: *Y* dataset, D Groups, t=0.5, S=2

Figure 5.41 shows the session-to-session graph for the Y dataset using the E grouping. This graph has a *t-val* of T=0.75 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

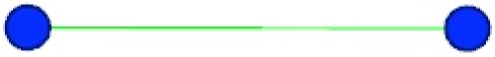


Figure 5.41: Y dataset, E Group, t=0.75, S=2

Figure 5.42 shows the session-to-session graph for the Y dataset using the S-only grouping. This graph has a *t-val* of T=0.5 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 92.11%.

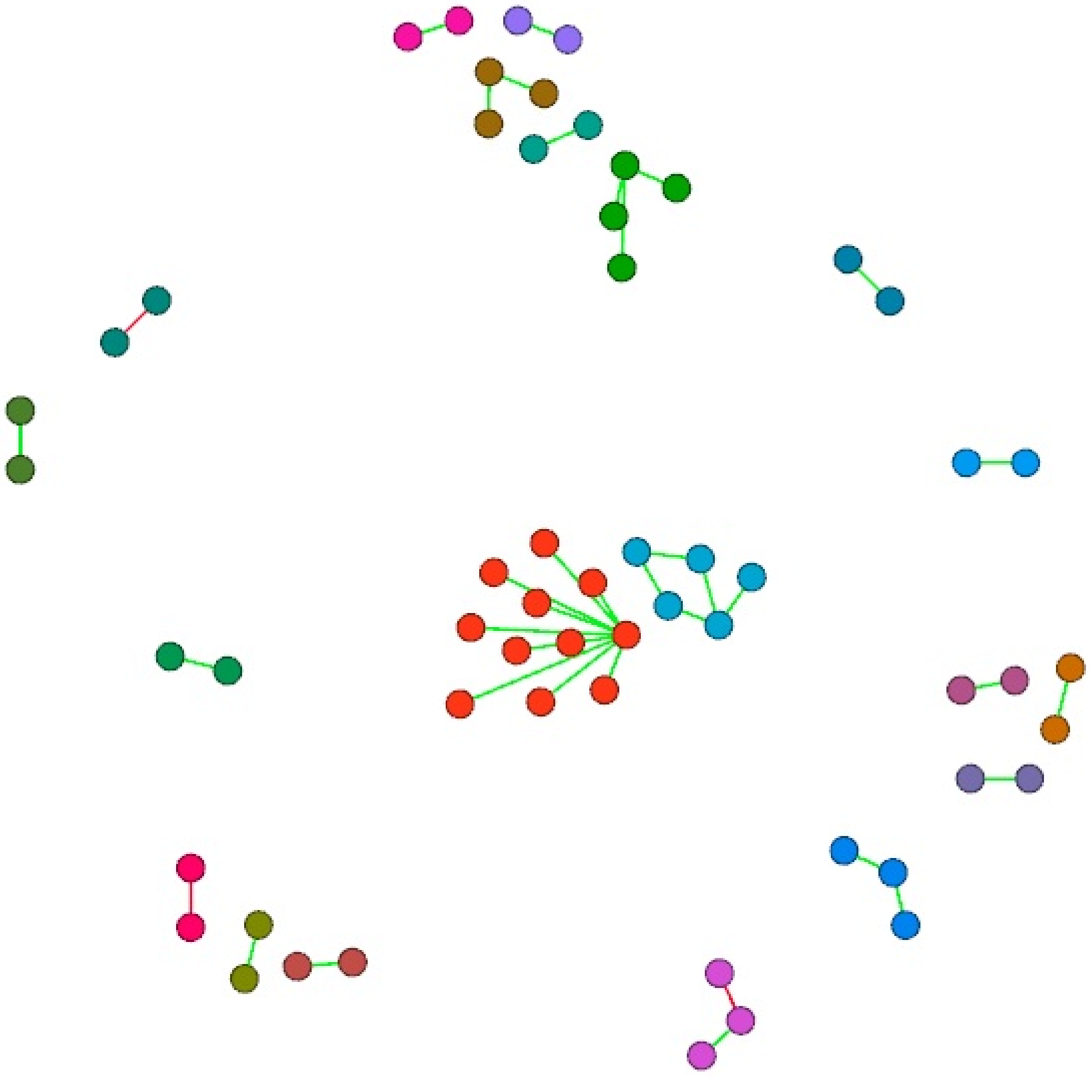


Figure 5.42: Y dataset, S-only Group, t=0.5, S=2

Figure 5.43 shows the session-to-session graph for the Y dataset using the L-only grouping. This graph has a *t-val* of T=0.75 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

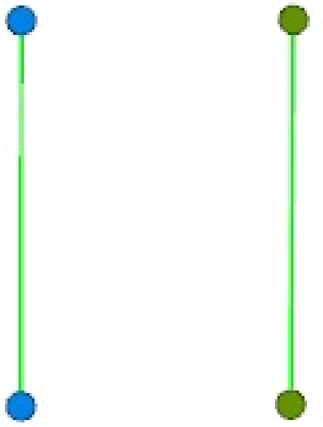


Figure 5.43: Y dataset, L-only Group, t=0.75, S=2

Figure 5.44 shows the session-to-session graph for the Y dataset using the C grouping. This graph has a *t-val* of T=1.0 and an *s-val* of S=3, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

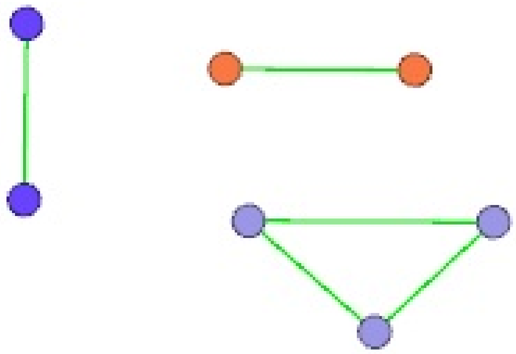


Figure 5.44: Y dataset, C Group, t=1.0, S=3

Figure 5.45 shows the session-to-session graph for the Y dataset using the Both grouping. This graph has a *t-val* of T=0.75 and an *s-val* of S=3, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

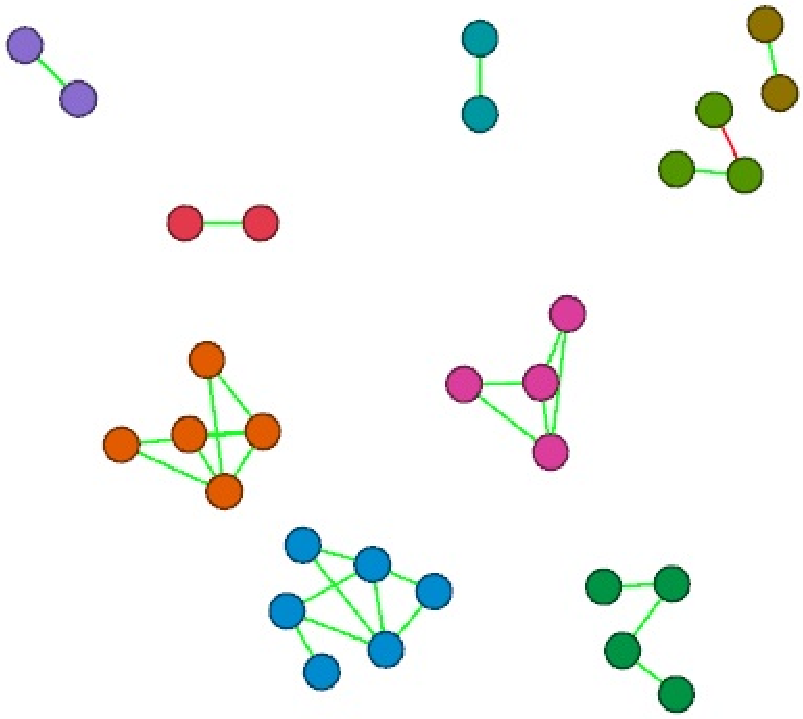


Figure 5.45: Y dataset, Both Group, t=0.75, S=3

5.4.3 *Z* DATASET GRAPHS

Figure 5.46 shows the session-to-session graph for the *Z* dataset using the D grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a ‘Beyond Reasonable Doubt (BRD) result of 93.14%.

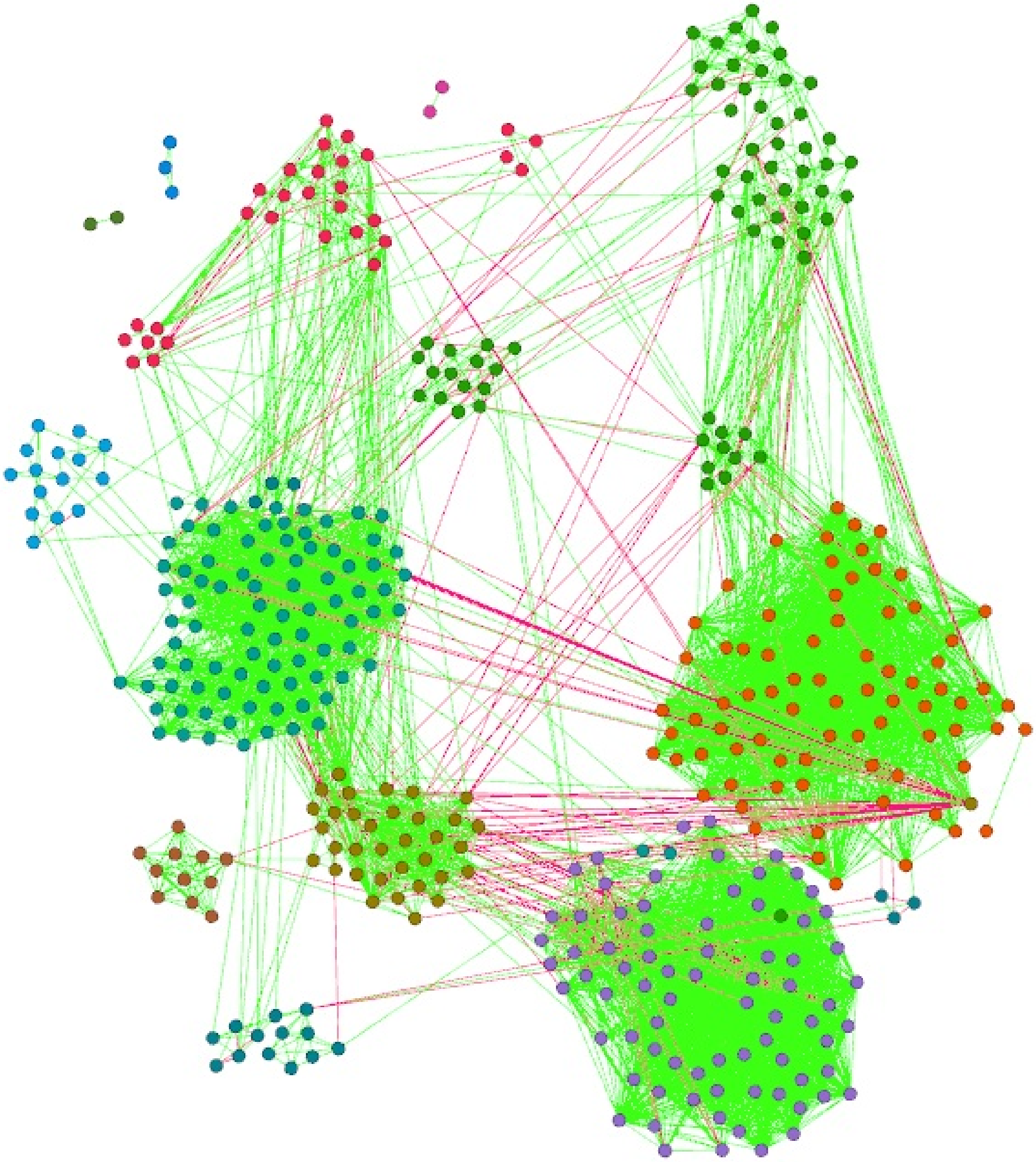


Figure 5.46: *Z* dataset, D Groups, t=0.25, S=1

Figure 5.47 shows the session-to-session graph for the Z dataset using the E grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 98.31%.

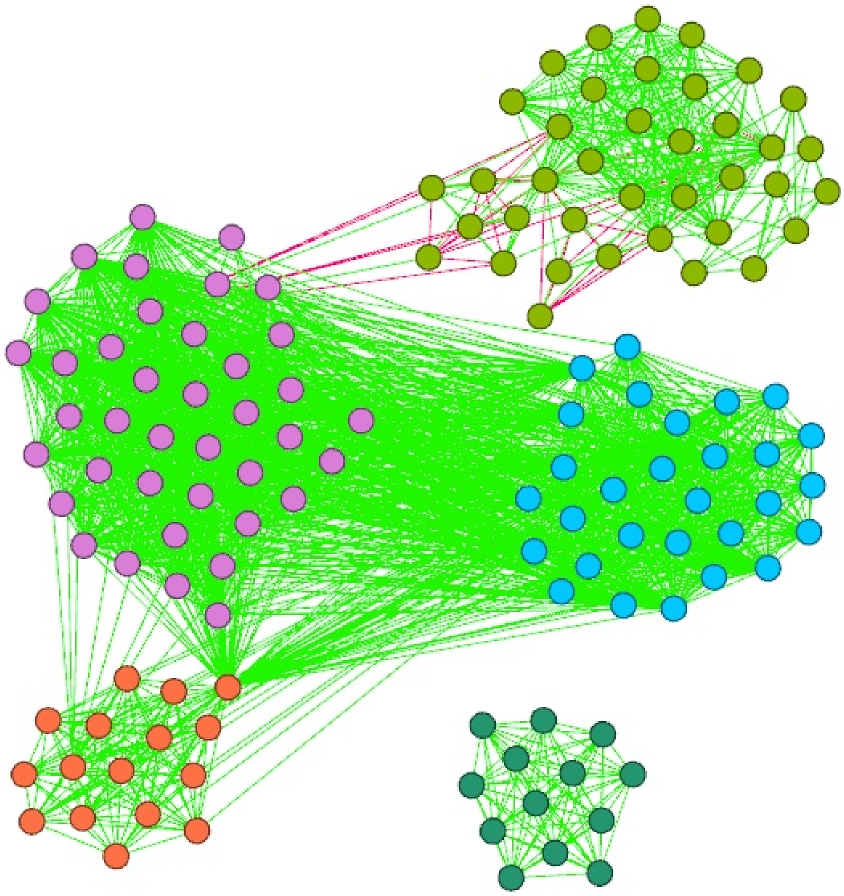


Figure 5.47: Z dataset, E Group, t=0.25, S=1

Figure 5.48 shows the session-to-session graph for the Z dataset using the S-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 96.95%.

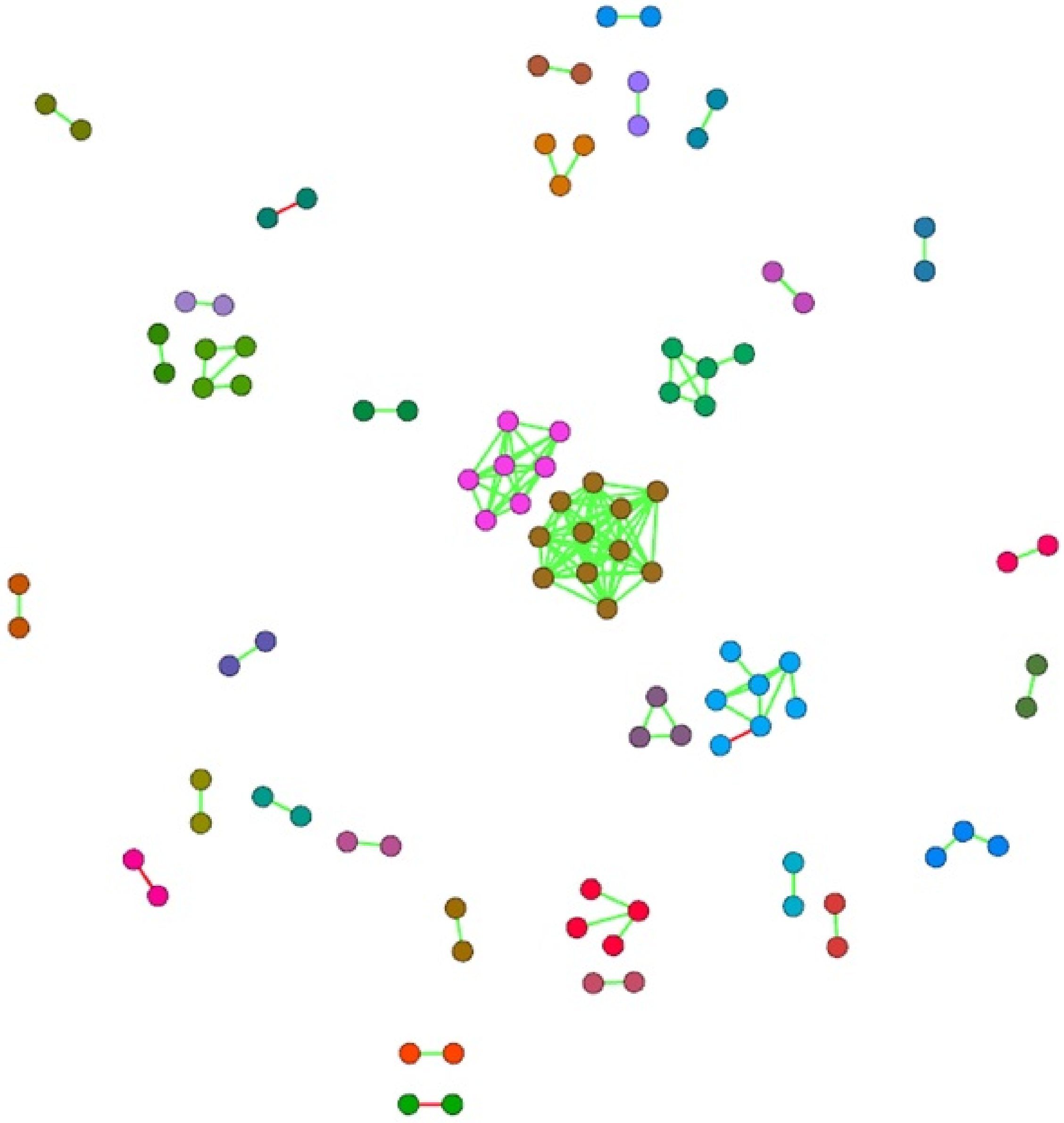


Figure 5.48: Z dataset, S-only Group, t=0.25, S=1

Figure 5.49 shows the session-to-session graph for the Z dataset using the L-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 93.07%.

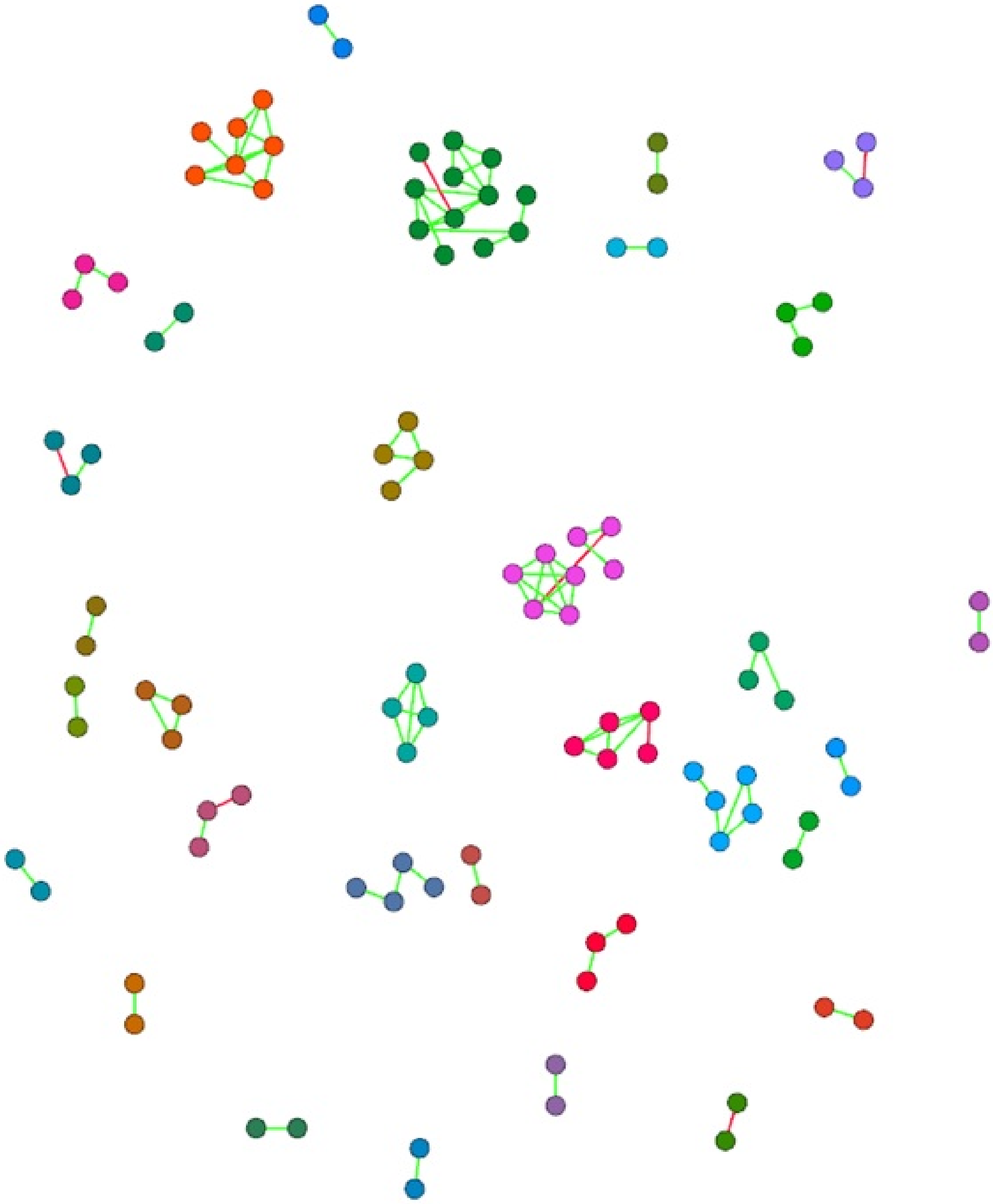


Figure 5.49: Z dataset, L-only Group, t=0.25, S=1

Figure 5.50 shows the session-to-session graph for the Z dataset using the C grouping. This graph has a *t-val* of T=0.75 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 93.64%.

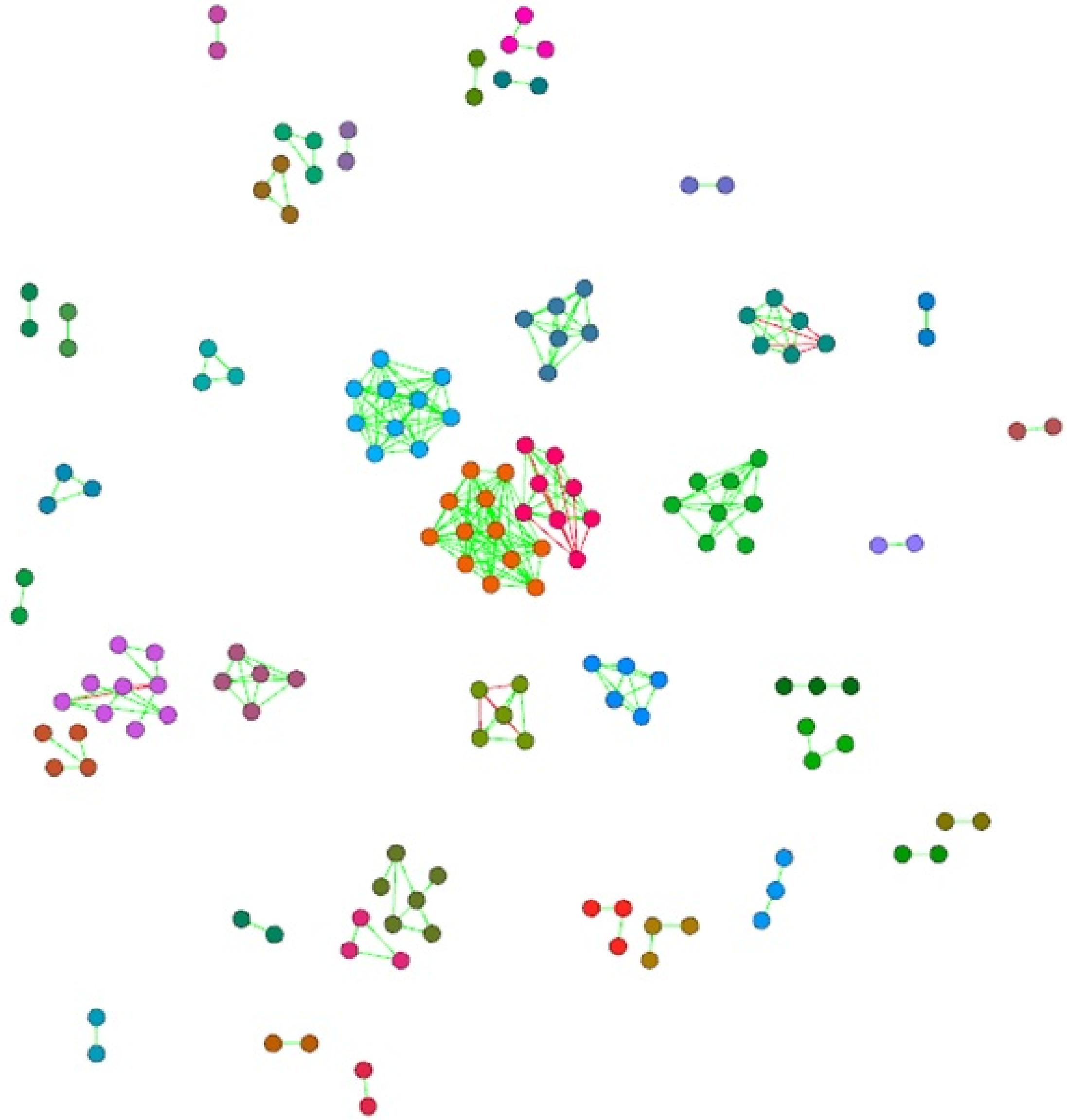


Figure 5.50: Z dataset, C Group, t=0.75, S=2

Figure 5.51 shows the session-to-session graph for the *Z* dataset using the Both grouping. This graph has a *t-val* of T=0.75 and an *s-val* of S=2, which is the lowest *t-val* and *s-val* to produce a BRD result of 97.08%.

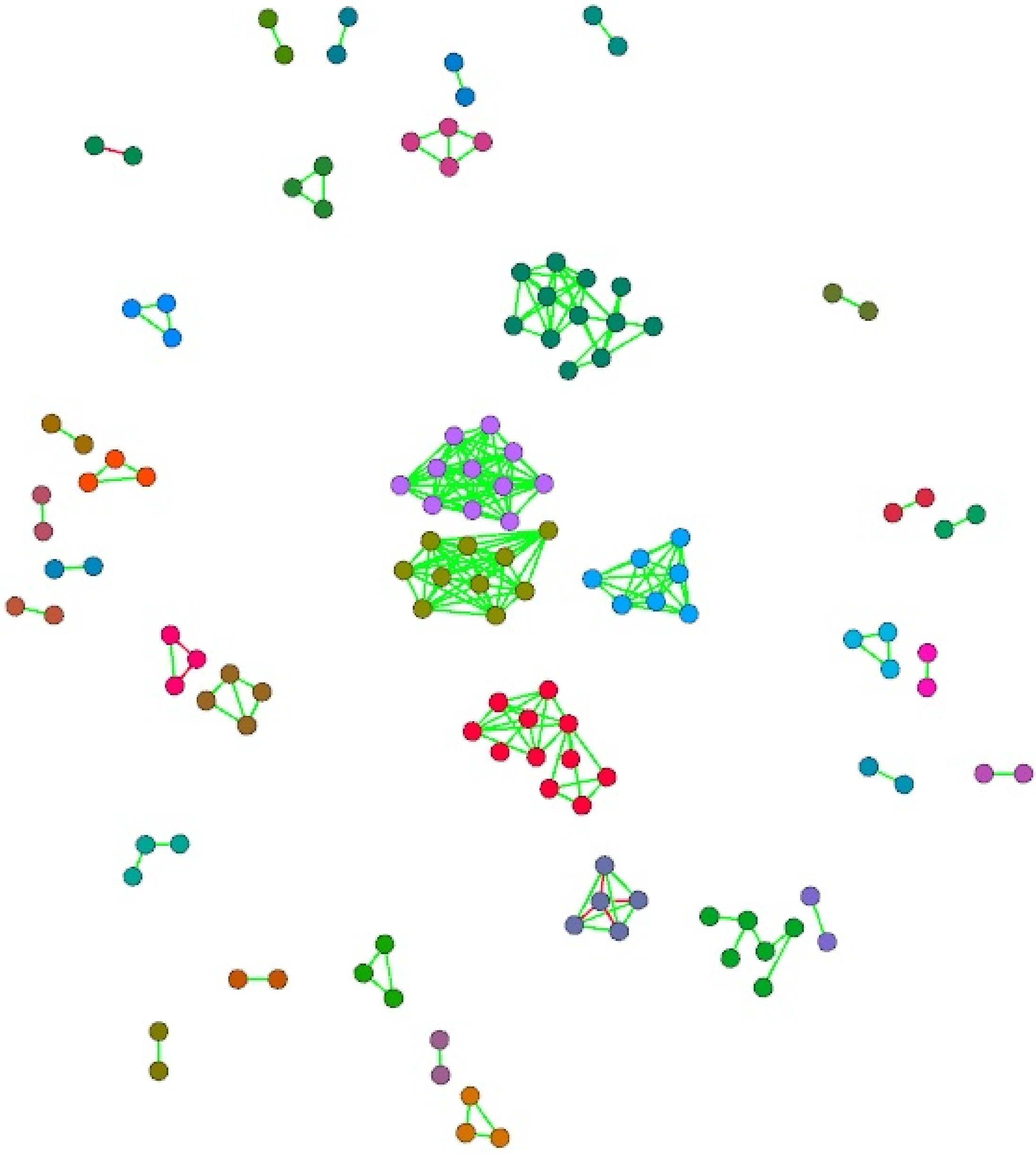


Figure 5.51: *Z* dataset, Both Group, t=0.75, S=2

5.4.4 *X* DATASET GRAPHS

Figure 5.52 shows the session-to-session graph for the *X* dataset using the D grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a ‘Beyond Reasonable Doubt (BRD) result of 100%.

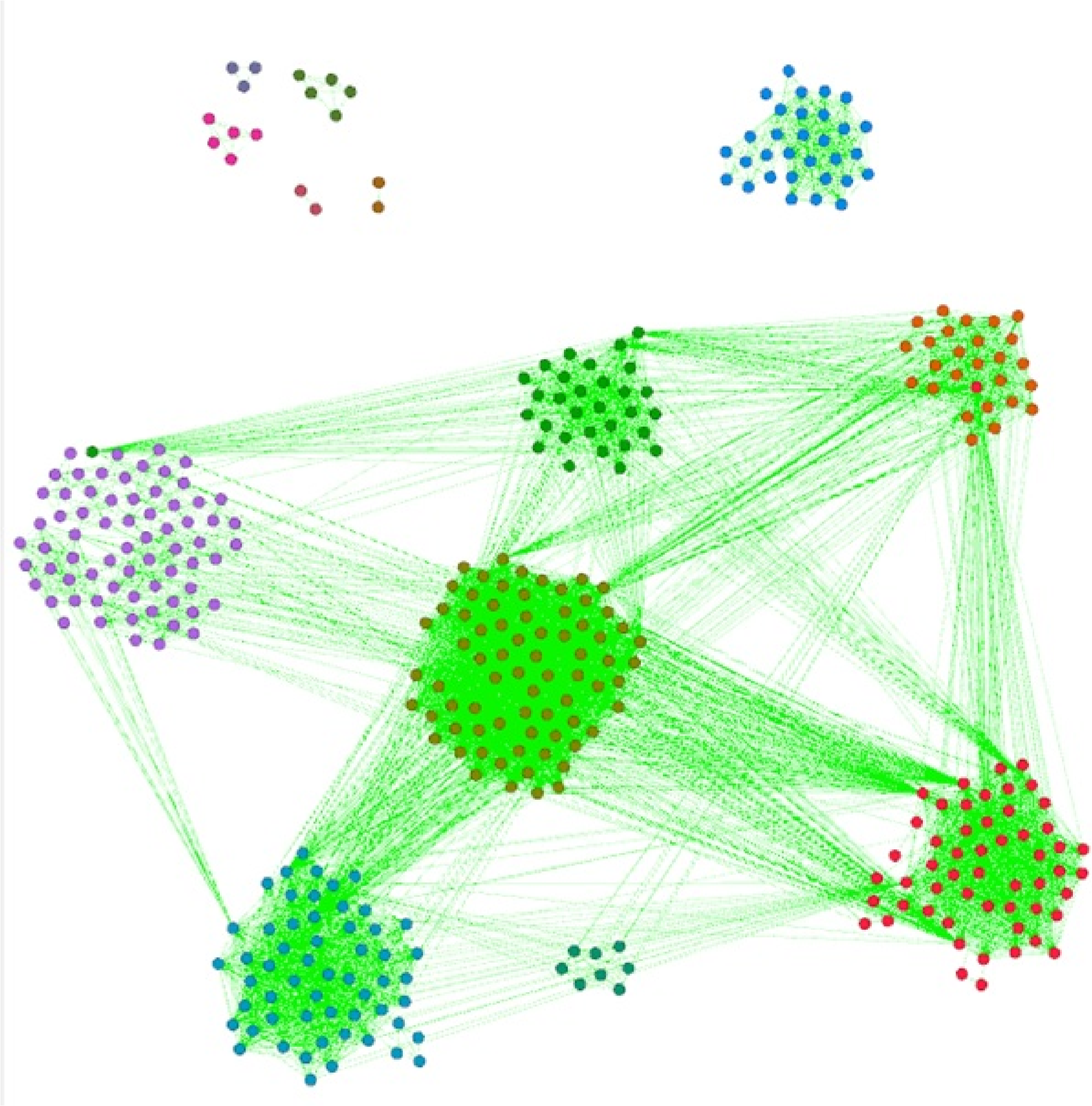


Figure 5.52: *X* dataset, D Group, t=0.25, S=1

Figure 5.53 shows the session-to-session graph for the X dataset using the E grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

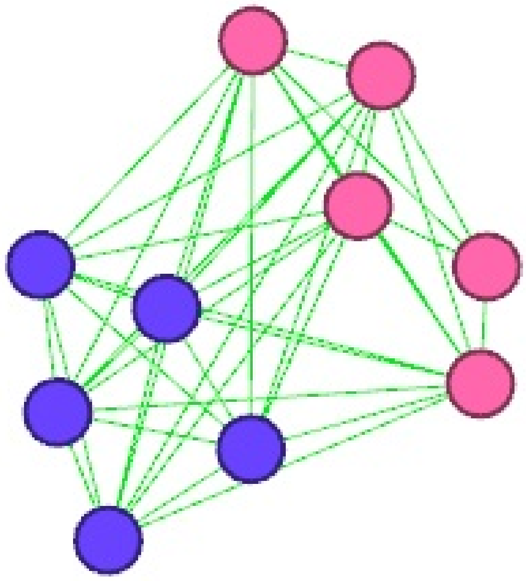


Figure 5.53: X dataset, E Group, t=0.25, S=1

This is an unusual situation where 10 seemingly interconnected nodes have been segregated into 2 communities, by the community detection algorithm. Functionally, as there is zero error there is no effect by this.

Figure 5.54 shows the session-to-session graph for the *X* dataset using the S-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 95.45%.

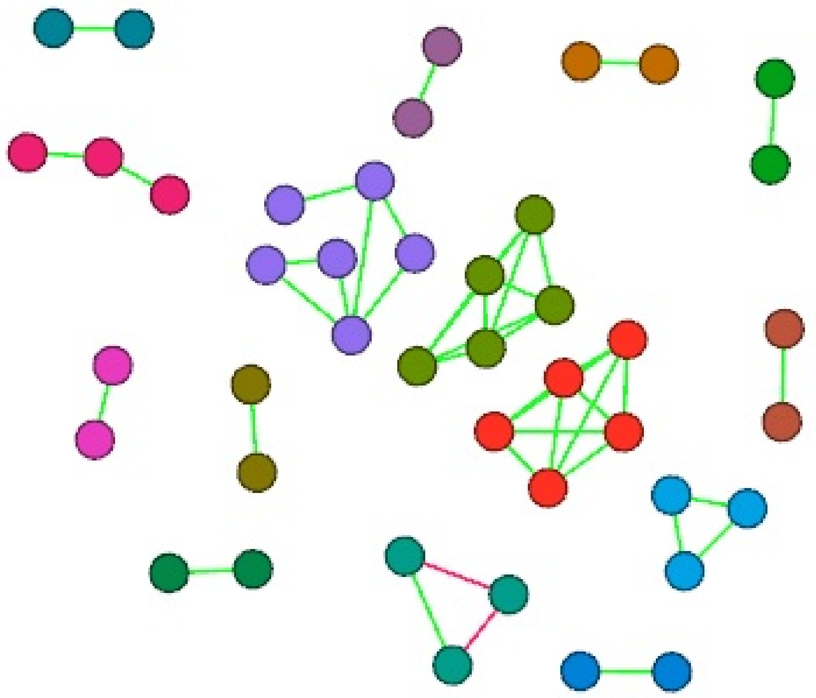


Figure 5.54: *X* dataset, S-only Group, t=0.25, S=1

Figure 5.55 shows the session-to-session graph for the X dataset using the L-only grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 100%.

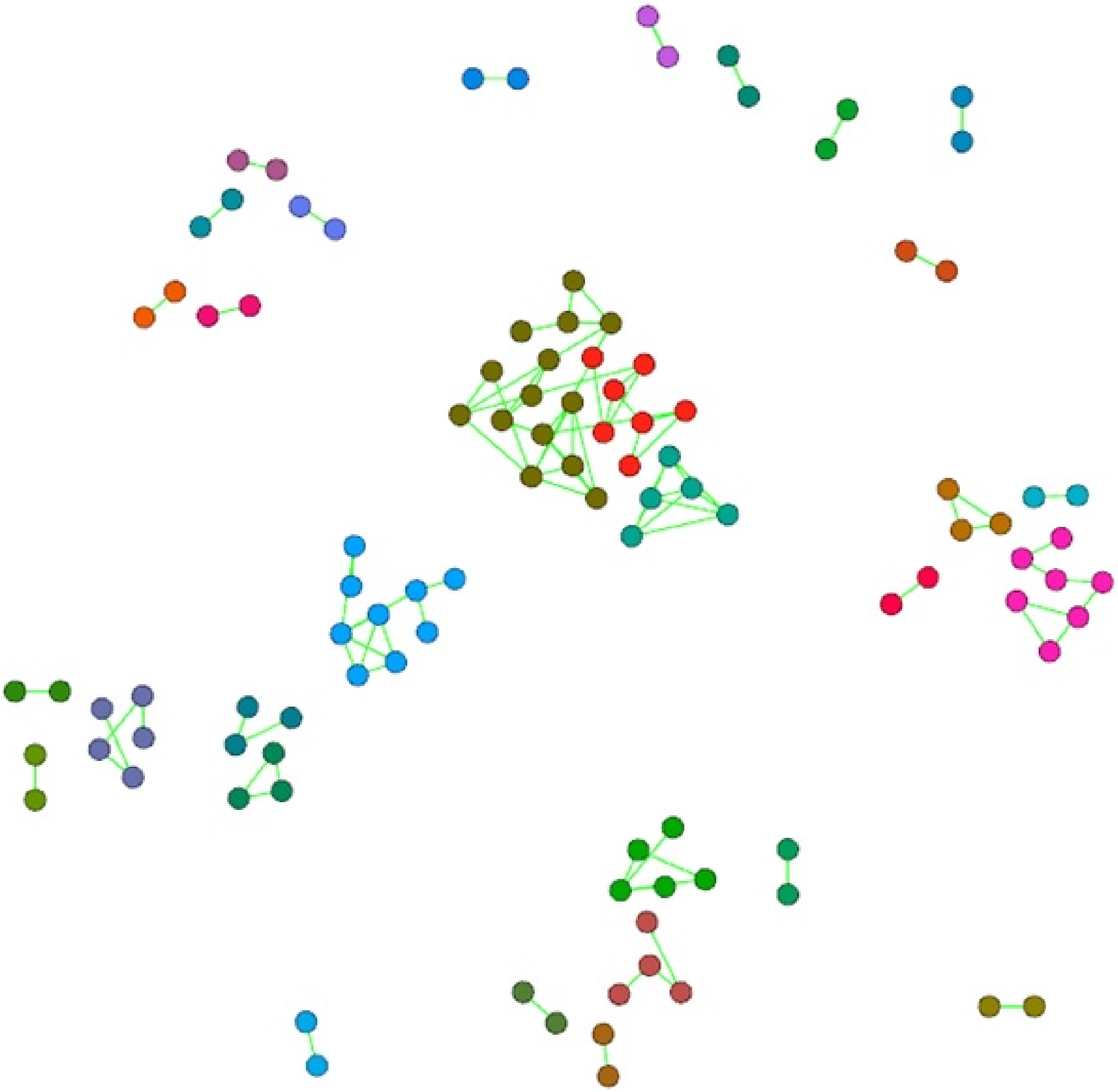


Figure 5.55: X dataset, L-only, t=0.25, S=1

Figure 5.56 shows the session-to-session graph for the X dataset using the C grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 96.39%.

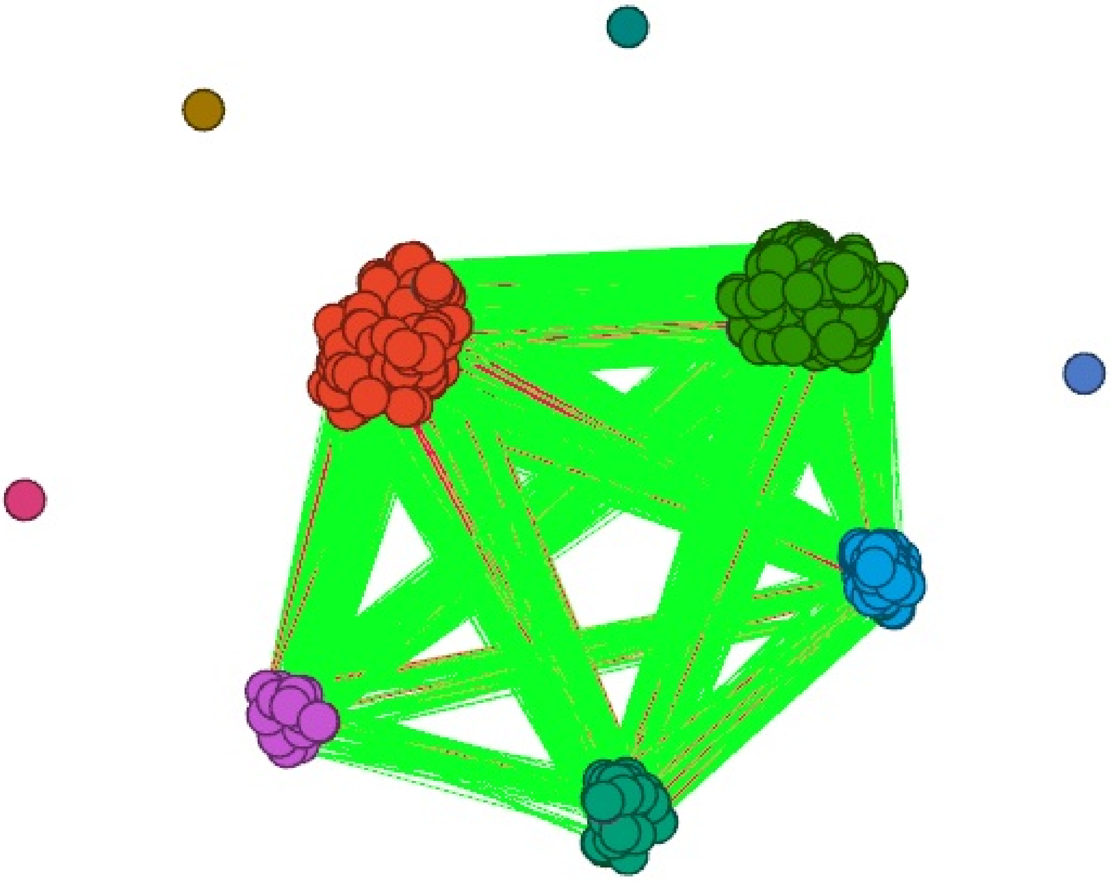


Figure 5.56: X dataset, C Group, t=0.25, S=1 n.b. this graph is extremely large (1098 sessions) and as such the communities have not been expanded

Figure 5.57 shows the session-to-session graph for the X dataset using the Both grouping. This graph has a *t-val* of T=0.25 and an *s-val* of S=1, which is the lowest *t-val* and *s-val* to produce a BRD result of 96.71%.

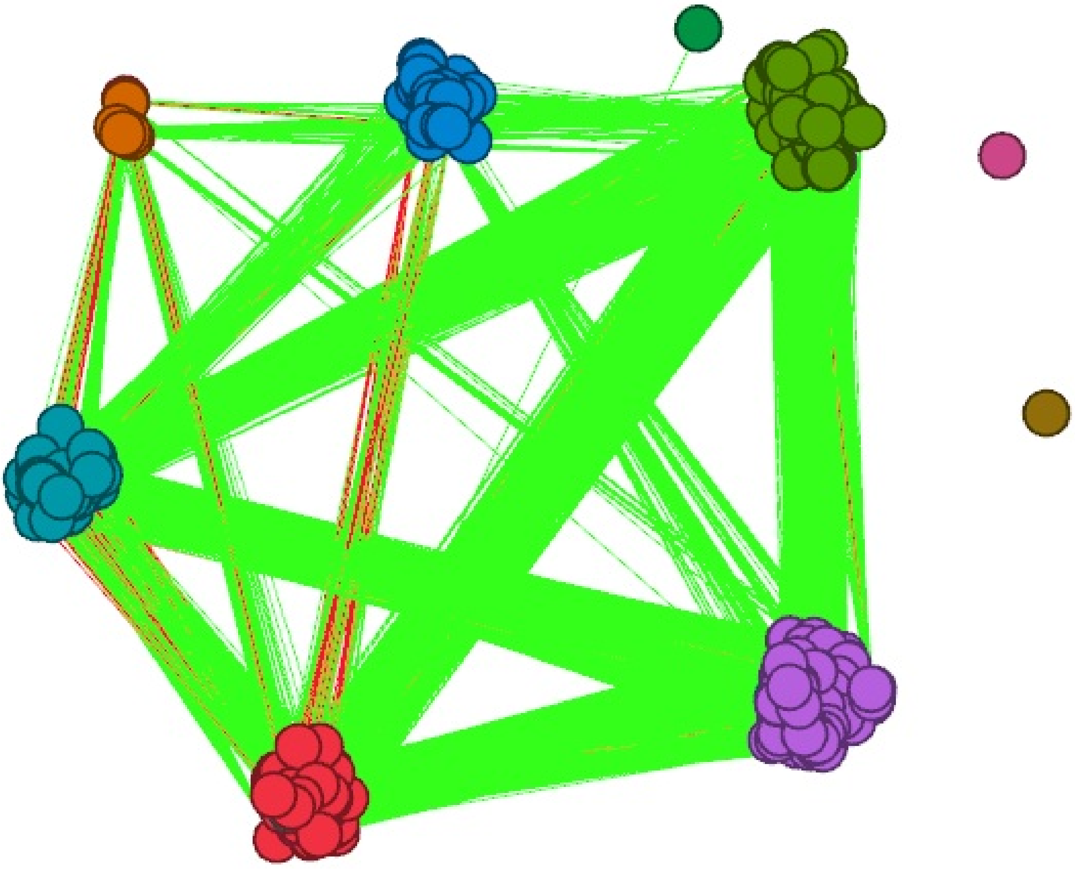


Figure 5.57: X dataset, Both Group, t=0.25, S=1 n.b. this graph is extremely large (1133 sessions) and as such the communities have not been expanded

5.4.5 OVERALL COMMENTS ON THE GRAPHS

The S-only and L-only graphs have similar performance to the D and E graphs in the sense that they reach the BRD level with the same *s-val* and *t-val*, however we can see the appearance of the graphs is quite different. S-only and L-only graphs tend to have a much higher number of communities containing smaller numbers of nodes per community. The D and E graphs have fewer communities and a much higher degree of interconnectedness between the communities generally.

We cannot necessarily predict the number of nodes and communities that will be available using different grouping schemes. For example, the Z dataset (at S=1 and *t-val* 0.25) has 427 nodes in the D group, 135 nodes in the E group, 97 nodes in the S-only group and 107 nodes in the L-only group). The Y dataset (at S=1 and *t-val* 0.25) has 185 nodes in the D group, 82 nodes in the E group, 324 nodes in the S-only group and 39 nodes in the L-only group). In one case the S-only grouping produced much more than the D grouping, and in the other case the reverse is true.

We therefore cannot say that X grouping scheme will produce Y numbers of nodes as this is far too dependent on the Internet history used as input. We can predict that the S-only and L-only scheme will produce a higher number of pairwise communities and an overall greater number of communities than the D and E groupings.

Practically this means that the D and E schemes are likely to better answer questions if a particular session is 'regular behaviour' rather than a 'one-off event' as it will be part of a larger interconnected group. S-only and L-only grouping schemes are likely to provide good reliability if we are trying to say session X and session Y belong to the same user, but it is probabilistically less likely that any two sessions in the dataset will be interconnected with those types of grouping schemes, than with the D and E groupings.

We note that the C and Both grouping schemes produce somewhat similar results, with similar numbers of nodes, communities and accuracies.

5.5 ILLUSTRATING THE EFFECT OF THE *S-VAL* AND *T-VAL* ON THE NETWORK GRAPHS

Although we discuss earlier the issue of selecting the correct *s-val* and *t-val*, we illustrate here the effect of dialing different *s-val* and *t-val*s. We propose that if analysts are attempting to show the relationship between two or more sessions, they can initially select a high *s-val* and *t-val* and then 'walk back' the variables until either a connection has been made between the two (or more sessions), or that the setting of the *s-val* or *t-val* is so low that the analyst has insufficient confidence that the produced graph is accurate 'beyond reasonable doubt'.

We show here as an example the Y dataset plotted for comparison, with all the nodes/sessions coloured grey for each data grouping. As above the edges between the nodes are green if they represent a correct match between two sessions belonging to the same user, and red if they represent an incorrect match.

In figure 5.58 we see the highest level of filtering where the *s-val* is S=3, i.e. there must be three or more matching components and the *t-val* is T=1.0, i.e. there must be an exact match between the components in the matching sessions. We see that in the figure there are only 2 sessions coloured black which meet this criteria and all the remaining sessions are unavailable.

As we 'dial down' the *t-val* settings from T=1.0 to T=0.25 in figures 5.59 to 5.61 we see more and more sessions turn black which means they become available for analysis at the quite stringent S=3 level.

For the 'D' Group of data, the sessions coloured black are present and available in the analysis at that *s-val* and *t-val* setting:

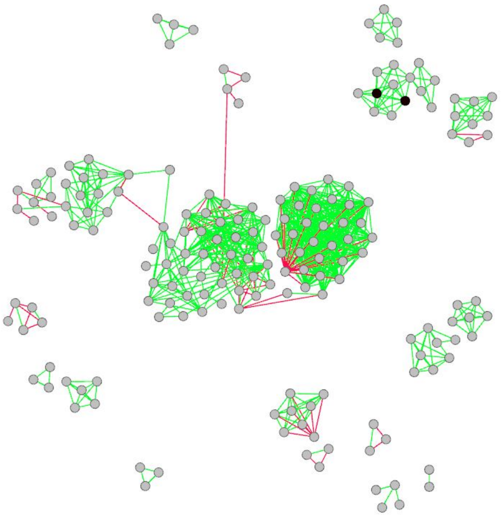


Figure 5.58: S=3 T=1

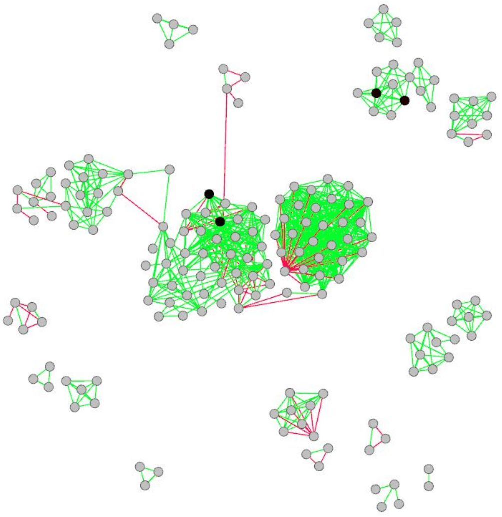


Figure 5.59: S=3 T=0.75

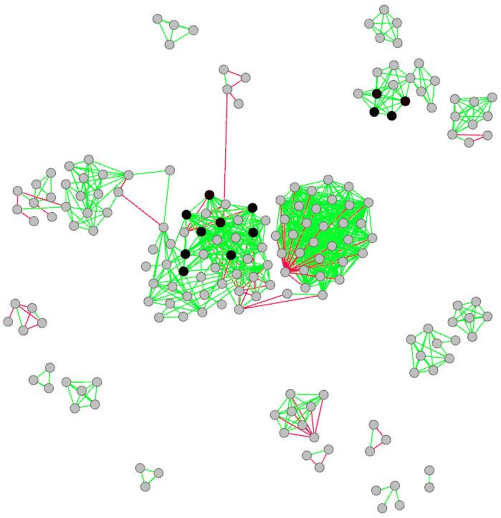


Figure 5.60: S=3 T=0.5

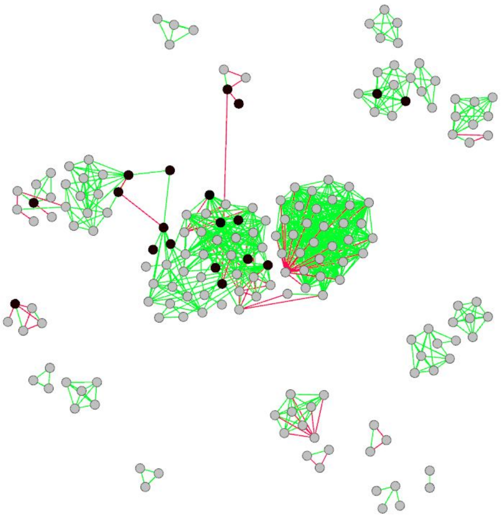


Figure 5.61: S=3 T=0.25

For the ‘D’ Group of data, the sessions coloured black and blue are present and available in the analysis at that *s-val* and *t-val* setting:

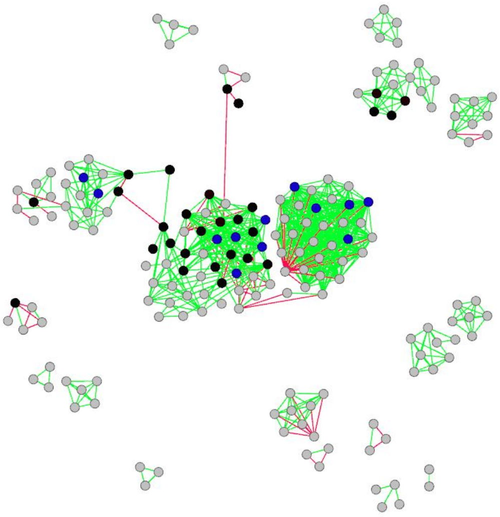


Figure 5.62: S=2 T=1

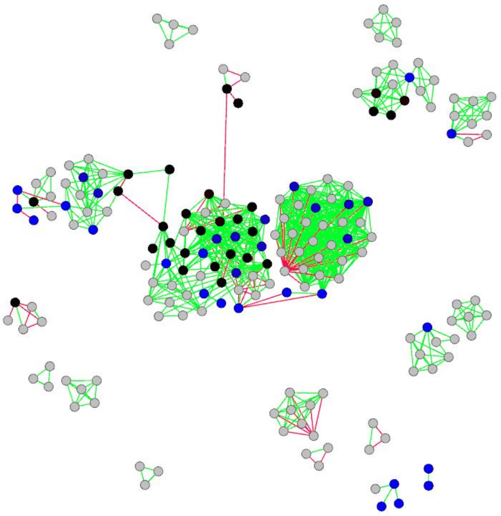


Figure 5.63: S=2 T=0.75

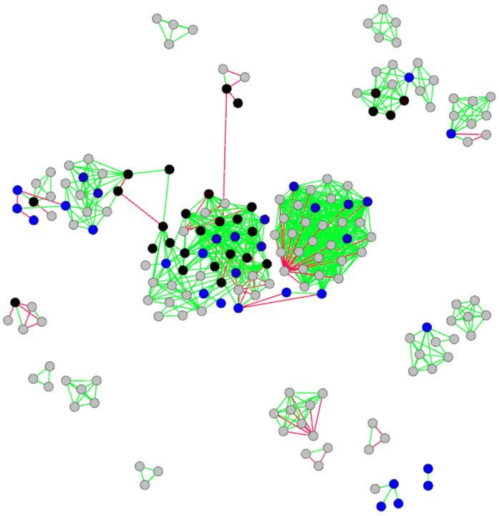


Figure 5.64: S=2 T=0.5

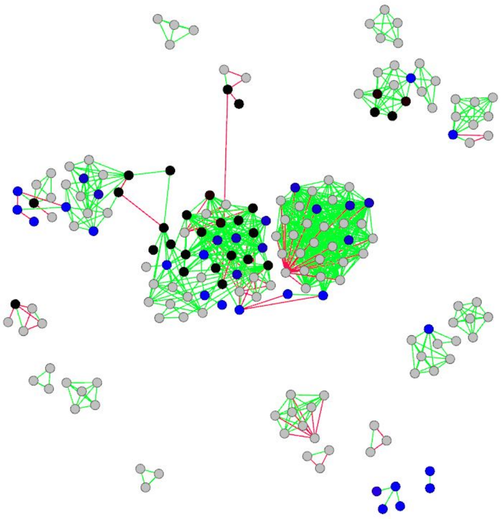


Figure 5.65: S=3 T=0.25

In figures 5.62 through 5.65 we see the same process as done in figures 5.58 through 5.61, but at the lower *s-val* of S=2. It is worth noting that the black coloured nodes are still available during this process, but the addition of the blue coloured nodes represent the S=2 sessions.

For the ‘D’ Group of data, the sessions coloured black, blue and pink are present and available in the analysis at that *s-val* and *t-val* setting:

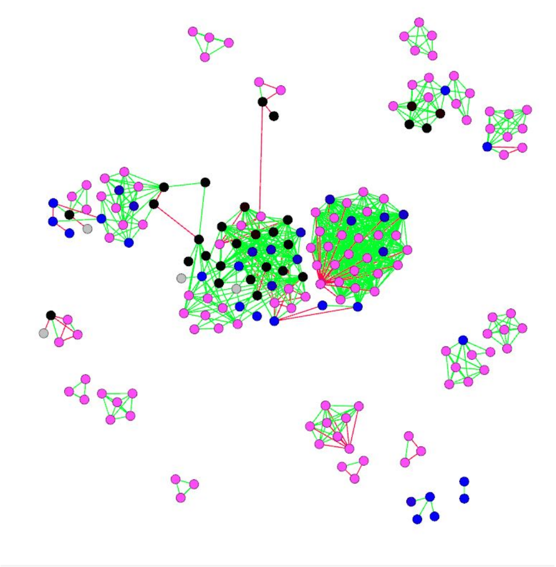


Figure 5.66: S=1 T=1

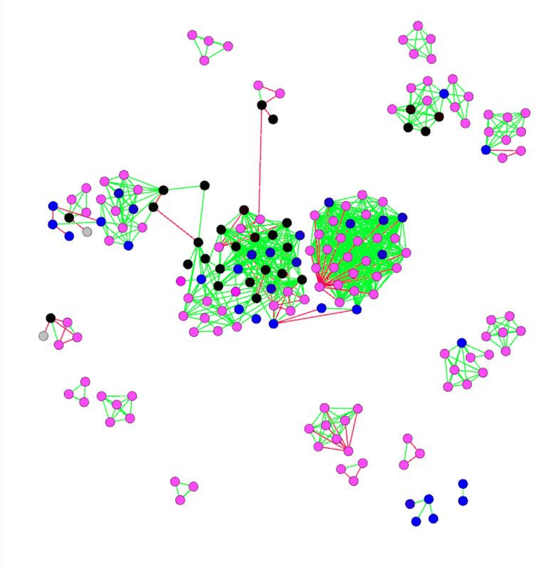


Figure 5.67: S=1 T=0.75

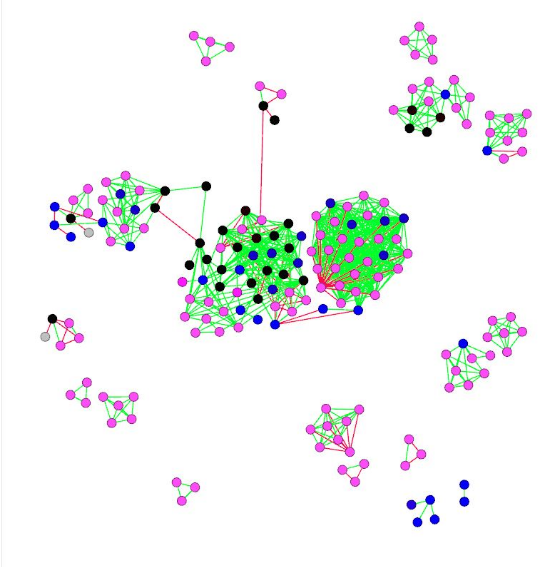


Figure 5.68: S=1 T=0.5

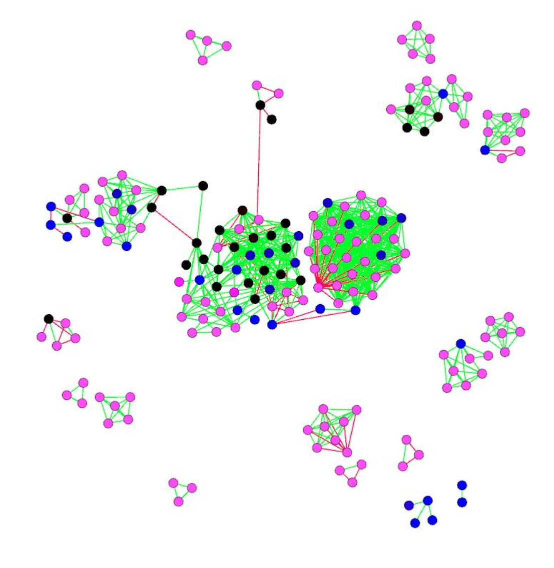


Figure 5.69: S=1 T=0.25

For the ‘E’ Group of data, the sessions coloured black, blue and pink are present and available in the analysis at that *s-val* and *t-val* setting:

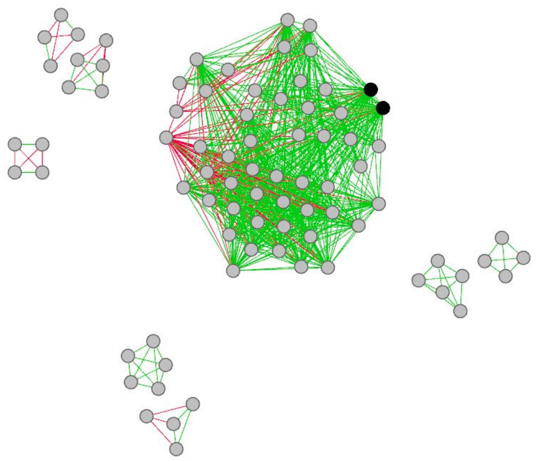


Figure 5.70: S=3 T=1

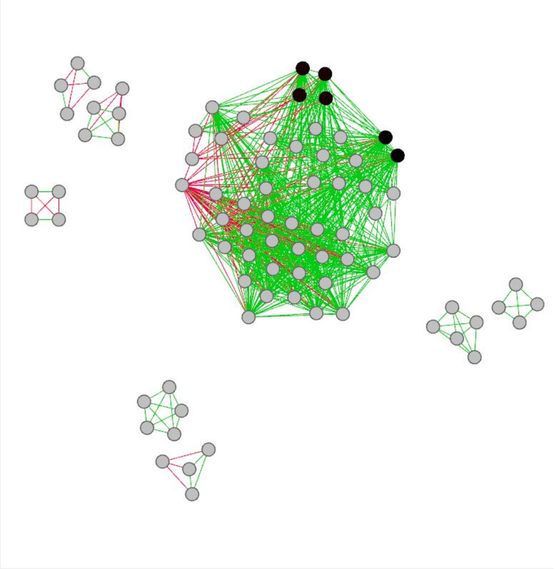


Figure 5.71: S=3 T=0.75, T=0.5 and T=0.25

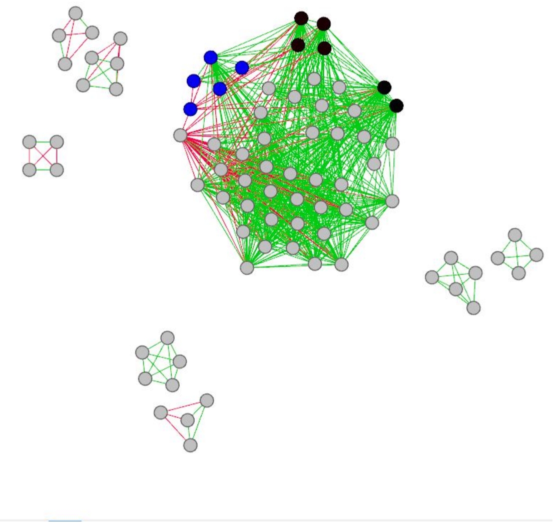


Figure 5.72: S=2 T=0.5 and T=0.25

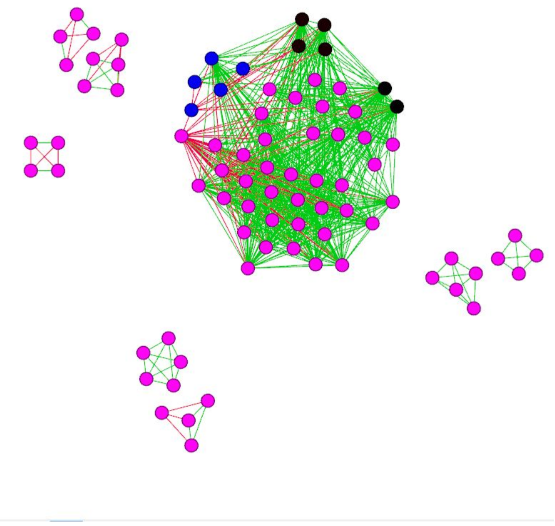


Figure 5.73: S=1 for all *t-vals*

In figures 5.66 through 5.69 we see S=1 and the introduction of the pink coloured nodes, which shows a substantial increase of available nodes for analysis, but given that in this test data we know the ground truth of the correctness of these interconnections, we also see large numbers of incorrect edges in the graphs.

For the 'L' Group of data, the sessions coloured black, blue and pink are present and available in the analysis at that *s-val* and *t-val* setting:

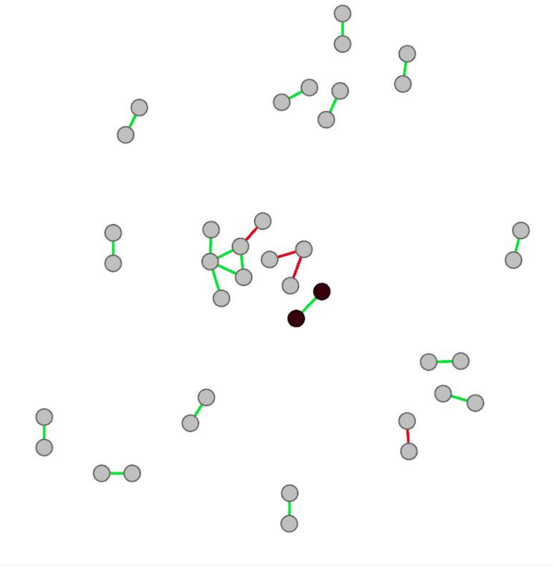


Figure 5.74: S=3 T=0.75 and T=0.5

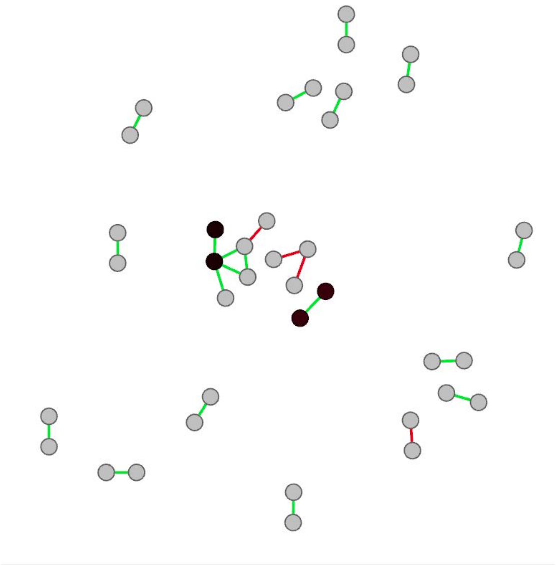


Figure 5.75: S=3 T=0.25

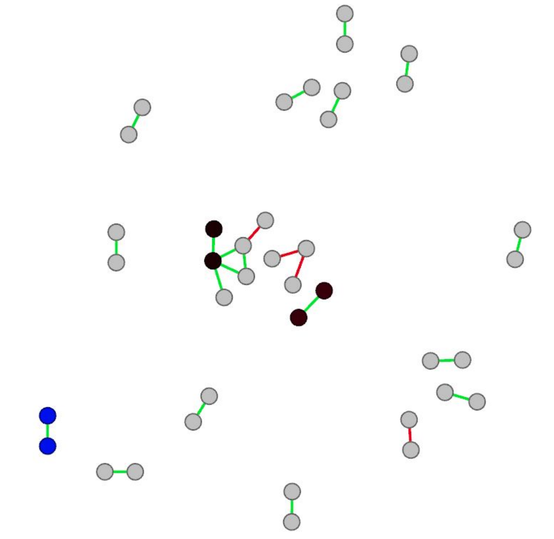


Figure 5.76: S=2 T=1

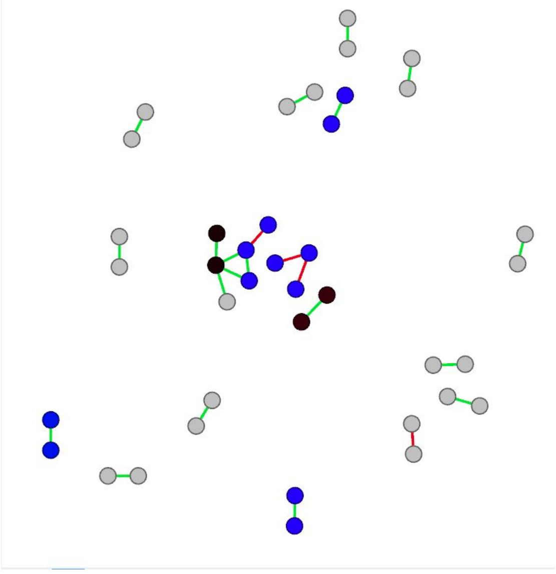


Figure 5.77: S=2 T=0.5 and T=0.25

We see in figures 5.70 to 5.73 the same approach used on this dataset for the 'E' group, however we show a smaller number of figures for this set as there is less data, which is less sensitive to changes of the *s-val* and *t-val* variables.

Similarly to figures 5.70 to 5.73, we show here the effect of the 'L' Grouping in figures 5.74 to 5.79. This data is also not as sensitive as the 'D' grouping and as

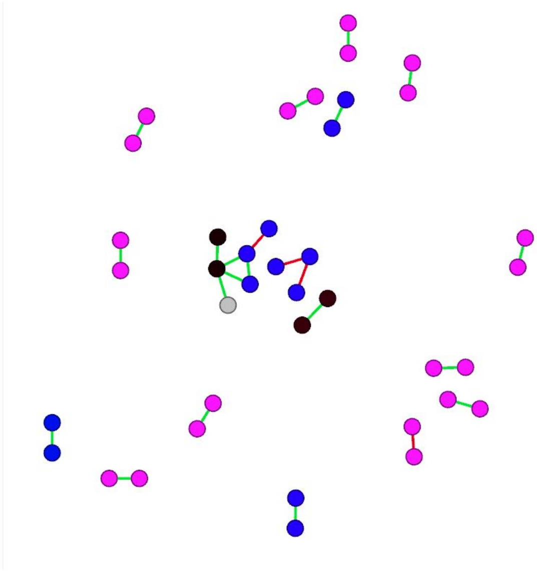


Figure 5.78: S=1 T=1

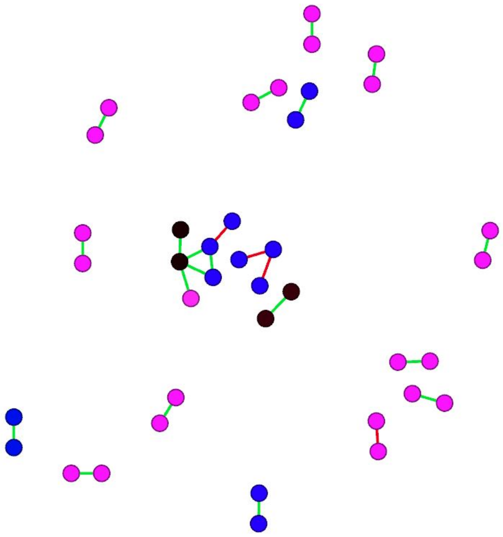
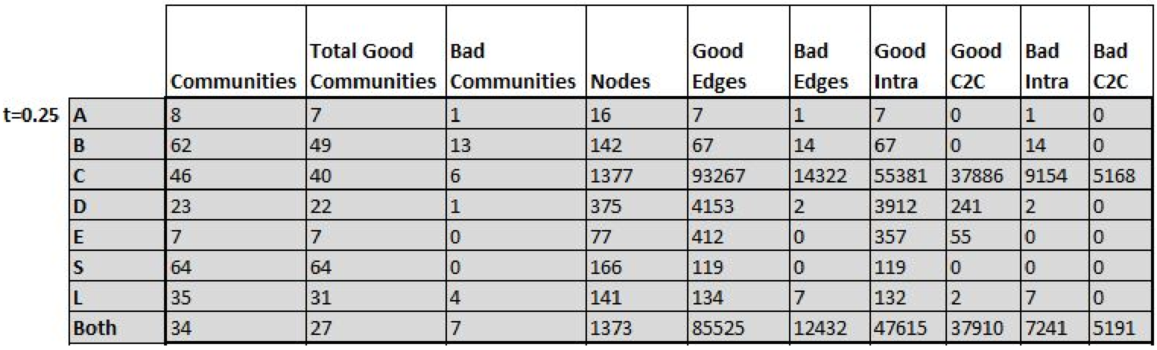


Figure 5.79: S=1 T=0.75, T=0.5 and T=0.25

such there are fewer figures presented here for the ‘L’ grouping than were presented for the ‘D’ grouping. We can see that this approach is visually simple and could be implemented in software with a simple dial or slider, but with these groupings of the datasets, D, E and L, they are sufficiently small and manageable for a human analyst or investigator, where this would not be as simple or easy to use for the larger scale groupings such as the ‘Both group’ and ‘C group’ and the somewhat large size of the ‘S group’, these have been omitted from the thesis.

5.6 USING COMMUNITIES TO IMPROVE PERFORMANCE

The ‘correctness’ of data is determined essentially during the group selection phase, coupled with the appropriate choice of the *t-val* and *s-val* for that group. The community detection algorithm does not at that point improve or modify the overall session-to-session, rather it clusters the sessions together and there may be errors within a community (what we have referred to as Intra-community errors) and there may be error between communities (what we have referred to as community-to-community or C2C errors).



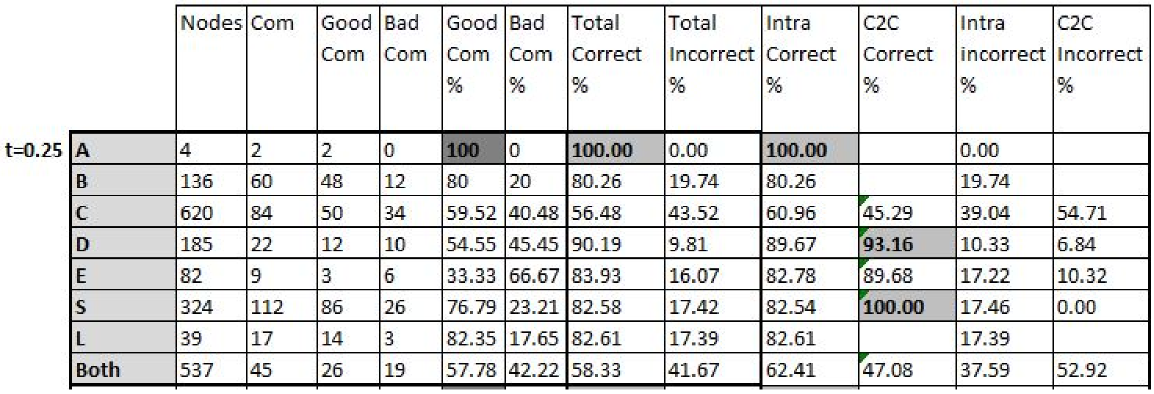
		Communities	Total Good Communities	Bad Communities	Nodes	Good Edges	Bad Edges	Good Intra	Good C2C	Bad Intra	Bad C2C
t=0.25	A	8	7	1	16	7	1	7	0	1	0
	B	62	49	13	142	67	14	67	0	14	0
	C	46	40	6	1377	93267	14322	55381	37886	9154	5168
	D	23	22	1	375	4153	2	3912	241	2	0
	E	7	7	0	77	412	0	357	55	0	0
	S	64	64	0	166	119	0	119	0	0	0
	L	35	31	4	141	134	7	132	2	7	0
	Both	34	27	7	1373	85525	12432	47615	37910	7241	5191

Figure 5.80: Raw data from the *W* dataset, where S=1

We can see in figure 5.80, an excerpt from the *W* dataset, the D group of data which has 23 communities (determined using the Unweighted Louvain method for community detection (noted in section 5.3.1) with a Resolution of 1.0). There are 3912 correct edges within communities, there are 2 edges that are incorrect within the communities (the Intra-community error) and there are no errors between communities with the D group (the C2C error). Indeed, with this dataset the C2C error occurs only with the C group and the Both group. As can be seen in the appendices, for the dataset used in this thesis the L grouping scheme never produced C2C errors, and the S, D and E grouping schemes only produced it where there were lower *t-val*s.

We can therefore propose a method where all community-to-community edges are severed, which has the effect of improving the overall ‘correctness with the C and

Both grouping, particularly in the lower *t-val*s. An example of this can be seen in figure 5.81:



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	B	136	60	48	12	80	20	80.26	19.74	80.26		19.74	
	C	620	84	50	34	59.52	40.48	56.48	43.52	60.96	45.29	39.04	54.71
	D	185	22	12	10	54.55	45.45	90.19	9.81	89.67	**93.16**	10.33	6.84
	E	82	9	3	6	33.33	66.67	83.93	16.07	82.78	89.68	17.22	10.32
	S	324	112	86	26	76.79	23.21	82.58	17.42	82.54	**100.00**	17.46	0.00
	L	39	17	14	3	82.35	17.65	82.61	17.39	82.61		17.39	
	Both	537	45	26	19	57.78	42.22	58.33	41.67	62.41	47.08	37.59	52.92

Figure 5.81: Data from the Y dataset, where S=1

In figure 5.81 from the Y dataset (using the weighted Louvain method with a Resolution of 1.0), we can see for the 'Total Correct%' column that the C group has a 56.48% correctness and we see for the 'C2C Incorrect %' has 54.71% of the C2C edges being incorrect. By severing all of these edges (correct and incorrect) the overall is raised to 60.96%, which although far below the 91% correctness for us to call the grouping data 'Beyond Reasonable Doubt' it does show a method of using communities to improve performance, but at the cost of not being able to associate sessions in adjacent communities together.

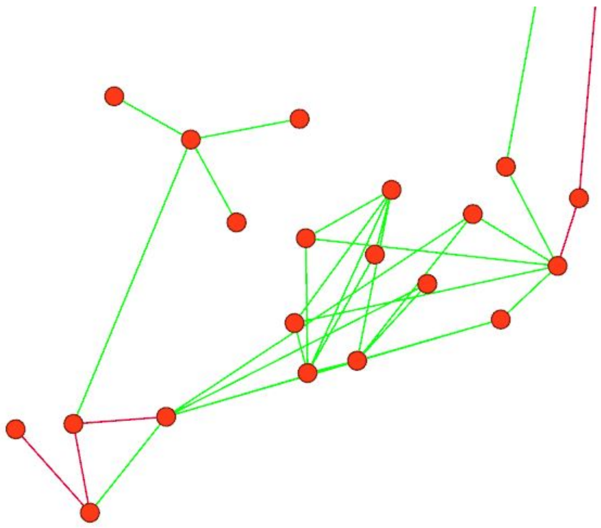


Figure 5.82: Data from the Y dataset, where S=1 and showing a single community of nodes all coloured red

Figure 5.82 shows an excerpt from the Y dataset where all sessions belong to the same community. The communities were calculated using the weighted Louvain method with a Resolution of 1.0. If the resolution is reduced to 0.1, as we can see in figure 5.83 the same network does not contain a single session, but rather contains four communities, two of which contain no intra-community error. By severing the C2C edges for the light green community from the blue community at ‘A, we have automatically improved the ‘correctness of this set of data.

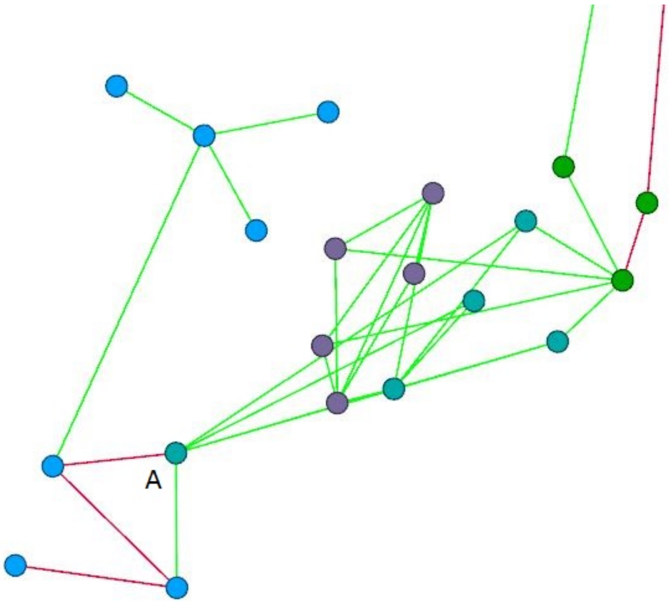
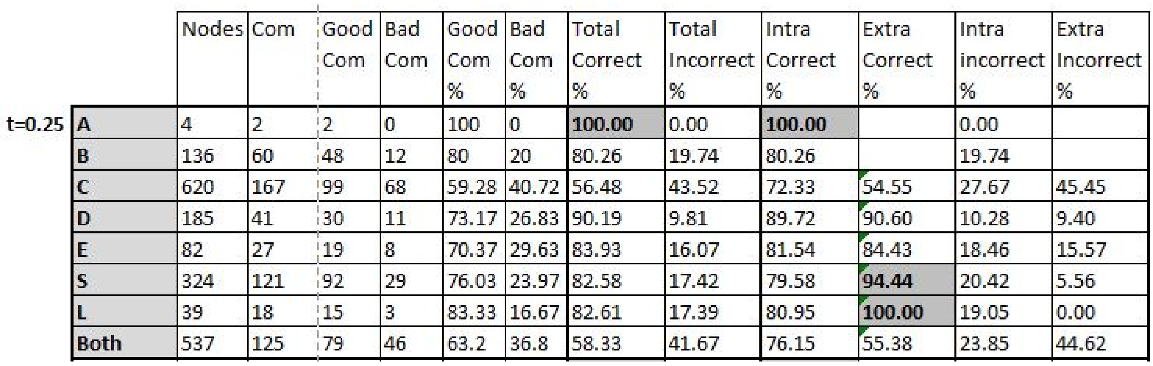


Figure 5.83: Data from the Y dataset, where S=1 the same as figure 5.82 however this time showing 4 different coloured communities

We can see in figure 5.84 that the same Y data from the above figure 5.81, however this time the resolution of 0.1 is used to calculate the communities (using the weighted Louvain method) which produces substantially more communities, 167 rather than 84 for the C grouping. The 'Intra Correct %' (i.e. the correctness after all of the C2C edges have been severed) is 72.33% which is a substantial improvement on the 'Total Correct %' of 56.48% with all of the C2C edges retained.



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	Extra Correct %	Intra incorrect %	Extra Incorrect %
t=0.25	A	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	136	60	48	12	80	20	80.26	19.74	80.26		19.74	
	C	620	167	99	68	59.28	40.72	56.48	43.52	72.33	54.55	27.67	45.45
	D	185	41	30	11	73.17	26.83	90.19	9.81	89.72	90.60	10.28	9.40
	E	82	27	19	8	70.37	29.63	83.93	16.07	81.54	84.43	18.46	15.57
	S	324	121	92	29	76.03	23.97	82.58	17.42	79.58	**94.44**	20.42	5.56
	L	39	18	15	3	83.33	16.67	82.61	17.39	80.95	**100.00**	19.05	0.00
	Both	537	125	79	46	63.2	36.8	58.33	41.67	76.15	55.38	23.85	44.62

Figure 5.84: Data from the Y dataset, where S=1

The difference between using the resolution of 1.0 (which had an 'Intra Correct %' of

60.96%) and a resolution of 0.1 (which had an 'Intra correct % of 72.33%) strongly indicates that the use of finer granularity community detection algorithms is effective when dealing with the larger groups of data such as the C and Both groups. The same approach when used with the other grouping systems was largely ineffective, predominantly as we noted above that relatively few grouping systems had much C2C error present.

We show in the appendices A.0.4 to A.0.7 the Y dataset processed using the Louvain method for community detection (noted in section 5.3.1), Weighted, Unweighted, at Resolution 1.0 and Resolution 0.1. for the *t-vals* T=0.25 to T=1.0 and for *s-val* of S=1 to S=3. There is very little difference between weighted and unweighted algorithms and as we note above there is substantial difference for the 'Intra correct % versus the 'Total Correct % when using the finer granularity resolution, but only for the C and Both grouping systems.

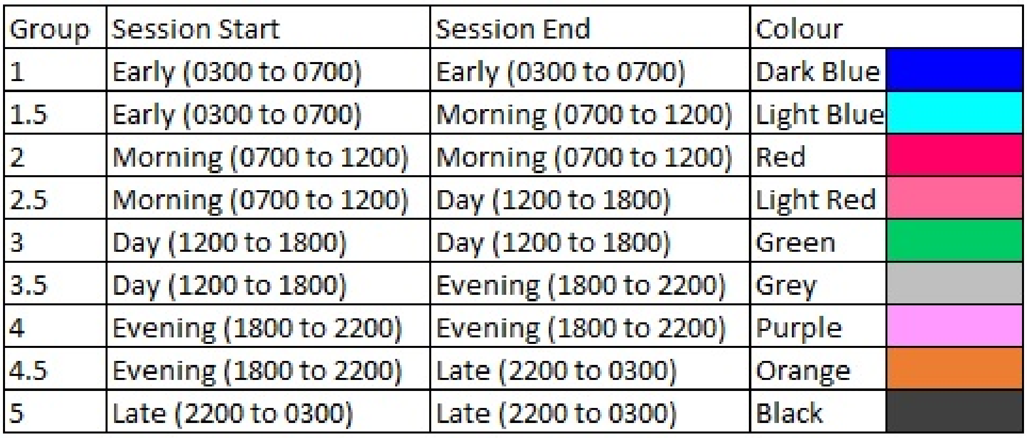
5.7 PATTERNS OF LIFE

We define Patterns of Life within our Context Analysis Approach as a model of how a device was used in time and space. As we have noted, the spatial component is largely left as future work for this thesis, but forms a generally important consideration for future research with mobile computing devices, and for the use of these techniques with general communications records data, for example Cell Site analysis.

We need to discover, or impose a model of the operation of the device in the 'real world' and then see if that model identifies relevant patterns that can be used by the analyst or investigator.

As we showed in section 5.4 above, we can plot the session-to-session matches onto a graph and then perform community detection on the clusters of nodes/session. If we then apply some model of the real world onto those nodes we can identify sub-graphs within the communities or across communities.

We show here an example Pattern of Life model where the sessions are categorised by a time of day (as proposed in Chapter 4.5.3), as shown in figure 5.85.



Group	Session Start	Session End	Colour	
1	Early (0300 to 0700)	Early (0300 to 0700)	Dark Blue	🟦
1.5	Early (0300 to 0700)	Morning (0700 to 1200)	Light Blue	🟦
2	Morning (0700 to 1200)	Morning (0700 to 1200)	Red	🟥
2.5	Morning (0700 to 1200)	Day (1200 to 1800)	Light Red	🟥
3	Day (1200 to 1800)	Day (1200 to 1800)	Green	🟩
3.5	Day (1200 to 1800)	Evening (1800 to 2200)	Grey	⬜
4	Evening (1800 to 2200)	Evening (1800 to 2200)	Purple	🟪
4.5	Evening (1800 to 2200)	Late (2200 to 0300)	Orange	🟧
5	Late (2200 to 0300)	Late (2200 to 0300)	Black	⬛

Figure 5.85: A Time of day model for Pattern of Life analysis

Note, unlike in chapter 4.5.3 we have included here a session start group and session end group, taking what was 5 categories and increasing this to 9 (theoretically 10, although group 5.5 did not appear in the example data that we used) as a response to the real-world data not fitting into the pre-determined time boxes. The problem of starting in one time group and ending in another is caused by the use of Variable-length session classification (chapter 3.4) where the total length of the session is unknown in advance, whereas if analysis approach used Fixed-length sessions the Pattern of Life model could easily be fitted to those fixed time 'buckets'.

We illustrate here in figure 5.86, the application of this temporal model by using the W dataset, showing the D and E groups of the relative popularity grouping method that has had community detection applied to the graph.

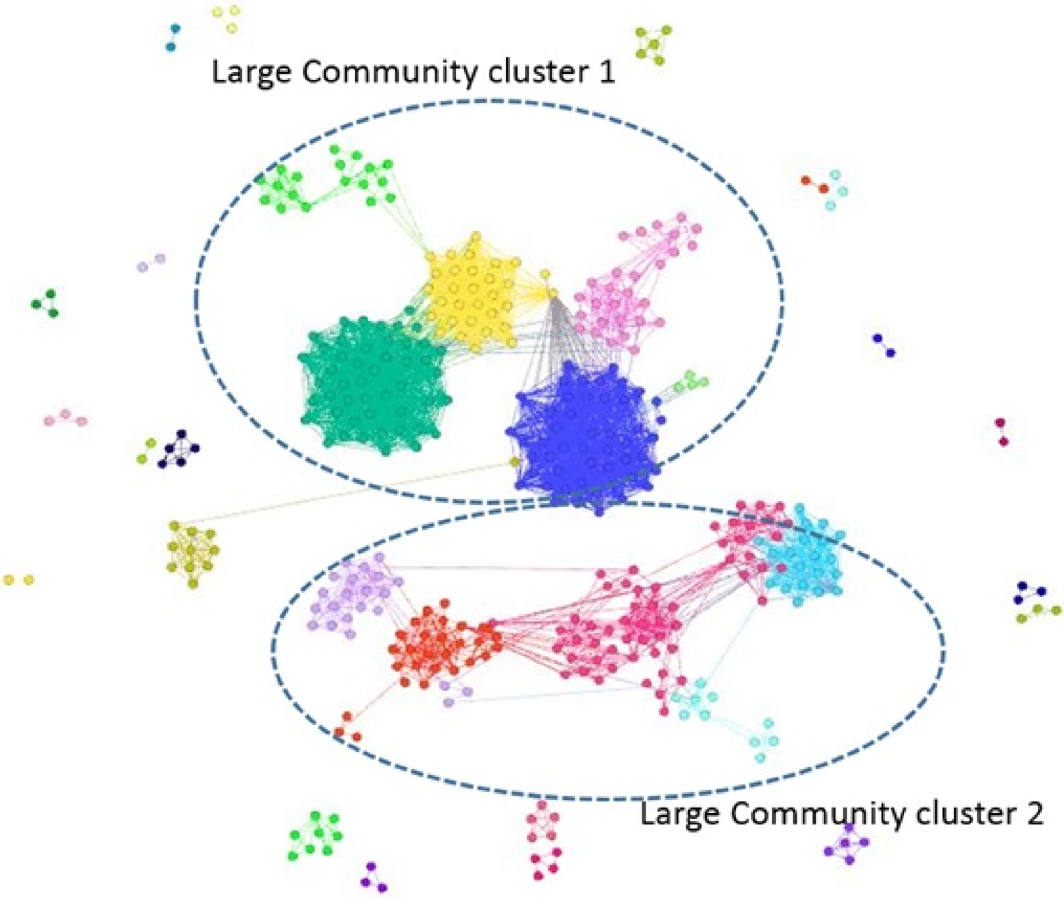


Figure 5.86: The *W* dataset's DE Groups with an *s-val* of 1 and a *t-val* of 0.50, showing the communities

By applying the temporal model from table 5.85 in figure 5.86, we get the result shown in figure 5.87. What we can note is that only one of the large communities and two of the small outlier communities has group 1 (and 1.5) data present in them, which is emphasised in figures 5.87 and 5.88.

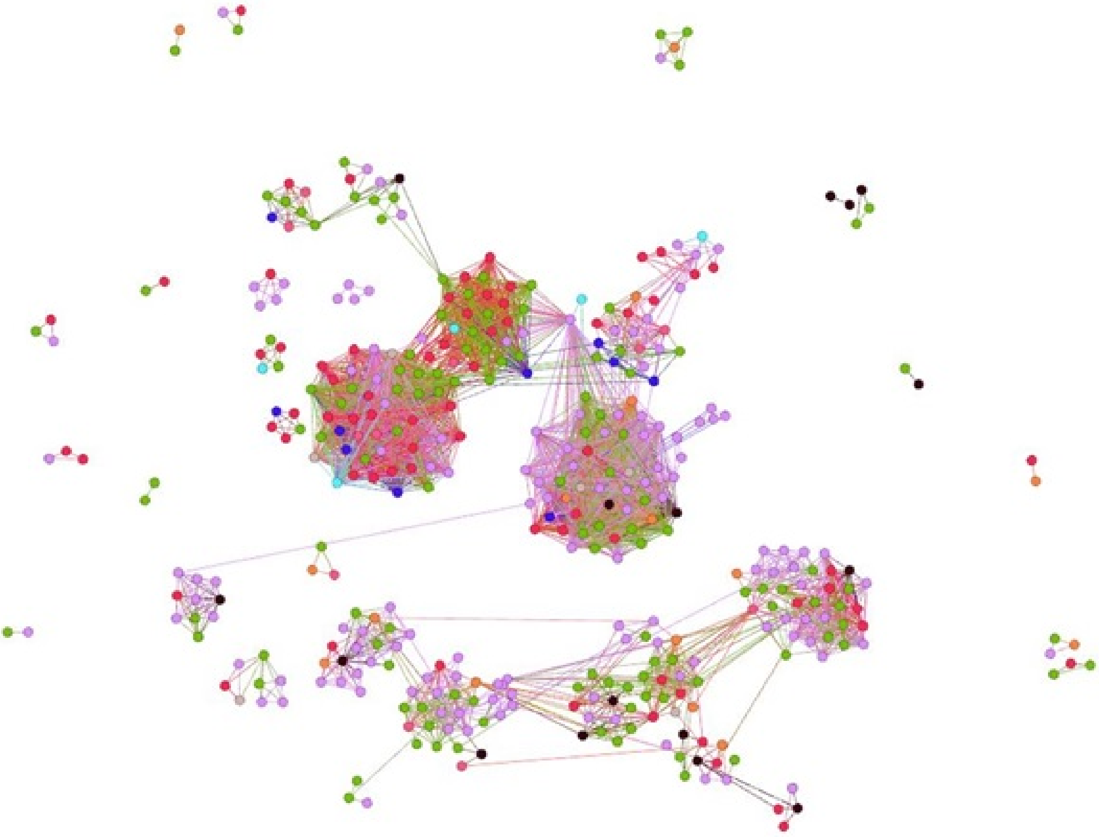


Figure 5.87: The temporal model overlaid onto Figure 5.86

Figure 5.88 shows that the two sessions in the outlier sessions are not connected, within this dataset, to any other sessions and therefore are excluded from the pattern of life analysis. It does suggest however that the outliers may belong to the user associated with the activity seen in large community 1, and as such this could be a situation when the edges just below the *t-val* would be looked at to see if they were connected.

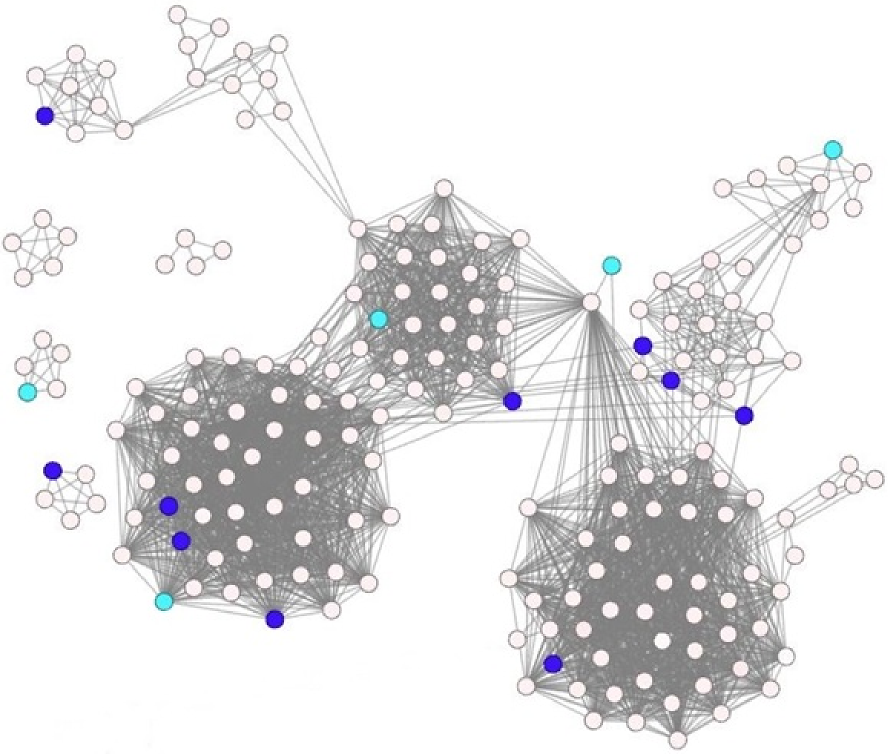


Figure 5.88: A section of the *W* dataset's DE Group with the Early (Blue) and Early/Morning (Light Blue) sessions highlighted

We can remove all of the sessions that have not been highlighted to create the subgraph and determine if there are direct and indirect connections between the remaining blue sessions, as can be seen in figure 5.89.

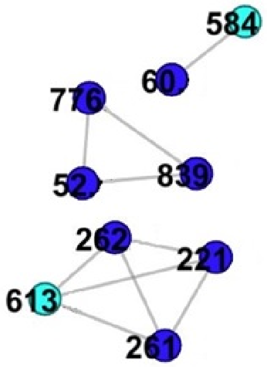


Figure 5.89: The *W* dataset's DE Group with a *t-val* of 0.5, showing only the interconnected dark blue and light blue nodes

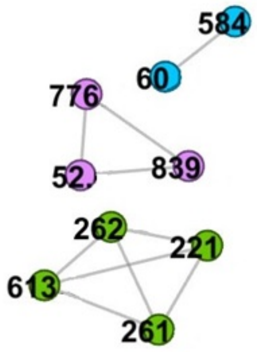


Figure 5.90: shows the communities for the nodes highlighted in Figure 5.89

Figure 5.90 shows the original communities that had been identified using the community detection approach with this dataset, and we can see that although there are almost no communities that are exclusively of one group of temporal data in the original graph (in figure 5.87), there is a general trend that can be observed of Purple/Evening (34%) and Green/Day (31%) being mixed with Red/Morning (18%). Again, we note that the Blue/Early (2.43%) and Light Blue/Early and Morning (1.22%) are exclusively grouped with one of the large communities with a small number of outliers.

When the sessions belonging to all of the other Pattern of Life groups are removed we find in figure 5.89 that there are 9 sessions remaining, in three subgraphs. Figure 5.90 shows the Pattern of Life data with respect to the original community data. We see that the members of the three subgraphs all consistently belong to the same communities.

We can show the discovery of interconnected direct relationship sub-graphs within communities. This can be used to further strengthen the case when testing a relationship between two or more sessions, or discovering trends within a community. The testing of investigative hypotheses and the discovery of lifestyle data is described in the next section.

5.8 INVESTIGATIVE REASONING

An investigator may want to use the proposed context analysis approach to either discover information about the operation of a device in time and space, which may be used to provide lines of enquiry for an investigator, or to test an investigative hypothesis/question which supports or refutes a line of enquiry, specifically with respect to the time or place that a device could have been used.

To use this approach, the investigator must be able to frame all questions as relating to the test of a session in a community or matching with another session. The session(s) the investigator in interested in must also belong to a particular grouping and consequently adjusting the *s-val* and *t-val* may be required such that the session(s) appear in the analysis, with the consequent understanding that if the adjustment is very low then there is the realistic possibility that there are precision errors, especially if the sessions appear in the C or Both groupings.

5.8.1 TESTING THE INVESTIGATIVE HYPOTHESIS

The following are tests that we can do well with our approach using the graphs:

- Membership Within a community.
- Direct Relationship within a community between 2 members.
- Indirect Relationship within a community between 2 members via 1 or more other members.
- Direct Relationship between members of 2 different communities.
- Indirect Relationship between members of 2 different communities via 1 or more other members.

To illustrate these tests, figure 5.91 shows two communities taken from the W dataset DE group, and we list here a set of tests we can perform across these two communities:

- Test of membership in a community (A belongs to Blue community)
- Test direct relationship between sessions within a community (A has a relationship with B)
- Test indirect relationship between sessions within a community (C has a Relationship with E, via D)
- Test direct relationship between sessions, between different adjacent communities (A has a relationship with C)
- Test indirect relationship between sessions, between different adjacent communities (B has a relationship with C via A)

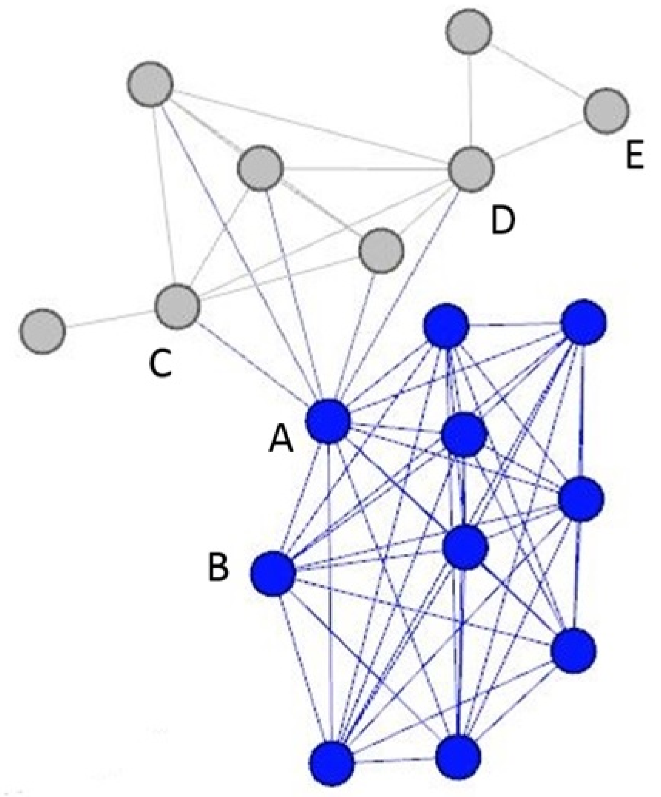


Figure 5.91: Two communities from the W dataset

Can we frame the typical kinds of questions an investigator would want to ask as questions relating to the test of a session's membership within a community or relationship to other sessions? The following presents a sample of investigative questions that an investigator could simply test using our approach:

- At time X, was the usage of the device
  - normal,
  - or abnormal?
- If at time X we have a notable event (an offence) and at time Y we have information relating to the identity of the user, can we tie the user at X to the user at Y?
- Who are the users of a device?

  - Who was the user of the device at time X?
  - How many regular users of the device?
- Can we exclude sessions and communities as part of an investigative question?

At time X, was the usage of the device normal, or abnormal?

A reasonably straightforward question if time X corresponds to a session, although what is 'normal' may be contentious. This is a test of membership in a community and if the session belongs to a large community (or communities if grouped in such a way), the size and interrelatedness of the communities would determine if this was particularly 'normal'. A Pattern of Life model may also be used to show that the community also consists of a large number of sessions occurring within a similar time box, day of the week etc.

If at time X we have a notable event (an offence) and at time Y we have information relating to the identity of the user, can we tie the user at X to the user at Y?

This is a test of direct relationship, or a test of how distant an indirect relationship is. Although the test is fairly straight forward, we have discussed at the beginning of chapter 5 the dangers of associating two sessions indirectly connected where there is absolutely no direct relationship. However, we also see in figure 5.86 that where there is very clear segregation between the user's behaviour, even a low *t-val* is enough to emphasise the difference.

Therefore, if a direct relationship is present between two sessions that have personably identifiable information and notable events then this can be considered good. If there is also an indirect relationship alongside a direct relationship this could be considered a 'good and supported' relationship. If there are multiple 'good' relationships the analyst should be in 'good confidence' of the relationship. If there are only indirect relationships then these relationships should be considered 'possible' but have a declining level of confidence for each level of indirectness, particularly with respect to crossing community boundaries.

Who are the users of a device?

The questions relating to users are both the most desirable from an investigate standpoint, yet also the most technically difficult to prove and largely do require some knowledge prior to the analysis about the identity of one or more users of the system. The questions of users can be the general question of who are the pool of possible users, and then the specific case of who was the user at a specified time:

Who was the user of the device at time X?

This question requires that a session contains personal identification for the most reliable answer. If the personal information is not available during the time X session then we have to look at if the session can be associated directly or indirectly with another session, within the same community, or if possible, or desirable, within an alternative community. How many regular users of the device?

The approach presented in this thesis does not, at least at this time, attempt to predict the number of users of a device and this remains an area of future work. We can however see in figure 5.86 that by adjusting the *t-val*, it may be possible to identify clear segregation in the data suggesting, accurately in figure 5.86, the number of distinct users, i.e. a large community of interconnected sessions for each of the users. We have shown in our experiment that the D and E groups perform well at establishing the most idiosyncratic data relating to the users, and as such there may be much less clear cut partitioning of the data if for example, using the C or Both group data.

Can we exclude sessions and communities as part of an investigative question?

We can classify broadly based upon some prior knowledge that has been determined during the investigation, for example "The suspect is at work between 9 to 5, Monday to Friday and has taken no other time off during the year". Such information can be used to identify all the sessions that occur between 0900 and 1700 hours, Monday to Friday, and determine if there are patterns of usage during that time, if there are patterns between 1701 and 0859 hours and for all of the weekend.

In such a scenario, we have two groups or patterns of life: Group 1) 0900 to 1700, Monday to Friday; Group 2) all other sessions. The investigator would naturally want to exclude communities where both of these patterns are present as being either, belonging to a second user, or being sufficiently non-unique to the suspect. Indeed, if any artefacts of note appear within a community that contains patterns from both groups - and the investigator is sufficiently happy that the data is likely to be idiosyncratic - then this would be a strong exculpatory line of enquiry, suggesting that the suspect is unrelated to the notable artefacts.

5.8.2 DISCOVERY OF LIFESTYLE INFORMATION

It is entirely possible, particularly when considering Internet history derived from server-side Internet Communications Records, that an Internet history could be analysed, before any of the witnesses have been interviewed. The results of an early analysis may be used to direct lines of enquiry, to frame the questions that need to be asked by investigators to most effectively establish ownership of a device at any particular time or place. There is also the iterative nature of investigations where witnesses and suspects can be re-interviewed as an investigation progresses.

Without any prior knowledge of the circumstances relating to a device an analyst can extract the Internet history, compile sessions, perform grouping and session-to-session analysis. The results of the session-to-session analysis can we have shown in this chapter be graphically represented and community discovery can be performed on that data. Normally this would be the point to test investigative hypotheses, but in an early analysis it is possible to unpick and discover characteristics and components within the communities, allowing an investigator to ask probing questions about the individuals' interests, or times they are active on a device etc. For example, a general question about "who uses the device in the mornings, who uses it late at night?" would be standard fare for an experienced investigator. However, the addition of "Who in the house likes Motorcycles?" becomes available as a line of enquiry because the investigator has noted that one of the communities contains sessions that contain components relating to motorcycles.

We can see this approach with figure 5.92 where we have three communities that have been determined from session-to-session analysis, that have then been reduced back to showing only the components that belong to those sessions in figure 5.93.

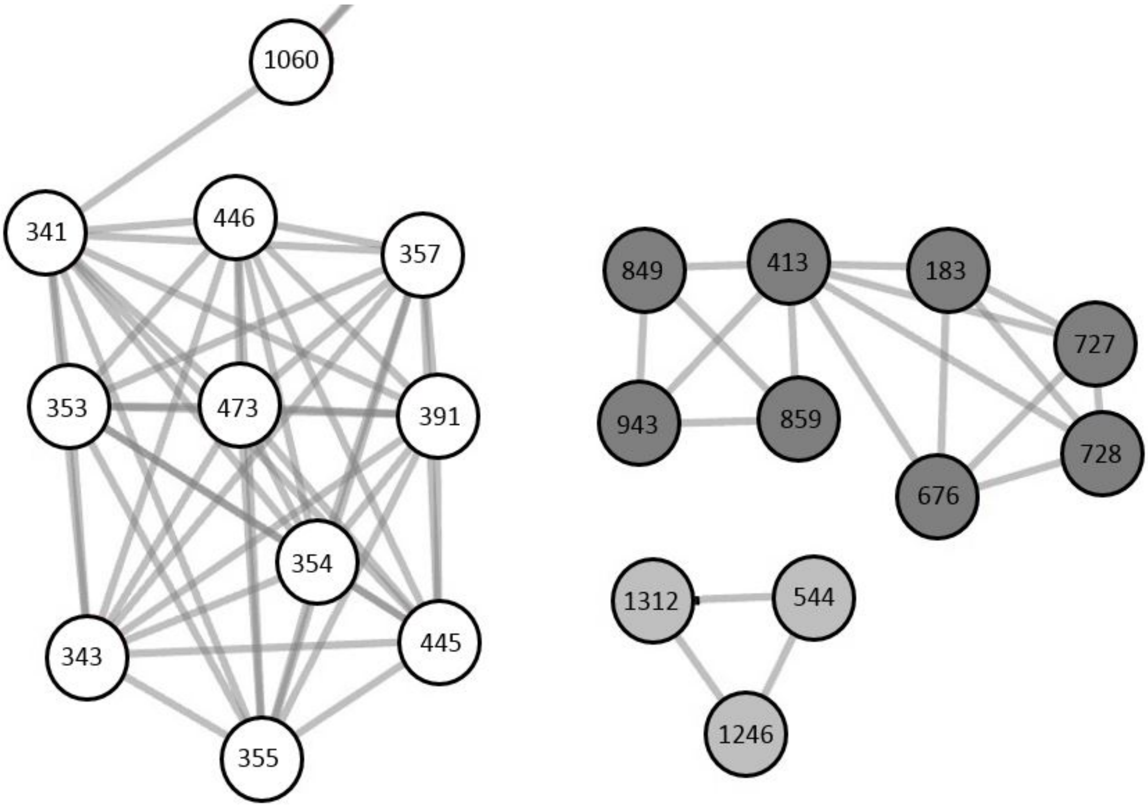
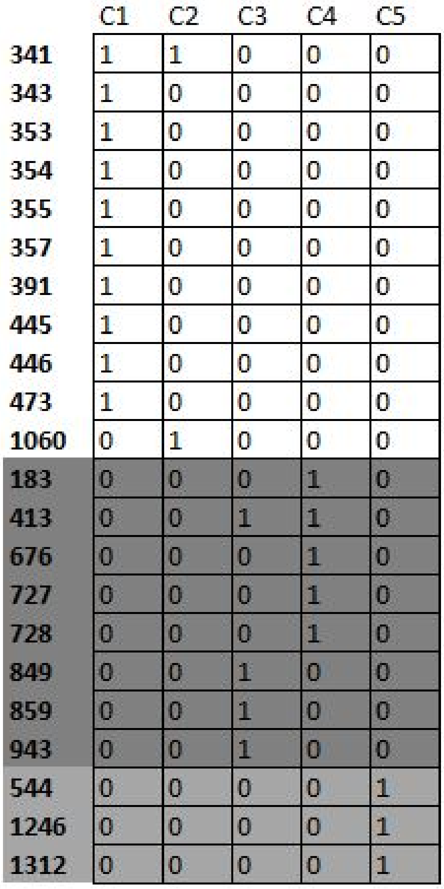


Figure 5.92: Three Communities from *W* dataset, DE Group with *s-val* 1, *t-val*=0.5

We can see in figure 5.93 that there are 5 components (C1 to C5) that relate to these 22 sessions in the 3 communities. These sessions have an *s-val* of 1, which is to say that a single component was required to make a session and the largest number of components in any of these sessions is 2.

We can see that the uncoloured community is based upon primarily C1 with some C2, the dark grey community is split between C3 and C4, and the light grey community is based purely upon access to C5.



	C1	C2	C3	C4	C5
341	1	1	0	0	0
343	1	0	0	0	0
353	1	0	0	0	0
354	1	0	0	0	0
355	1	0	0	0	0
357	1	0	0	0	0
391	1	0	0	0	0
445	1	0	0	0	0
446	1	0	0	0	0
473	1	0	0	0	0
1060	0	1	0	0	0
183	0	0	0	1	0
413	0	0	1	1	0
676	0	0	0	1	0
727	0	0	0	1	0
728	0	0	0	1	0
849	0	0	1	0	0
859	0	0	1	0	0
943	0	0	1	0	0
544	0	0	0	0	1
1246	0	0	0	0	1
1312	0	0	0	0	1

Figure 5.93: The Session table for Figure 5.92

In this example, the ground truth for these 3 communities is that the light grey community has an error in it that cannot be detected purely by examining the data and graph. Session '544' belongs to 'User 1', whereas the remaining two sessions in the light grey community belong to 'User 2'. The light grey community is however a small community and that should be considered when using this approach for discovering lines of enquiry.

An analyst could see from the graphical representation that session '1060' is potentially the odd-one-out in the uncoloured community and then looking at the data to see the C1 component is the most crucial information within that community. Similarly, in the dark grey community an analyst can see from the graphical representation that session '413' sits between two smaller communities and as such that might be point where the analyst would not be prepared to show indirect connect between sessions on either side unless question could show disclosure from witnesses or suspects to suggest either side of that community had relevance.

This approach to discovery offers an interesting way to work backwards and highlight potentially significant components, but as there are hundreds of components in a typical dataset, it may be desirable to only use this approach on a handful of components from the larger, or the ostensibly most significant communities.

By performing discovery of lifestyle information:

- The investigator can use this to interview witnesses and suspects about what is seemingly general questions about their overall Internet accessing habits, but with the specific goal of determining if there is disclosure about critical components, sessions or communities.
- Analysts can identify the most significant components for the communities that they may rely upon for evidence. They can then investigate those components in more detail, see if the original web pages are available, cross reference these against bookmarks or other intelligence to strengthen any arguments that they represent crucial behavioural indicators.

5.9 CONCLUSIONS

We have shown in this chapter that the Session-to-Session comparisons that were grouped in the previous chapter can be graphically displayed in a visually accessible way, and that community detection algorithms can be performed on these graphs to automate the grouping of sessions into communities of, what is ostensibly similar behaviour.

We have performed analysis in this chapter showing that the groups can have the *t-val* Jaccard similarity coefficient manipulated and the *s-val* minimum number of components that must appear in a session to produce a Beyond Reasonable Doubt (BRD) of 91% accuracy of the edges for the whole grouping of data. It is important to emphasise that the BRD is a match across all of the data in that grouping rather than the individual edges, as we have seen in chapter 3 that there are proportionally few session-to-session matches that are so similar that they have a match at greater or equal to 0.91.

We have shown in this chapter that number of sessions that are available for analysis, and consequently the number of communities that we automatically detect falls as the precision of the *s-val* and *t-val* are raised.

We do however show in this chapter that the groupings that we proposed in chapter 4, the relative popularity grouping and the Short-only and Long-only session groupings, do appear to extract 'idiosyncratic' sessions and group them into communities. We can make this statement by considering the overlap in the datasets: The Z and Y datasets have overlaps in the low 30% area, yet by grouping the data we can get a BRD (i.e. 9% or less error) for the groups D, E, S-only and L-only with quite small *s-val* and *t-val* manipulation for the Z dataset and a larger but reasonable manipulation for the Y dataset. With the W and X datasets that have an overlap of greater than 10%, but the same modest manipulations used for the Z data set produces a BRD of significantly higher accuracy, 95% to 100%.

There is a general trend in the regression graphs in section 5.3.5 that shows that the smaller the number of sessions used in the analysis (something which is controlled by the tuning of the *t-val* and *s-val* variables), then generally the higher the accuracy in the correct matching between the correct user identification. Depending upon the grouping scheme and differences within the data this we have shown can be accurate, but these variables are being applied across the whole dataset and it may be appropriate to provide finer levels of control at the community-level based upon domain-knowledge of activities. For example, we see in section 5.6 the higher possibility of error between communities and therefore the use of higher threshold *t-val* between communities and a lower value within communities may be appropriate rather than a general setting applied across the whole dataset. Two users with similar interests are likely to access similar websites if not the same websites. If the shared interests are what we would classify as sufficiently niche, then it would be difficult to disentangle those two users without domain-level knowledge of the actual individuals. The closer the overall similarity between the user characteristics, the greater the difficulty in separating the difference between the data that we would categorise as idiosyncratic. We therefore conclude that future work to automatically select appropriate *s-val* and *t-val* would be based upon:

- The domain-level knowledge of the communities or activities.
- The perceived similarity between the possible users of the device.

Notably both of these approaches are not zero-knowledge solutions, where the knowledge of the websites is present in the former option and knowledge of the possible users is in the latter option. We have noted that in the majority of our results (see section 5.3.3) the *s-val* of S=1 and *t-val* of T=0.25 is sufficient to differentiate the users within datasets such as the D and E groups of the W dataset. The original sources of data the D1 and S1 datasets were clearly interested in different sites therefore the techniques work very well with low *s-val* and *t-val* settings. Comparing this with the Y dataset where much higher values are required to get the same accuracy, which has a consequence of fewer sessions in the analysis. We therefore propose in future work experiments can be developed for two or more sets of users on a single device and determine the similarity of those users based upon the following criteria:

- Location Visiting websites that are specific to a geographic area, such as local news or regional transport websites. Schools, community centres and other subjects of limited relevance to people outside of the region.

- Friends and Family Shared friends, family and associates which may be geographically diverse (family members on the other side of the world) but of shared relevance to the users of the device.

- Occupation Visiting news, current events and industry/school/university related websites.

- Interests Sports, hobbies and activities

- Language On some systems/households there are members of the household that speak languages that may not be spoken by other members of the same household and consequently this is a significant discriminator between users on the system.

- Popular Culture Gender and Age may influence what popular culture, music and media that is consumed by the different users of the system.

Where there is overlap between the users we can see to what extent this impacts the *s-val* and *t-val*. This approach would still somewhat be a zero-knowledge approach as we are assessing if a model of the users could be used to dictate an appropriate setting for the variables. However, such an approach must also be considered alongside domain-level knowledge of the data and the communities this produces, otherwise we could still be subjecting our whole dataset to analysis with variables that may or may not be suitable for all intra-community relationships and community-to-community relationships.

Chapter 6

CONCLUSIONS

"An inconvenience is only an adventure wrongly considered; an adventure is an inconvenience rightly considered."

G. K. Chesterton - All Things Considered, 1908

6.1 CONTRIBUTIONS

From our objectives outlined in chapter 1, we have provided research in this thesis to achieve the following objectives:

- **Objective 1.** We have identified the state of the art and challenges in event modelling in multi-user computing environments and proposed a model of assessing this research called 'Context Analysis'.
- **Objective 2.** Identified Internet history artefacts and the levels of resolution we can expect to find for Internet history records on a standard multi-user computing environment or server-side records that might be retained by a Communications Service Provider.
- **Objective 3.** We have evaluated the feasibility of aggregating multi-user Internet history sessions without prior knowledge of the user and produced schemes for aggregating the history into sessions. We have found that the Session-to-Session Context Analysis approach gives a broad-based model of how the system

was used and a similarity comparison between any two periods of time in the history.

- **Objective 4.** Developed novel methods for grouping a computer system's Internet history without prior knowledge about the websites that are visited, their structure, and the users of the system, so as to identify and extract idiosyncratic features of the history with an accuracy we have classified as 'beyond reasonable doubt'.

- **Objective 5.** We have provided a novel visualisation of the grouped Internet history records, so the results of our analysis can be used for investigative reasoning and analysis of the aggregated history sessions.

The contributions of the research are therefore:

- Context Analysis: We propose an approach that considers how artefacts are related to other artefacts, such that they can be Identified, Interpreted, Verified and the Activity of the artefacts be analysed with respect to their peer artefacts. This has the novelty within Digital Forensic Science in that we consider the modelling of the whole Internet history of the system, rather than focusing around specific events.

- Sessions and Session-to-Session Analysis: We formally defined a novel aggregation method for Internet history as Fixed-length and Variable-length sessions. This approach allows us to evaluate periods of time in the Internet history and make meaningful comparisons between those periods of time. The process of aggregating into sessions can not only be used for analysing history from a computer, or smart phone device, but can also be used to analyse history from gateway devices such as firewalls, routers and the Internet Connection Records that new UK legislation [110] will compel Communication Service Providers to retain. We propose this approach could be used for analysing a variety of other types of activities that relate to events that can be aggregated together, such as sensors or access control logs etc.

- Grouping methods: We have proposed two novel methods of grouping or characterising the components within our sessions using zero-knowledge about the characteristics of the users of the device or the websites that they are accessing:
  - Short and Long Sessions: A simple but powerful threshold method of determining if the session is 'short', or 'long'. Although this method is not technically sophisticated it is fast to compute, requires no external reference and because of the way we have aggregated the history into session we have a novel way of analysing the website/components that appear only in short sessions, components that appear only in long sessions and the else condition of components that appear in both the short and long sessions, which is an interesting way of characterising the behaviour of the user.
  - Relative Popularity: This method of grouping our data contrasts the difference between the rank order of the popularity of components taken from a third-party reference source, and the rank order of the popularity of the component across the local system that is under investigation. This approach also lends itself to the statistical investigation of the 'normality' of web browsing behaviour and what is "a normal person's web browsing behaviour"?
- Graphing Session-to-Session Comparisons and Community Detection: We have taken the graphing of similarity coefficients between two or more nodes and used it in a novel way to graph the activity of Internet activity on a device. Using this new approach to Internet history analysis we can automatically detect the communities within these graphs which has the implication of showing like-for-like website access behaviour, which we have shown during our experiments in chapter 5.

6.2 DISCUSSION

The four sets of data represented four slightly different scenarios, but they are consistent with the types of scenarios that an investigator would want to model a large body of Internet history, such as outline is section 1.3:

- The W set was a large set of data split 2/3 for one user and 1/3 for a second user with a 12.9% in the components.
- The X set had a majority user (92.5% of the sessions) and a minority user (7.5% of the sessions) and this represented a scenario where there were two users that shared some activity (10.65% of the components).
- The Z set was approximately 50/50 usage between two users and the overlap between the users was 31.53%.
- The Y set was a 2/3 and 1/3 split between users with a 32.3% overlap in components.

We saw that the X, although having a similar amount of overall overlap as the W set, a small amount of *s-val* and *t-val* adjustment to the data drastically reduced the possibility of error. Similarly, the overall performance between Y and Z is quite different despite both sets of data having similar overall overlap. The total overlap of the components does not appear to indicate the correctness of grouping or the variable settings that must be used to achieve high precision and availability.

We can therefore conclude that to better evaluate the correctness of the grouping we could use the volume (in the Local Popularity sense) of the overlapping components to better estimate the similarity of the two datasets. That would allow us to better measure the reduction of error, rather than the correctness of matching.

Although this is interesting from an experimental point of view where we are considering how best to measure the accuracy and correctness of our proposed method on a set of data with a known ground truth, we ultimately are left with the issue that given any set of data, we cannot necessarily predict the probability of error in determining if a session was made by the same user as another session. What we can do is apply the techniques proposed in this thesis, determine if the sessions appear in groups that are likely to distinguish a user, and apply increasingly harsh variable manipulation. If two sessions still have a high similarity coefficient after all of that, then we can conclude that the two sessions were made by the same user "beyond a reasonable doubt", but we cannot, without the ground truth assess the exact BRD accuracy for that dataset. This is in large part because of what we describe below in 6.3.1, where we discuss we do not exactly know what is 'normal' or the volumes of 'typical' activity. With a theoretical model of behaviour, we can group and manipulate data in the way we propose in this thesis.

The development of such a model and the investigation of normal behaviour, along with other refinements to the grouping and community detection are necessary next steps in this research and are necessary before our approach could be used as a standard analytical technique, unquestioned in the courts of law.

6.3 FUTURE WORK

This thesis is a proof of concept that Internet history can be automatically and reliably broken up into periods of activity, and those periods can be accurately related to other periods of activity created by the same user, by grouping them based on the websites visited, with zero knowledge of the content and type of website. However, as a result of our work we have noted a number of issues to consider which due to the scope of this project we leave as avenues of further research.

6.3.1 NORMALITY OF BEHAVIOUR

The approach we have used has no knowledge about the type of websites, with the exception of the relative popularity grouping approach, which did require an external ranking of global popularity. Even with the global popularity ranking, there is no explicit knowledge about the type of website that is visited, although there is some implicit knowledge with respect to popularity such that the website may be search related, commerce, social media, pornography etc., i.e. popular websites.

This leaves a very important open question about what is 'normal' browsing behaviour. We can describe any session with respect to its normality on that system (Local Popularity), but at the moment we cannot describe the session with respect to a model of what we would expect to find on a typical system. If we are interested in a particular session, we can show if a session belongs to a community, if it belongs to communities across different groupings (for example, if the particular session contains components in communities for the C and D groups), and we can describe the size of the communities and look at pattern of life information relating to time of day that the sessions were made, location etc. What we do not have at the moment is a model that gives us an expectation about what components we should find in someone's Internet history, what time of day they were made and so on.

An investigation into the normality of Internet history would be interesting and would not only be able to assist the type of work we are doing here, removing or identifying potentially incorrect session-to-session matches, but would also form the basis of user behavioural profiling where characteristics such as gender, age, etc. could be used to anticipate the user at any time.

6.3.2 COMMUNITIES

We have used community detection in a relatively simplistic way, and we have not examined in great detail characteristics of the communities such as size, shape, density, efficiency and type of algorithms for finding the communities. However, we note that community detection on groups such as the S-only, L-only, A and B tend towards a large number of communities with few members that are clearly segregated from each other. The D and E groups typically have larger sized communities, fewer of them, but the communities are also fairly segregated. This segregation is clearly an artefact of the small numbers of components in each sessions.

Consequently, our initial testing of different community methods, weighted edges and different resolution settings did not drastically affect the results for the S-only, L-only, A, B, D and E groups. Where we believe there is scope for enhancing our findings is with the C and Both groups as these have many more components in the sessions and to remove false matching requires higher t-vals. Therefore we believe that we can substantially increase the number of communities, those communities will have good intra-community edges and those ‘sub communities’ will relate to what we talk about next which is categorising the behaviour within the sessions and communities.

6.3.3 CATEGORISING A SESSION

A logical next step in this research is the automatic categorisation of the behaviour in the communities. The approach thus far has used zero explicit knowledge about the components, but we see that in chapter 4.5.1 we propose that if ‘type’ data was known for the components we could use that for matching. A perhaps more useful approach would be to enable an analyst to highlight or colour a community based upon the dominant type of websites that make up a community.

It is also worth considering that there are generally few components that make up the communities for the S-only, L-only, A, B, D and E groupings. It would be reasonably simple task to group those communities, if, and this is where it becomes a non-trivial task, there was the appropriate 'type' data for the components, given the difficulties we note in chapter 4.5.1. As the S-only, L-only, A, B, D and E groupings are likely to contain 'niche' or regional specific websites, the possibility of having the 'type' data is much lower than say with the C or Both group data which would likely be much more commonly available, even with prior knowledge from a different case.

6.3.4 DISPLAY AND PRESENTATION

There is still an open area of research in how to create an effective and useful front-end for the different stakeholders and users of this approach:

- The analyst is technologically literate, but may not be fully versed in the details of the case. This person would normally be expected to search and manually extract Internet history so they will want to be able to apply keyword searches to the history, find the sessions this best relates to and then associate that with other sessions and other files and events on the digital devices. The analyst would be happy with graph views and being able to manipulate the variables so as to be confident that they have good level of precision and recall.

- An investigator may not be as technically literate as an analyst (if a different person), but will be fully familiar with the circumstances, offences, points that they need to prove to the court and witness testimony. The investigator will want to test alibis, find out if there is history for times and dates that they have statements about and consequently they will be interested in relating this back to the real-world and will want calendar views etc.

- The court. Professional lawyers and lay people such as a jury may have very little technical knowledge and they will only have the case-specific knowledge that has been presented to them as part of the proceedings. If they are being told that some period of time is like another period of time they will want to know from an expert if that is a fact or an opinion, and how confident the

> opinion is. This may, indeed should, involve an expert being able to demonstrate in a clear non-technical way why two or more sets of data are related. We have presented in this research a functional way of graphing data that satisfies the majority of the analyst-level view. The Jaccard similarity tables (such as seen in figure 3.5) are fairly straightforward and can provide a visually simple way to demonstrate similarity with court-level view. We have focused our experiments in Chapter 5 at a level "beyond reasonable doubt", but we are still left with open questions about determining the precision and recall/availability, so there may be desirable views for the investigator where they can 'dial it up' to demonstrate the highest degree of assurance. The practical usability of our approach for these different levels therefore remains an interesting future area of work.

To address if the techniques presented in this research is immediately admissible in the UK legal system it very much depends upon if the work is being used to identify an individual, or if it is being used to test either an investigative hypothesis or an affirmative defence made by a defendant/suspect.

Admissibility is very much the grant of the court and with all expert techniques and testimony it is subject to the test of how accepted within the scientific community. Therefore, at this stage the admissibility of the whole proposed method would be subjected to scrutiny.

The first major stage is if the court would accept the aggregation of Internet history into a 'session'. If this was accepted then the use of session-to-session analysis is straightforward and based on Set Theory, and although the application in this research is novel, the concept is well accepted, available for inspection and explicable even to a lay jury. Therefore, we suggest that the testing of affirmative defences, e.g. "It wasnt me, it must have been someone else" can be easily tested for the occurrence of a repetitive pattern.

For proactively determining the identity of a user we suggest that the research requires further experimentation, enhancement and data before it would be considered sufficiently robust for use in a trial, despite our results demonstrating high degrees of accuracy and confidence. We present a number of areas of further work in this chapter and these can only improve the confidence and admissibility of the research.

We would consider a practical software implementation of the research presented in this thesis could be used by an investigator in testing the intersection of activity on the device with other forms of witness testimony, such as financial transactions, eyewitness testimony etc. For analysts, such a tool would provide an overview of the system, allowing them to highlight websites and periods of time that they would not necessarily be aware of, which is a substantial advantage of this kind of approach over a traditional approach of searching for known keywords only. The use of this kind of analysis may be useful in the courtroom as the community detection is visual and a prosecutor could put to a defendant "Each of these circles represents a period of time. Each of those circles that are the same colour that are connected together are very, very similar use of the computer. Are you telling me that it was some random person off the street using your computer rather than as the diagram shows it was someone that was using it the same way that it has been used all these times before?". Such an argument may be persuasive not only to the defendant such that they drop the pursuit of such a defence, but also to a jury. The dynamics of the use of expert evidence in court however beyond the scope of this thesis.

6.3.5 SEQUENTIAL ANALYSIS

From the outset of our research we identified that we could analyse our data sequentially or in aggregate. We have always held that the complexity of sequential analysis may be simplified if sufficiently similar aggregates can be identified, then sequentially parsed. The evidence we present in this thesis suggests that users are not sufficiently repetitive in their behaviour each time that they log in such that sequential analysis is even necessary (this is summarised below in section 6.4).

Internet history we contend may not lend itself to sequence analysis, but if using Context Session-to-Session Analysis on other types of data where there is greater predictability in the sequences such as with operating system log files, or cell phone call records, or cyber-physical systems where sensors can be activated in sessions but there are potentially multiple pathways through the sensors that could indicate idiosyncrasy, then we believe that there is scope for using sequential analysis alongside aggregate session analysis.

6.4 FINAL COMMENTS

This research project has not only produced the high-level concept of 'Context Analysis', which has value to researchers, but has produced a significant low-level concept, which we assert as having immediate value to the researcher, the analyst and the investigator: the formally defined 'Session' of Internet history.

Sessions can be calculated without knowing the ground truth of who was using the device, so analysts and Investigators can immediately use Session-to-Session analysis and the novel grouping schemes that we have developed during this research, to identify the components that have significance to the users without having to have domain-level knowledge of the users, their interest or even specific websites or keywords to search for. Such information can guide an analyst or investigator to periods of time that they may want to focus their enquiries upon.

Although ultimately the further work in this research for automating the detection of the individual users is outside the scope of the results in this thesis, we have demonstrated that with zero-knowledge about the users we can automatically detect websites/components that can discriminate between two different users that share a device, with a high degree of accuracy. This work and the methods we have developed should then form the basis of future research that we can pursue where there is domain-level knowledge about the potential users of the device and the types of websites that are being visited.

In addition to the practical analytical techniques that we have presented we have identified that there are significant questions about assessing the normality of Internet history and Sessions. We have extensively investigated the difference between a session where there is only a single repeating component, what we have referred to as S=1, and how there are many more of those kinds of sessions than the sessions where there are two or three repeating components, S=2, S=3 respectively. What we have seen is that although people have websites that they visit throughout the day, week etc. we see that they do not necessarily have a regular pattern of behaviour they will perform everytime they go online. This leads to a number of interesting questions that a researcher may want to pursue about those clusters of activity which we have called communities. Do those communities appear with the same frequency, size, components, time of day etc. across different datasets? That is to say, can we with sufficient data draw conclusions about those communities across all users? We have as part of our research into the validity of our datasets been able to show statistical correlation with the global norms for the appearance of those websites, but much finer levels of analysis would be a fascinating area of further research, that is now available because of the formal specification of sessions and Relative Popularity.

Consequently, the high-level concept of ‘Context Analysis’, and the low-level techniques and methods for splitting up the specific Internet history artefacts has gone a significant way to formally define the investigation of the user of a device at any particular time and establish if observable patterns of regularly occurring behaviour are present, which may be significant to analysts, investigators and future researchers of Internet history and Context Analysis.

REFERENCES

[1] Abraham, T. (2006). Event sequence mining to develop profiles for computer forensic investigation purposes. Proceedings of the 2006 Australasian workshops on Grid computing and e-research, 54: 145153.

[2] Al Awawdeh, S., Baggili, I., Marrington, A. and Iqbal, F. (2013). 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.

[3] Al Awadhi, I., Read, J.C., Marrington, A. and Franqueira, V.N. (2015). Factors Influencing Digital Forensic Investigations: Empirical Evaluation of 12 Years of Dubai Police Cases. The Journal of Digital Forensics, Security and Law: JDFSL, 10(4), p.7.

[4] Al Fahdi, M., Clarke, N.L. and Furnell, S.M. (2013). Challenges to digital forensics: A survey of researchers & practitioners attitudes and opinions. InInformation Security for South Africa, 2013 (pp. 18). IEEE.

[5] Amato, F., Cozzolino, G., Mazzeo, A. and Mazzocca, N. (2017). Correlation of Digital Evidences in Forensic Investigation through Semantic Technologies. In Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on (pp. 668-673). IEEE.

[6] Baggili, I., BaAbdallah, A., Al-Safi, D. and Marrington, A. (2012). Research trends in digital forensic science: an empirical analysis of published research. In International Conference on Digital Forensics and Cyber Crime (pp. 144-157). Springer Berlin Heidelberg.

[7] Bartlett, J., Norrie, R., Patel, S., Rumpel, R. and Wibberley, S. (2014). Misogyny on twitter. Demos.

[8] Baryamureeba, V. and Tushabe, F. (2004). The Enhanced Digital Investigation Process Model Digital Forensics Research Workshop.

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

[10] Boyd, C. and Forster, P. (2004). Time and date issues in forensic computinga case study. Digital Investigation, 1(1), pp.18-23.

[11] Brenner, S.W, Carrier, B. and Henninger J. (2004). Trojan Horse Defense in Cybercrime Cases. Santa Clara High Technology Law Journal 21.1.

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

[13] Buchholz, F. (2007). An Improved Clock Model for Translating Timestamps. Department of Computer Science, Technical Report: JUM-INFOSEC-TR-2007-001.

[14] Bunting, S. and Wei, W. (2006). EnCase Computer Forensics: The Official EnCE: EnCase? Certified Examiner Study Guide. John Wiley & Sons.

[15] Carbone, R. and Bean, C. (2011). Generating computer forensic super-timelines under Linux. SANS Reading Room. 1-136.

[16] Carney, M. and Rogers, M. (2004). The Trojan Made Me Do It: A First Step in Statistical Based Computer Forensics Event Reconstruction. International Journal of Digital Evidence (IJDE), 2(4).

[17] Carrier, B. and Spafford, E.H. (2003). Getting Physical with the Investigation Process International Journal of Digital Evidence. Fall 2003, Volume 2, Issue 2.

[18] Carrier, B.D. and Spafford, E.H. (2005). Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence. DFRWS.

[19] Carrier, B.D. and Spafford, E.H. (2006). Categories of digital investigation analysis techniques based on the computer history model. Digital Investigation, 3 (1), 121130. 2006.

[20] Casey, E. (2004). Digital Evidence and Computer Crime, 2nd Edition, Elsevier Academic Press.

[21] Casey, E. (2006). Cutting corners: Trading justice for cost savings. Digital investigation, 3(4), pp.185-186.

[22] Casey, E., Ferraro, M. and Nguyen, L. (2009). Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence. Journal of forensic sciences, 54(6), pp.1353-1364.

[23] Casey, E. (2011). The increasing need for automation and validation in digital forensics. Digital Investigation 7.3 103-104.

[24] Chabot, Y., Bertaux, A., Nicolle, C. and Kechadi, T. (2014). A complete formalized knowledge representation model for advanced digital forensics timeline analysis. Digital Investigation, 11: S95S105.

[25] Chabot Y., Bertaux A., Nicolle C. and Kechadi T. (2014). Automatic Timeline Construction For Computer Forensics Purposes. IEEE Joint Intelligence and Security Informatics Conference.

[26] Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D. and Kleinberg, J. (1999). Mining the Web's link structure. Computer, 32(8), pp.60-67.

[27] Chakrabarti, S., Dom, B., Gibson, D., Kumar, S.R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1998). Experiments in topic distillation. In ACM SIGIR workshop on Hypertext Information Retrieval on the Web.

[28] Chesney, E.J. (1939). The Concept of mens rea in the Criminal Law. Journal of Criminal Law and Criminology (1931-1951), 29(5), pp.627-644.

[29] Chisum, W.J. and Turvey, B. (2000). Evidence dynamics: Locards exchange principle & crime reconstruction. Journal of Behavioral Profiling, 1(1), pp.1-15.

[30] Dhami, M.K., Lundrigan, S. and Mueller-Johnson, K. (2015). Instructions on reasonable doubt: Defining the standard of proof and the jurors task. Psychology, Public Policy, and Law, 21(2), p.169.

[31] Dice, L.R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), pp.297-302.

[32] Ding, X. and Zou, H. (2011). Time Based Data Forensic and Cross-Reference Analysis. Proceedings of the 2011 Symposium on Applied Computing, ACM.

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

[34] Galbraith, C. and Smyth, P. (2017). Analyzing user-event data using score-based likelihood ratios with marked point processes. Digital Investigation, 22, pp.S106-S114.

[35] Garfinkel, S. (2013). Digital Forensics, American Scientist (Sept e Oct 2013, volume 101, number 5).

[36] Gladyshev, P. and Patel, A. (2005). Formalising Event Time Bounding in Digital Investigations. International Journal of Digital Evidence, 4(2): 114.

[37] Gogolin, G. (2010). The digital crime tsunami. Digital Investigation, 7(1), pp.3-8.

[38] Goss, J. and Gladyshev, P. (2010). Forensic triage: managing the risk. Master of Science, University College Dublin.

[39] Grajeda, C., Breitinger, F. and Baggili, I., (2017). Availability of datasets for digital forensics And what is missing. Digital Investigation, 22, pp.S94-S105.

[40] Gresty, D.W., Gan, D. and Loukas, G. (2014). Digital Forensic Analysis of Internet History Using Principal Component Analysis. Proceedings of the 15th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, pp.237-242, Liverpool, UK, June 23-24.

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

[42] Greenacre, M. and Primicerio, R. (2014). Multivariate analysis of ecological data. Fundacion BBVA. Chapter 7.

[43] Gudjonsson, K. (2010). InfoSec Reading Room Mastering the Super Timeline With log2timeline. P. 84.

[44] Harichandran, V.S., Breitinger, F., Baggili, I. and Marrington, A. (2016). A cyber forensics needs analysis survey: Revisiting the domain's needs a decade later. Computers & Security, 57, pp.1-13.

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

[46] Hamming, R.W. (1950). Error detecting and error correcting codes. Bell Labs Technical Journal, 29(2), pp.147-160.

[47] Hargreaves, C. and Patterson, J. (2012). An automated timeline reconstruction approach for digital forensic investigations. Digital Investigation, 9: S69S79.

[48] Ho, S.M., Kao, D. and Wu, W.Y. (2018). Following the breadcrumbs: Timestamp pattern identification for cloud forensics. Digital Investigation.

[49] Hoy, J. (2014). Forensic radio survey techniques for cell site analysis. John Wiley & Sons.

[50] Irons, A.D., Stephens, P. and Ferguson, R.I. (2009). Digital Investigation as a distinct discipline: A pedagogic perspective. Digital Investigation, 6(1), pp.82-90.

[51] Irons, A. and Lallie, H.S. (2014). Digital forensics to intelligent forensics. Future Internet, 6(3), pp.584-596.

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

[53] James, J., Gladyshev, P., Abdullah, M. and Zhu, Y. (2010). Analysis of evidence using formal event reconstruction. Digital Forensics and Cyber Crime, 31: 8598.

[54] James, J.I. and Gladyshev, P. (2013). Challenges with automation in digital forensic investigations. arXiv preprint arXiv:1303.4498.

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

[56] James, J.I. and Jang, Y. (2017). Inferring Action Instances with No Prior Knowledge in Digital Investigations. International Information Institute (Tokyo). Information, 20(6A), pp.4153-4162.

[57] Kalber, S., Dewald, A. and Idler, S. (2014). Forensic Zero-Knowledge Event Reconstruction on Filesystem Metadata. Sicherheit, 331-343.

[58] Kaye, D.H. (2009). Probability, individualization, and uniqueness in forensic science evidence: Listening to the academies.

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

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

[61] Kirchler, M., Herrmann, D., Lindemann, J. and Kloft, M. (2016). Tracked without a trace: linking sessions of users by unsupervised learning of patterns in their DNS traffic. In Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security (pp. 23-34). ACM.

[62] Kohn, M., Olivier, M.S. and Eloff, J.H. (2006). Framework for a Digital Forensic Investigation. In ISSA (pp. 1-7).

[63] Kim, S.J. and Lee, S.H. (2002). An improved computation of the pagerank algorithm. In European Conference on Information Retrieval (pp. 73-85). Springer Berlin Heidelberg.

[64] Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).

[65] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W.Y. (2008). Mining user similarity based on location history. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in geographic information systems, ACM.

[66] Lillis, D. and Scanlon, M. (2016). On the Benefits of Information Retrieval and Information Extraction Techniques Applied to Digital Forensics. In Advanced Multimedia and Ubiquitous Engineering (pp. 641-647). Springer Singapore.

[67] Lundrigan, S. The verdict on reasonable doubt, Barrister Magazine. http://www.barristermagazine.com/barrister/index.php?id=562

[68] Lyman, P. (2002). Archiving the world wide web. Building a national strategy for digital preservation: Issues in digital media archiving, pp.38-51.

[69] Ma, H., Cao, H., Yang, Q., Chen, E., and Tian, J. (2012). A habit mining approach for discovering similar mobile users. Proceedings of the 21st international conference on World Wide Web.

[70] Marrington, A., Mohay, G., Clark, A., and Morarji, H. (2007). Event-based computer profiling for the forensic reconstruction of computer activity. AusCERT2007 R&D Stream 71: 7187.

[71] Marrington, A. (2009). Computer Profiling for Forensic Purposes. Queensland University of Technology.

[72] Marrington, A., Baggili, I., Mohay, G., and Clark, A. (2011). CAT Detect (Computer Activity Timeline Detection): A tool for detecting inconsistency in computer activity timelines. Digital Investigation, 8, S52S61. 2011.

[73] Miller v Minister of Pensions (1947). 2 All ER 372.

[74] Mislan, R.P., Casey, E. and Kessler, G.C. (2010). The growing need for on-scene triage of mobile devices. Digital Investigation, 6(3), pp.112-124.

[75] Oh, J., Lee, S. and Lee, S. (2011). Advanced evidence collection and analysis of web browser activity. digital investigation, 8, pp. S62-S70.

[76] Olsson, J. and Boldt, M. (2009). Computer forensic timeline visualization tool. Digital Investigation, 6: S78S87.

[77] Omychund v Barker (1745). 1 Atk, 21, 49; 26 ER 15, 33.

[78] Palmer, G. (2001). A Road Map for Digital Forensic Research: Report from the First Digital Forensic Research Workshop (DFRWS). Utica, New York.

[79] Palmer, I., Gelfand, B. and Campbell, R. (2017). Exploring Digital Evidence with Graph Theory. Annual ADFSL Conference on Digital Forensics, Security and Law. 9.

[80] Poland, W.C. (1954). Criminal Procedure-Proof of Corpus Delicti by Circumstantial Evidence. Wm. & Mary Rev. Va. L., 2, p.170.

[81] Pollitt, M. (1995). Computer forensics: An approach to evidence in cyberspace. In Proceedings of the National Information Systems Security Conference (Vol. 2, pp. 487-491).

[82] Raghavan, S. and Raghavan, S.V. (2013). Determining the Origin of Downloaded Files Using Metadata Associations. Journal of Communications, 8(12): 902910.

[83] Raghavan, S. and Raghavan, S.V. (2013). AssocGEN: Engine for analyzing metadata based associations in digital evidence. 8th International Workshop on Systematic Approaches to Digital Forensics Engineering (SADFE).

[84] Raghavan, S. and Saran, H. (2013) UniTIME: Timestamp interpretation engine for developing unified timelines. 8th International Workshop on Systematic Approaches to Digital Forensics Engineering.

[85] Reith, M., Carr, C. and Gunsch, G. (2002). An Examination of Digital Forensic Models, International Journal of Digital Evidence. Fall 2002, Volume 1, Issue 3.

[86] Rogers, M.K., Goldman, J., Mislan, R., Wedge, T. and Debrota, S. (2006). Computer forensics field triage process model. In Proceedings of the conference on Digital Forensics, Security and Law (p. 27). Association of Digital Forensics, Security and Law.

[87] Rogers, M.K. and Siegfried, K. (2004). The future of computer forensics: a needs analysis survey. Computers & Security, 23(1), pp.12-16.

[88] Rowe, N.C. (2012). Testing the national software reference library. Digital Investigation, 9, pp.S131-S138.

[89] Rowe, N.C. and Garfinkel, S.L. (2012). Finding anomalous and suspicious files from directory metadata on a large corpus. Digital Forensics and Cyber Crime, Springer Berlin Heidelberg, 115-130.

[90] Schaefer, M., Wanner, F., Mansmann, F., Scheible, C., Stennett, V., Hasselrot, A.T. and Keim, D.A. (2011). Visual pattern discovery in timed event data. In IS&T/SPIE Electronic Imaging (pp. 78680K-78680K). International Society for Optics and Photonics.

[91] Schatz, B., Mohay, G., and Clark, A. (2006) A correlation method for establishing provenance of timestamps in digital evidence. Digital Investigation, 3: 98-107.

[92] Smith, D.M. (2011). Hype cycle for cloud computing. Gartner Inc., Stamford, 71.

[93] Stevens, M.W., (2005). Unification of relative time frames for digital forensics. Digital Investigation, 1(3):225239.

[94] Stoll, C. (1988). Stalking the wily hacker. Communications of the ACM, 31(5), pp.484-497.

[95] Taylor, M., Haggerty, J., Gresty, D. and Hegarty, R. (2010). Digital evidence in cloud computing systems. Computer Law & Security Review, 26(3), pp.304-308.

[96] Taylor, M., Haggerty, J., Gresty, D. and Lamb, D. (2011). Forensic investigation of cloud computing systems. Network Security, 2011(3), pp.4-10.

[97] Thompson, W.C., Vuille, J., Biedermann, A. and Taroni, F. (2013). The role of prior probability in forensic assessments. Frontiers in genetics, 4, p.220.

[98] Thurstone, L.L. (1929). The measurement of psychological value. Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court, pp.157-174.

[99] Van Baar, R.B., van Beek, H.M.A., van Eijk, E.J. (2014). Digital Forensics as a Service: A game changer. Digital Investigation 11 (2014) S54S62

[100] Wang, P., Wang, H., Liu, M. and Wang, W. (2010). An algorithmic approach to event summarization. Proceedings of ACM SIGMOD International Conference on Management of data, ACM.

[101] Willassen, S.Y. (2008). Timestamp evidence correlation by model based clock hypothesis testing. Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop, ICST.

[102] Willassen, S.Y. (2008). Finding Evidence of Antedating in Digital Investigations. Proceedings of the 2008 Third International Conference on Availability, Reliability and Security, pp. 26-32.

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

[104] Ye, Y., Zheng, Y., Chen, Y., Feng, J. and Xie, X. (2009). Mining individual life pattern based on location history. IEEE International Conference on Mobile Data Management.

[105] The Association of Chief Police Officers of England, Wales and N. Ireland, ACPO (2012). Good practice guide for computer based electronic evidence, Version 5. ¡http://www.acpo.police.uk¿.

[106] The Association of Chief Police Officers of England, Wales and N. Ireland, ACPO (2009). ACPO Managers Guide: Good Practice and Advice Guide for Managers of e-Crime Investigation, V0.1.4. http://www.digital-detective.net/digital-forensics-documents/ACPO_Good_Practice_and_Advice_for_Manager_of_e-Crime-Investigation.pdf

[107] Computer Misuse Act 1990. C.18, Section 1.

[108] Criminal Justice Act 1967. C.80, Section 8.

[109] Criminal Justice Act 1988, C.33, Section 160.

[110] Investigatory Powers Act 2016. (c.25).

[111] Protection of Children Act 1978, C.37, Section 1.

[112] Sexual Offences Act 2003, C.42, Section 15.

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

[114] "Pistorius trial: Prosecutor outlines '13 inconsistencies"', http://www.bbc.co.uk/news/world-africa-28686756

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

[116] Forensic Tool Kit (FTK), http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk

[117] "Vincent Tabak researched killings and sentences after Joanna Yeatess death", 19th October 2011. https://www.theguardian.com/uk/2011/oct/19/vincent-tabak-joannayeates-death

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

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

[120] Safwi W. A. Forensic Sciences & Criminology. 2011. http://www.legalservicesindia.com/article/article/forensic-sciences-&-criminology-536-1.html

Appendix A

RESULTS AND TABLES

We have performed our analysis on the following criteria:

- Nodes - The number of sessions in the graph.
- Good Com - The number of communities that contain no edges that are incorrect.
- Bad Com - The number of communities that contain one or more edges that are incorrect.
- Total Correct % - The percentage of edges within the graph that have nodes that correctly match to the same user.
- Total Incorrect % - The percentage of edges within the graph that have nodes that incorrectly as belonging to the same user.
- Intra Correct % - The percentage of edges that reside within communities that are correct.
- C2C Correct % - The percentage of edges that cross community-to-community that are correct.
- Intra Incorrect % - The percentage of edges that reside within communities that are incorrect.

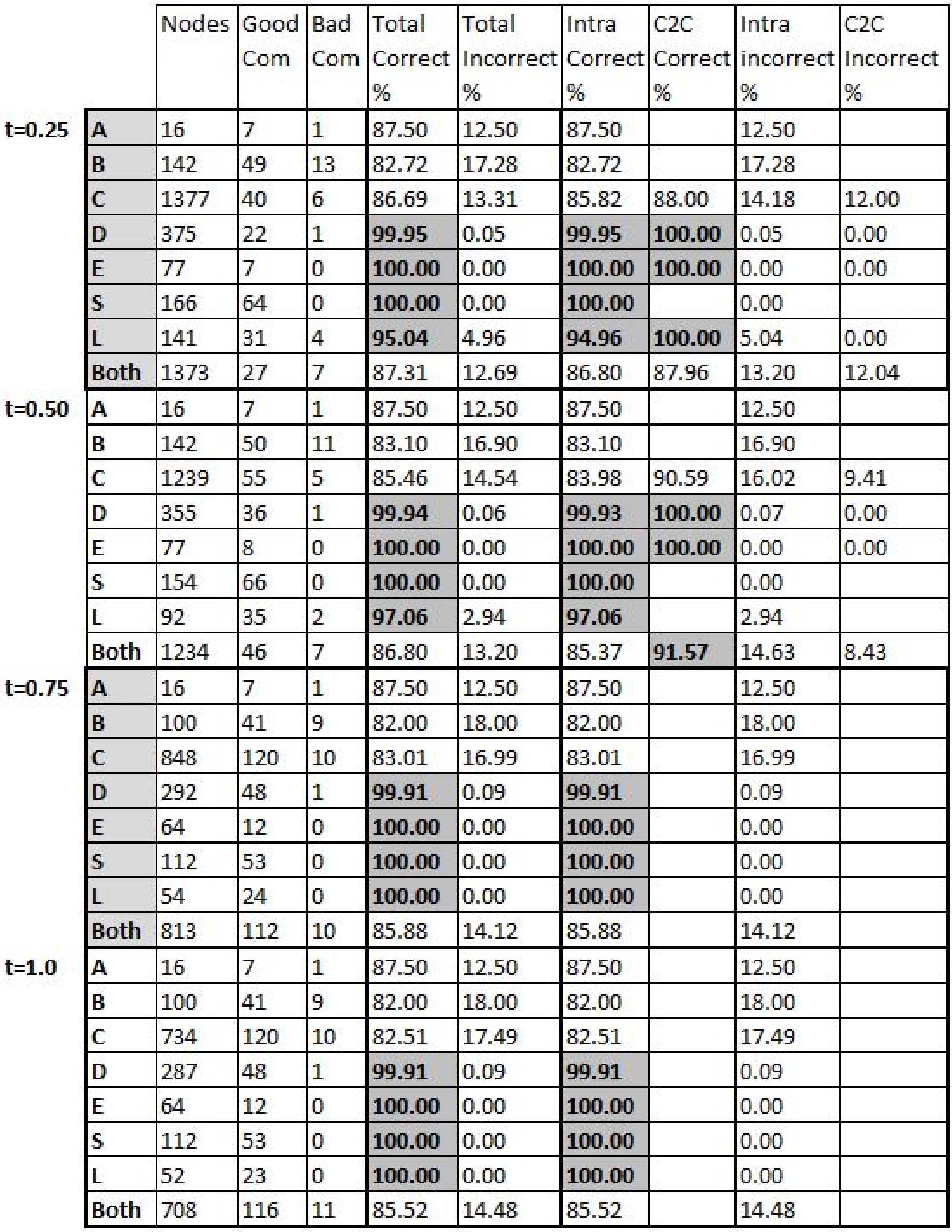
- C2C Incorrect % - The percentage of edges that cross community-to-community that are incorrect.

And with the Y dataset where we have provided details relating to the Weighted, Unweighted and 2 resolutions, so we have also explicitly stated:

- Com - Total number of communities in the graph
- Good Com % - The percentage of communities that contain no edges that are incorrect.
- Bad Com % - The percentage of communities that contain one or more edges that are incorrect.

A.0.1 *W* DATASET UNWEIGHTED, RESOLUTION 1.0

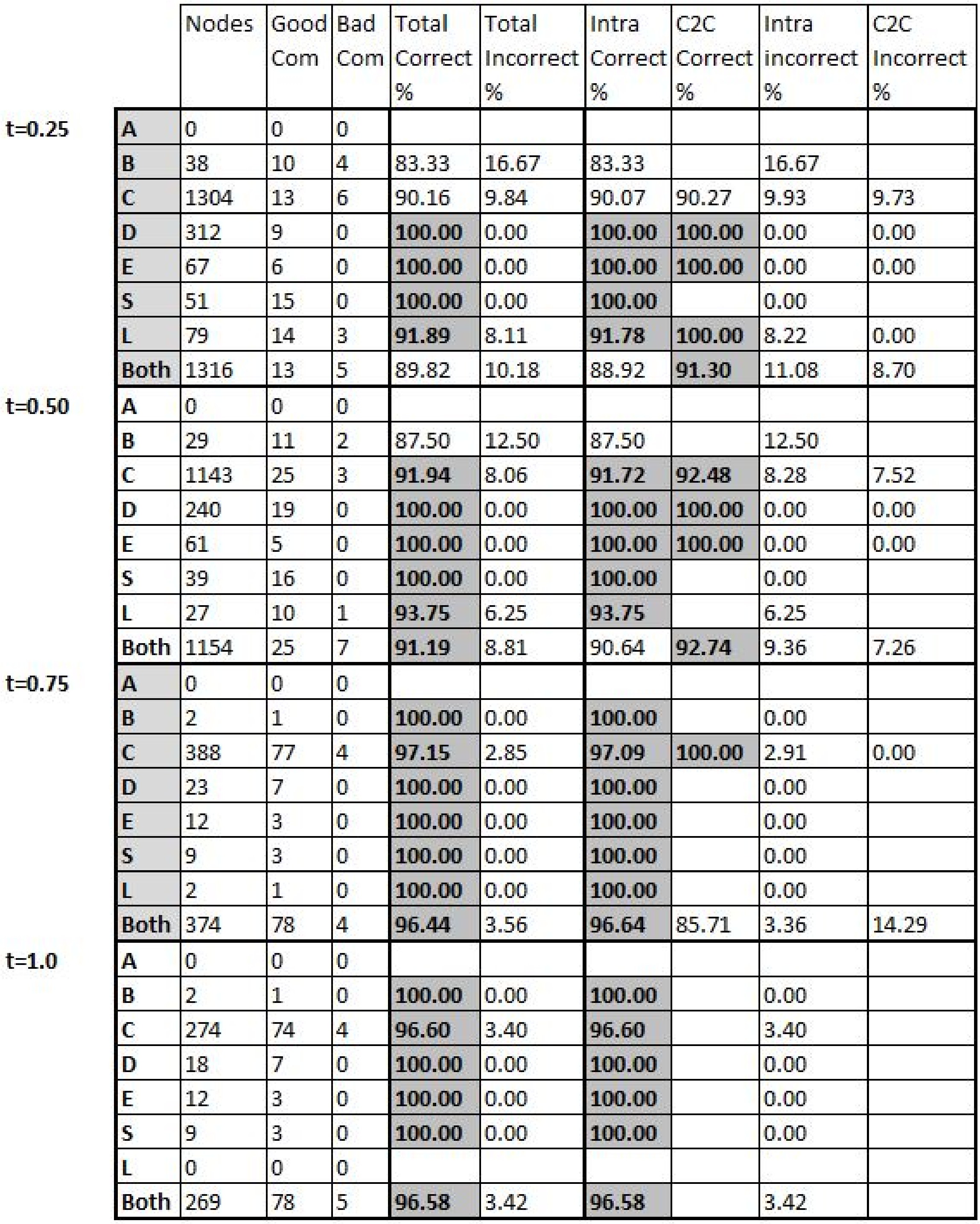
S-val **S=1**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	16	7	1	87.50	12.50	87.50		12.50	
	B	142	49	13	82.72	17.28	82.72		17.28	
	C	1377	40	6	86.69	13.31	85.82	88.00	14.18	12.00
	D	375	22	1	**99.95**	0.05	**99.95**	**100.00**	0.05	0.00
	E	77	7	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	166	64	0	**100.00**	0.00	**100.00**		0.00	
	L	141	31	4	**95.04**	4.96	**94.96**	**100.00**	5.04	0.00
	Both	1373	27	7	87.31	12.69	86.80	87.96	13.20	12.04
t=0.50	A	16	7	1	87.50	12.50	87.50		12.50	
	B	142	50	11	83.10	16.90	83.10		16.90	
	C	1239	55	5	85.46	14.54	83.98	90.59	16.02	9.41
	D	355	36	1	**99.94**	0.06	**99.93**	**100.00**	0.07	0.00
	E	77	8	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	154	66	0	**100.00**	0.00	**100.00**		0.00	
	L	92	35	2	**97.06**	2.94	**97.06**		2.94	
	Both	1234	46	7	86.80	13.20	85.37	**91.57**	14.63	8.43
t=0.75	A	16	7	1	87.50	12.50	87.50		12.50	
	B	100	41	9	82.00	18.00	82.00		18.00	
	C	848	120	10	83.01	16.99	83.01		16.99	
	D	292	48	1	**99.91**	0.09	**99.91**		0.09	
	E	64	12	0	**100.00**	0.00	**100.00**		0.00	
	S	112	53	0	**100.00**	0.00	**100.00**		0.00	
	L	54	24	0	**100.00**	0.00	**100.00**		0.00	
	Both	813	112	10	85.88	14.12	85.88		14.12	
t=1.0	A	16	7	1	87.50	12.50	87.50		12.50	
	B	100	41	9	82.00	18.00	82.00		18.00	
	C	734	120	10	82.51	17.49	82.51		17.49	
	D	287	48	1	**99.91**	0.09	**99.91**		0.09	
	E	64	12	0	**100.00**	0.00	**100.00**		0.00	
	S	112	53	0	**100.00**	0.00	**100.00**		0.00	
	L	52	23	0	**100.00**	0.00	**100.00**		0.00	
	Both	708	116	11	85.52	14.48	85.52		14.48	

Figure A.1: *W* Dataset Results - BRD of 91% Highlighted

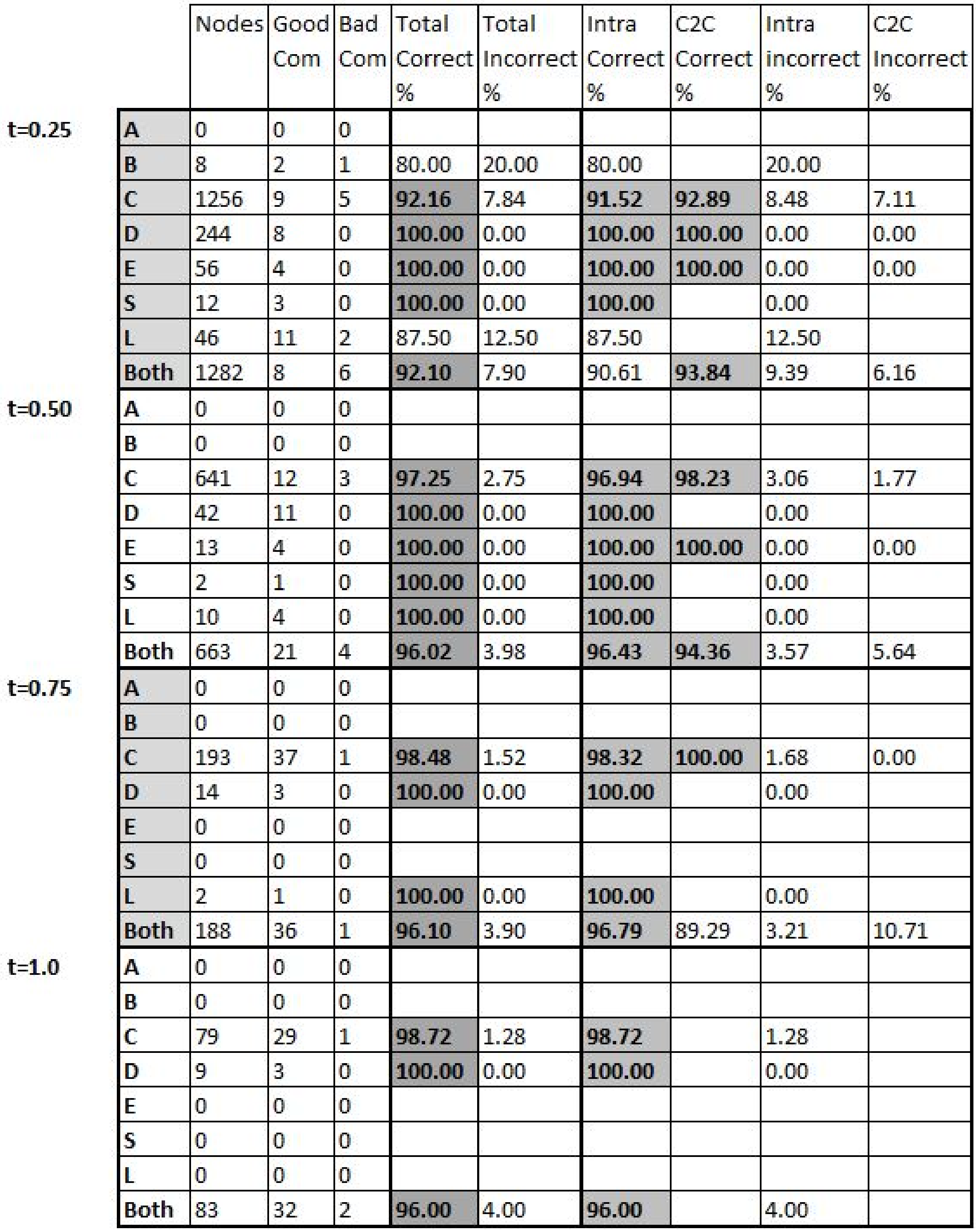
S-val **S=2**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	0	0	0						
	B	38	10	4	83.33	16.67	83.33		16.67	
	C	1304	13	6	90.16	9.84	90.07	90.27	9.93	9.73
	D	312	9	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	67	6	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	51	15	0	**100.00**	0.00	**100.00**		0.00	
	L	79	14	3	**91.89**	8.11	**91.78**	**100.00**	8.22	0.00
	Both	1316	13	5	89.82	10.18	88.92	**91.30**	11.08	8.70
t=0.50	**A**	0	0	0						
	B	29	11	2	87.50	12.50	87.50		12.50	
	C	1143	25	3	**91.94**	8.06	**91.72**	**92.48**	8.28	7.52
	D	240	19	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	61	5	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	39	16	0	**100.00**	0.00	**100.00**		0.00	
	L	27	10	1	**93.75**	6.25	**93.75**		6.25	
	Both	1154	25	7	**91.19**	8.81	90.64	**92.74**	9.36	7.26
t=0.75	**A**	0	0	0						
	B	2	1	0	**100.00**	0.00	**100.00**		0.00	
	C	388	77	4	**97.15**	2.85	**97.09**	**100.00**	2.91	0.00
	D	23	7	0	**100.00**	0.00	**100.00**		0.00	
	E	12	3	0	**100.00**	0.00	**100.00**		0.00	
	S	9	3	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	0	**100.00**	0.00	**100.00**		0.00	
	Both	374	78	4	**96.44**	3.56	**96.64**	85.71	3.36	14.29
t=1.0	**A**	0	0	0						
	B	2	1	0	**100.00**	0.00	**100.00**		0.00	
	C	274	74	4	**96.60**	3.40	**96.60**		3.40	
	D	18	7	0	**100.00**	0.00	**100.00**		0.00	
	E	12	3	0	**100.00**	0.00	**100.00**		0.00	
	S	9	3	0	**100.00**	0.00	**100.00**		0.00	
	L	0	0	0						
	Both	269	78	5	**96.58**	3.42	**96.58**		3.42	

Figure A.2: *W* Dataset Results - BRD of 91% Highlighted

S-val **S=3**

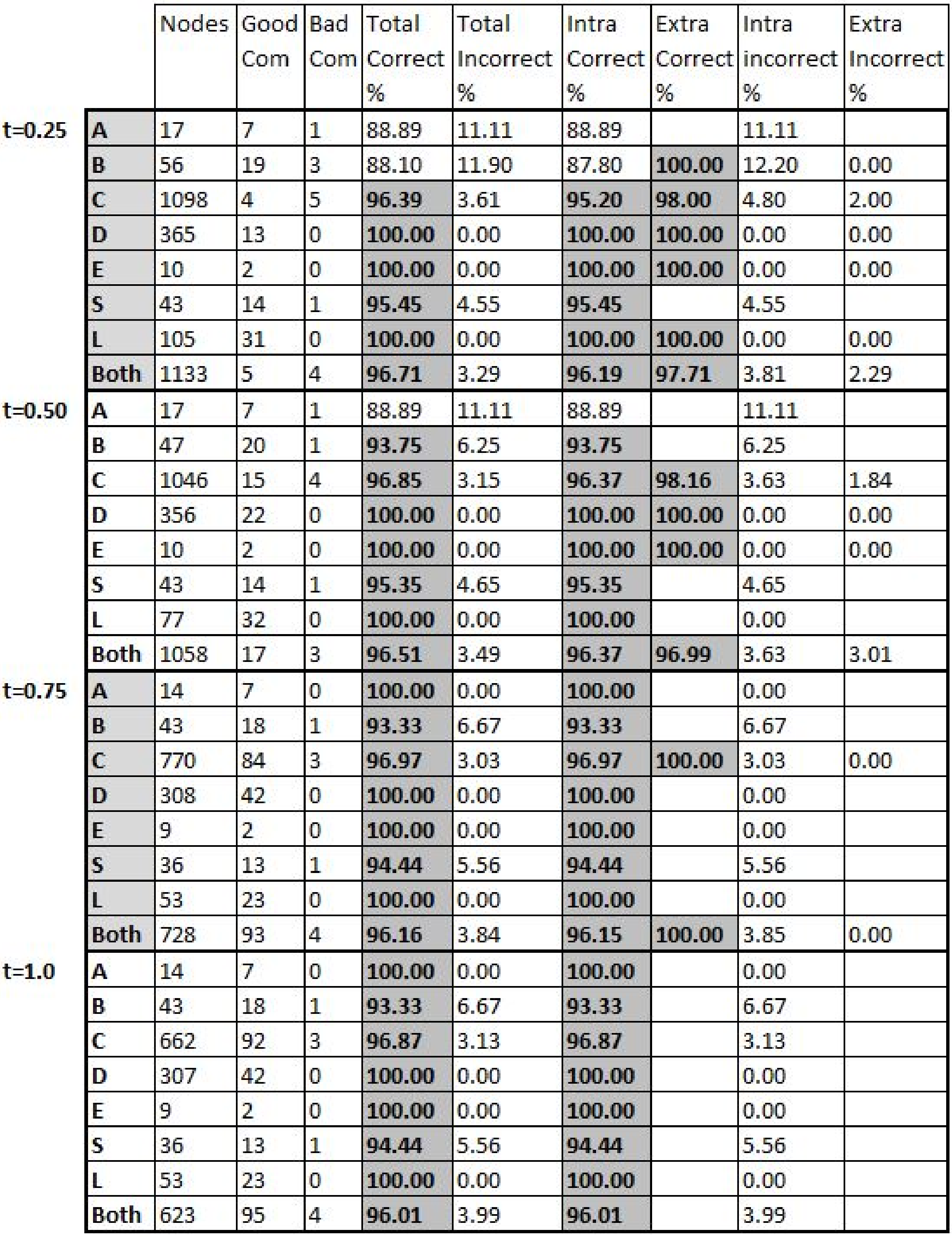


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	0	0	0						
	B	8	2	1	80.00	20.00	80.00		20.00	
	C	1256	9	5	**92.16**	7.84	**91.52**	**92.89**	8.48	7.11
	D	244	8	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	56	4	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	12	3	0	**100.00**	0.00	**100.00**		0.00	
	L	46	11	2	87.50	12.50	87.50		12.50	
	Both	1282	8	6	**92.10**	7.90	90.61	**93.84**	9.39	6.16
t=0.50	**A**	0	0	0						
	B	0	0	0						
	C	641	12	3	**97.25**	2.75	**96.94**	**98.23**	3.06	1.77
	D	42	11	0	**100.00**	0.00	**100.00**		0.00	
	E	13	4	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	2	1	0	**100.00**	0.00	**100.00**		0.00	
	L	10	4	0	**100.00**	0.00	**100.00**		0.00	
	Both	663	21	4	**96.02**	3.98	**96.43**	**94.36**	3.57	5.64
t=0.75	**A**	0	0	0						
	B	0	0	0						
	C	193	37	1	**98.48**	1.52	**98.32**	**100.00**	1.68	0.00
	D	14	3	0	**100.00**	0.00	**100.00**		0.00	
	E	0	0	0						
	S	0	0	0						
	L	2	1	0	**100.00**	0.00	**100.00**		0.00	
	Both	188	36	1	**96.10**	3.90	**96.79**	89.29	3.21	10.71
t=1.0	**A**	0	0	0						
	B	0	0	0						
	C	79	29	1	**98.72**	1.28	**98.72**		1.28	
	D	9	3	0	**100.00**	0.00	**100.00**		0.00	
	E	0	0	0						
	S	0	0	0						
	L	0	0	0						
	Both	83	32	2	**96.00**	4.00	**96.00**		4.00	

Figure A.3: *W* Dataset Results - BRD of 91% Highlighted

A.0.2 *X* DATASET UNWEIGHTED, RESOLUTION 1.0

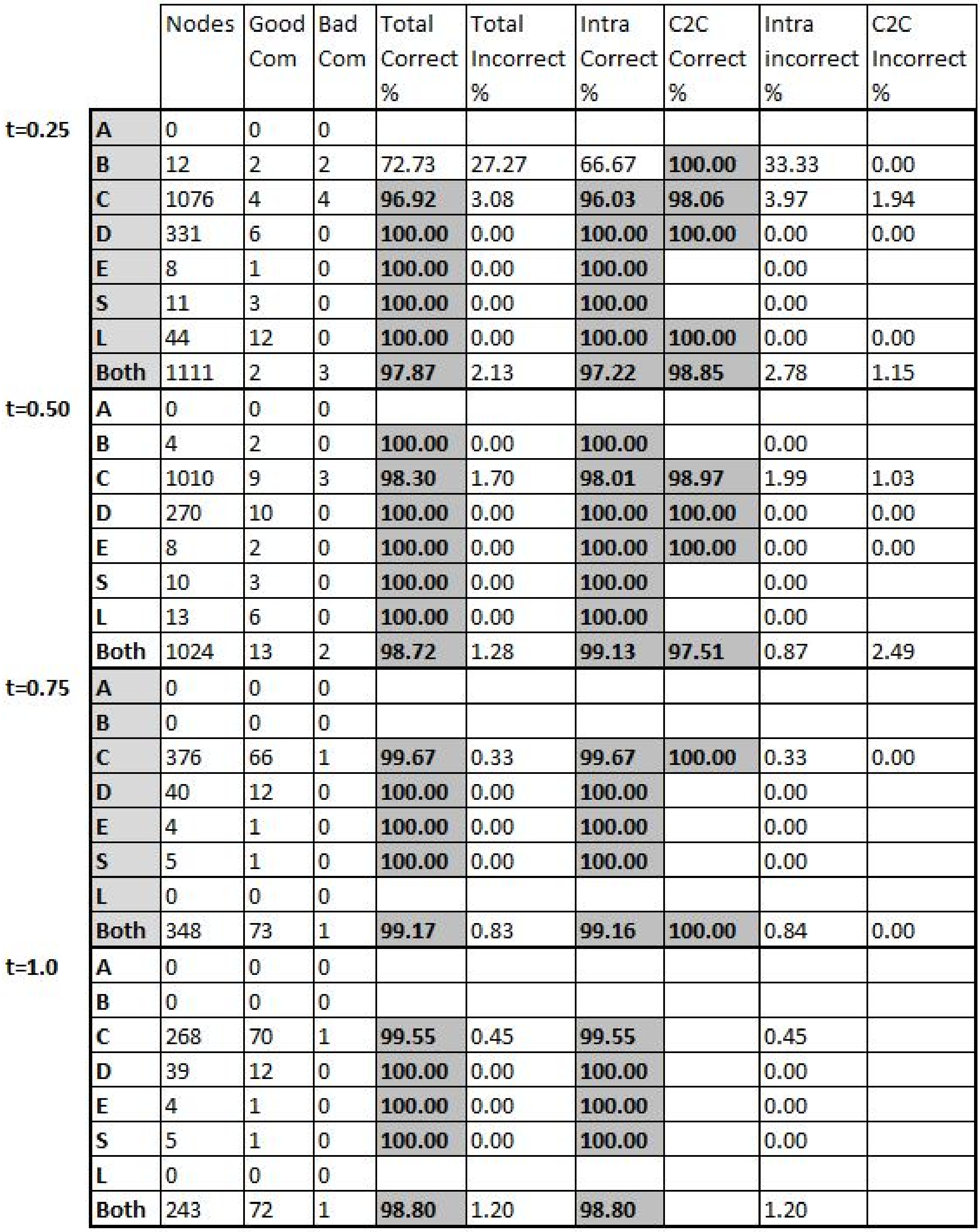
S-val **S=1**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	Extra Correct %	Intra incorrect %	Extra Incorrect %
t=0.25	A	17	7	1	88.89	11.11	88.89		11.11	
	B	56	19	3	88.10	11.90	87.80	100.00	12.20	0.00
	C	1098	4	5	96.39	3.61	95.20	98.00	4.80	2.00
	D	365	13	0	100.00	0.00	100.00	100.00	0.00	0.00
	E	10	2	0	100.00	0.00	100.00	100.00	0.00	0.00
	S	43	14	1	95.45	4.55	95.45		4.55	
	L	105	31	0	100.00	0.00	100.00	100.00	0.00	0.00
	Both	1133	5	4	96.71	3.29	96.19	97.71	3.81	2.29
t=0.50	A	17	7	1	88.89	11.11	88.89		11.11	
	B	47	20	1	93.75	6.25	93.75		6.25	
	C	1046	15	4	96.85	3.15	96.37	98.16	3.63	1.84
	D	356	22	0	100.00	0.00	100.00	100.00	0.00	0.00
	E	10	2	0	100.00	0.00	100.00	100.00	0.00	0.00
	S	43	14	1	95.35	4.65	95.35		4.65	
	L	77	32	0	100.00	0.00	100.00		0.00	
	Both	1058	17	3	96.51	3.49	96.37	96.99	3.63	3.01
t=0.75	A	14	7	0	100.00	0.00	100.00		0.00	
	B	43	18	1	93.33	6.67	93.33		6.67	
	C	770	84	3	96.97	3.03	96.97	100.00	3.03	0.00
	D	308	42	0	100.00	0.00	100.00		0.00	
	E	9	2	0	100.00	0.00	100.00		0.00	
	S	36	13	1	94.44	5.56	94.44		5.56	
	L	53	23	0	100.00	0.00	100.00		0.00	
	Both	728	93	4	96.16	3.84	96.15	100.00	3.85	0.00
t=1.0	A	14	7	0	100.00	0.00	100.00		0.00	
	B	43	18	1	93.33	6.67	93.33		6.67	
	C	662	92	3	96.87	3.13	96.87		3.13	
	D	307	42	0	100.00	0.00	100.00		0.00	
	E	9	2	0	100.00	0.00	100.00		0.00	
	S	36	13	1	94.44	5.56	94.44		5.56	
	L	53	23	0	100.00	0.00	100.00		0.00	
	Both	623	95	4	96.01	3.99	96.01		3.99	

Figure A.4: *X* Dataset Results - BRD of 91% Highlighted

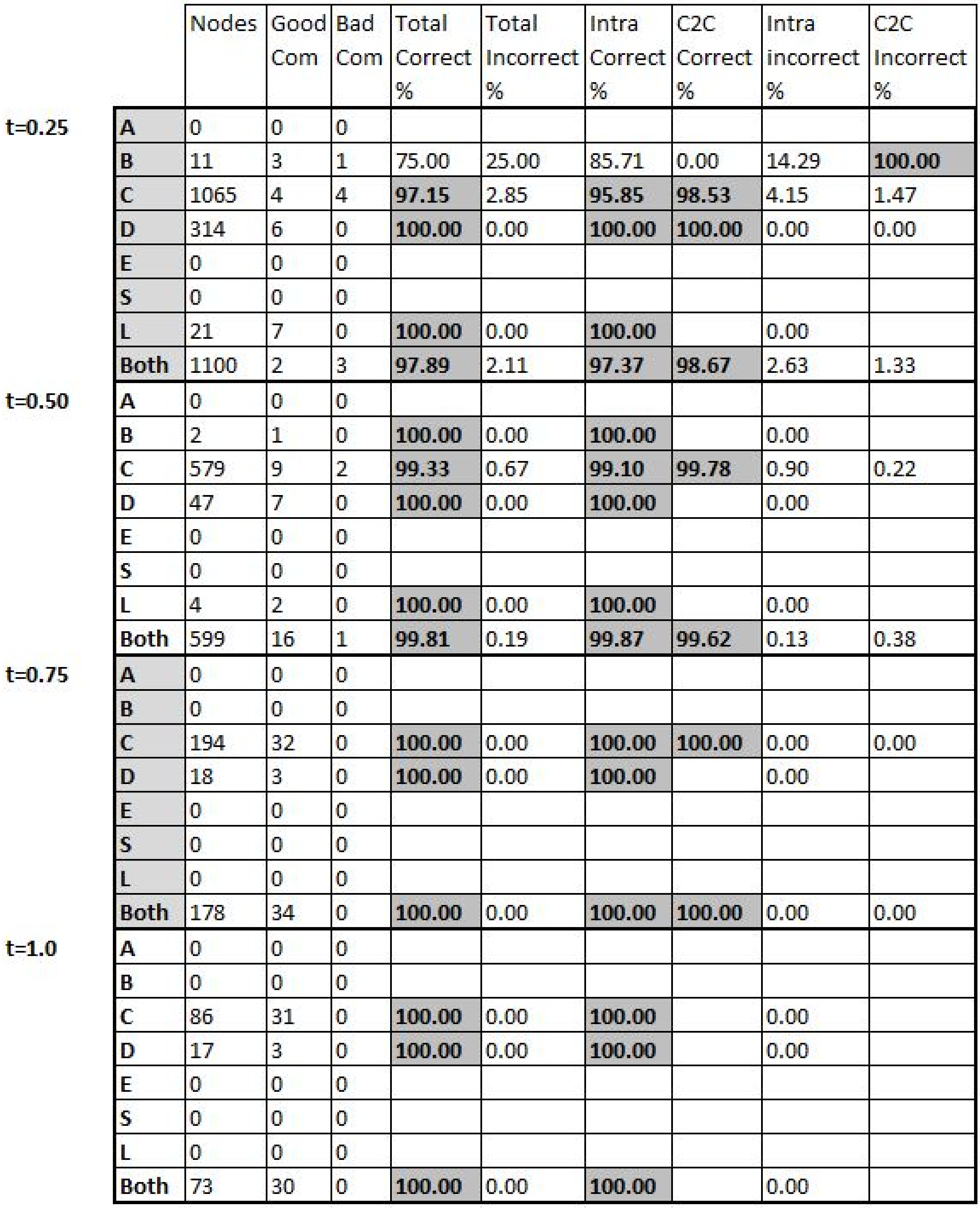
S-val **S=2**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	0	0	0						
	B	12	2	2	72.73	27.27	66.67	**100.00**	33.33	0.00
	C	1076	4	4	**96.92**	3.08	**96.03**	**98.06**	3.97	1.94
	D	331	6	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	8	1	0	**100.00**	0.00	**100.00**		0.00	
	S	11	3	0	**100.00**	0.00	**100.00**		0.00	
	L	44	12	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	Both	1111	2	3	**97.87**	2.13	**97.22**	**98.85**	2.78	1.15
t=0.50	**A**	0	0	0						
	B	4	2	0	**100.00**	0.00	**100.00**		0.00	
	C	1010	9	3	**98.30**	1.70	**98.01**	**98.97**	1.99	1.03
	D	270	10	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	8	2	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	10	3	0	**100.00**	0.00	**100.00**		0.00	
	L	13	6	0	**100.00**	0.00	**100.00**		0.00	
	Both	1024	13	2	**98.72**	1.28	**99.13**	**97.51**	0.87	2.49
t=0.75	**A**	0	0	0						
	B	0	0	0						
	C	376	66	1	**99.67**	0.33	**99.67**	**100.00**	0.33	0.00
	D	40	12	0	**100.00**	0.00	**100.00**		0.00	
	E	4	1	0	**100.00**	0.00	**100.00**		0.00	
	S	5	1	0	**100.00**	0.00	**100.00**		0.00	
	L	0	0	0						
	Both	348	73	1	**99.17**	0.83	**99.16**	**100.00**	0.84	0.00
t=1.0	**A**	0	0	0						
	B	0	0	0						
	C	268	70	1	**99.55**	0.45	**99.55**		0.45	
	D	39	12	0	**100.00**	0.00	**100.00**		0.00	
	E	4	1	0	**100.00**	0.00	**100.00**		0.00	
	S	5	1	0	**100.00**	0.00	**100.00**		0.00	
	L	0	0	0						
	Both	243	72	1	**98.80**	1.20	**98.80**		1.20	

Figure A.5: *X* Dataset Results - BRD of 91% Highlighted

S-val **S=3**

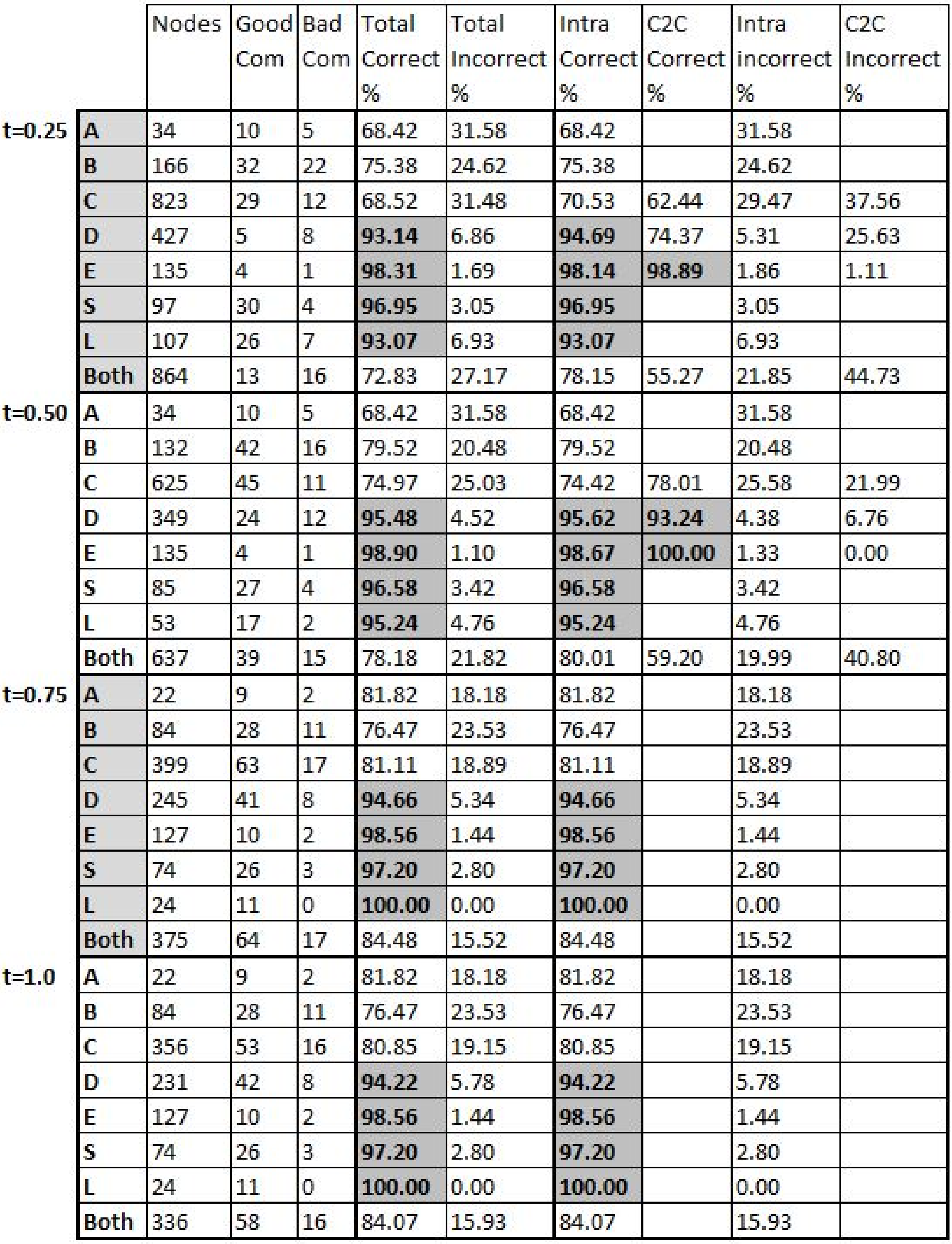


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	0	0	0						
	B	11	3	1	75.00	25.00	85.71	0.00	14.29	**100.00**
	C	1065	4	4	**97.15**	2.85	**95.85**	**98.53**	4.15	1.47
	D	314	6	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	0	0	0						
	S	0	0	0						
	L	21	7	0	**100.00**	0.00	**100.00**		0.00	
	Both	1100	2	3	**97.89**	2.11	**97.37**	**98.67**	2.63	1.33
t=0.50	**A**	0	0	0						
	B	2	1	0	**100.00**	0.00	**100.00**		0.00	
	C	579	9	2	**99.33**	0.67	**99.10**	**99.78**	0.90	0.22
	D	47	7	0	**100.00**	0.00	**100.00**		0.00	
	E	0	0	0						
	S	0	0	0						
	L	4	2	0	**100.00**	0.00	**100.00**		0.00	
	Both	599	16	1	**99.81**	0.19	**99.87**	**99.62**	0.13	0.38
t=0.75	**A**	0	0	0						
	B	0	0	0						
	C	194	32	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	D	18	3	0	**100.00**	0.00	**100.00**		0.00	
	E	0	0	0						
	S	0	0	0						
	L	0	0	0						
	Both	178	34	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
t=1.0	**A**	0	0	0						
	B	0	0	0						
	C	86	31	0	**100.00**	0.00	**100.00**		0.00	
	D	17	3	0	**100.00**	0.00	**100.00**		0.00	
	E	0	0	0						
	S	0	0	0						
	L	0	0	0						
	Both	73	30	0	**100.00**	0.00	**100.00**		0.00	

Figure A.6: *X* Dataset Results - BRD of 91% Highlighted

A.0.3 *Z* DATASET UNWEIGHTED, RESOLUTION 1.0

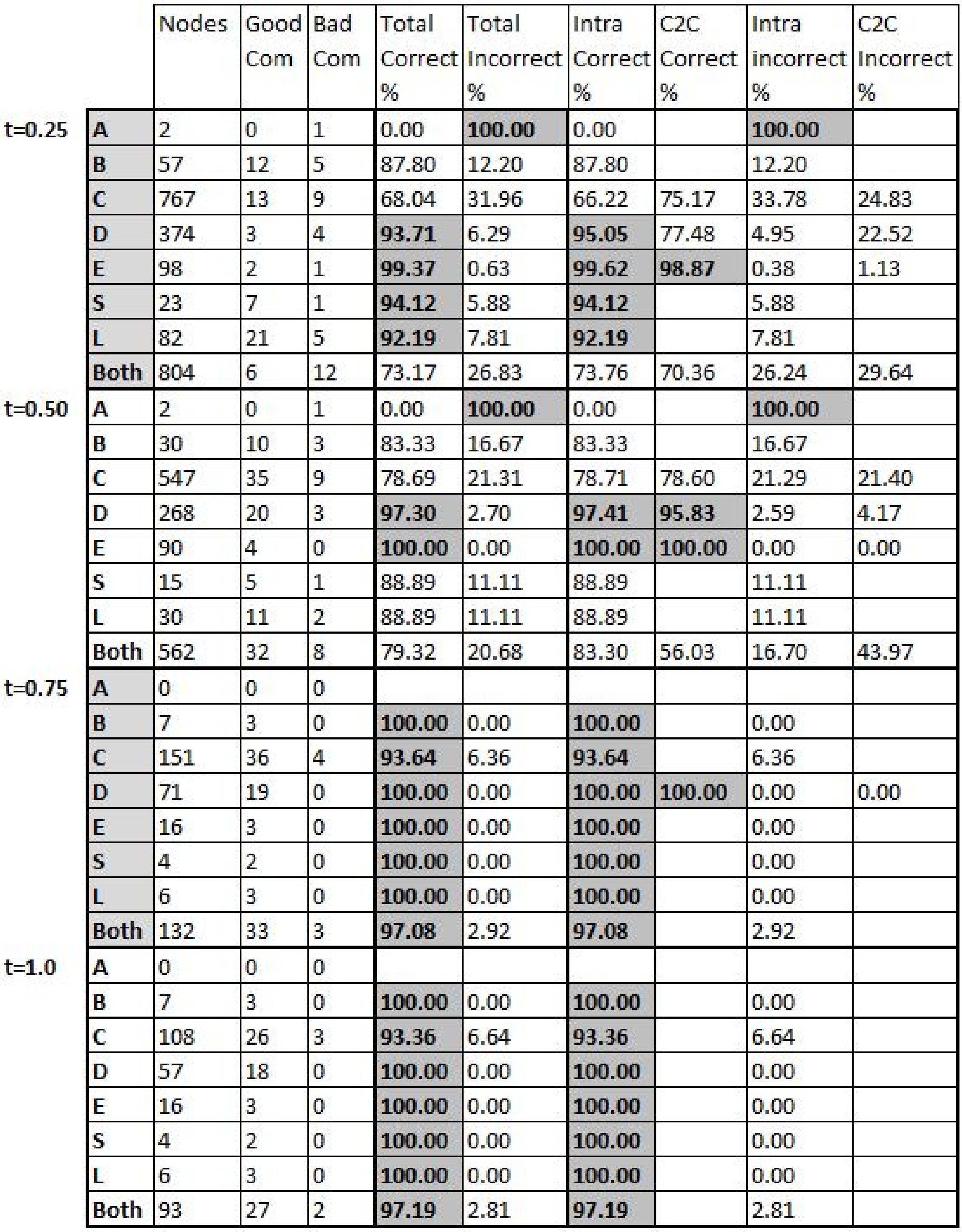
S-val **S=1**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	34	10	5	68.42	31.58	68.42		31.58	
	B	166	32	22	75.38	24.62	75.38		24.62	
	C	823	29	12	68.52	31.48	70.53	62.44	29.47	37.56
	D	427	5	8	**93.14**	6.86	**94.69**	74.37	5.31	25.63
	E	135	4	1	**98.31**	1.69	**98.14**	**98.89**	1.86	1.11
	S	97	30	4	**96.95**	3.05	**96.95**		3.05	
	L	107	26	7	**93.07**	6.93	**93.07**		6.93	
	Both	864	13	16	72.83	27.17	78.15	55.27	21.85	44.73
t=0.50	A	34	10	5	68.42	31.58	68.42		31.58	
	B	132	42	16	79.52	20.48	79.52		20.48	
	C	625	45	11	74.97	25.03	74.42	78.01	25.58	21.99
	D	349	24	12	**95.48**	4.52	**95.62**	**93.24**	4.38	6.76
	E	135	4	1	**98.90**	1.10	**98.67**	**100.00**	1.33	0.00
	S	85	27	4	**96.58**	3.42	**96.58**		3.42	
	L	53	17	2	**95.24**	4.76	**95.24**		4.76	
	Both	637	39	15	78.18	21.82	80.01	59.20	19.99	40.80
t=0.75	A	22	9	2	81.82	18.18	81.82		18.18	
	B	84	28	11	76.47	23.53	76.47		23.53	
	C	399	63	17	81.11	18.89	81.11		18.89	
	D	245	41	8	**94.66**	5.34	**94.66**		5.34	
	E	127	10	2	**98.56**	1.44	**98.56**		1.44	
	S	74	26	3	**97.20**	2.80	**97.20**		2.80	
	L	24	11	0	**100.00**	0.00	**100.00**		0.00	
	Both	375	64	17	84.48	15.52	84.48		15.52	
t=1.0	A	22	9	2	81.82	18.18	81.82		18.18	
	B	84	28	11	76.47	23.53	76.47		23.53	
	C	356	53	16	80.85	19.15	80.85		19.15	
	D	231	42	8	**94.22**	5.78	**94.22**		5.78	
	E	127	10	2	**98.56**	1.44	**98.56**		1.44	
	S	74	26	3	**97.20**	2.80	**97.20**		2.80	
	L	24	11	0	**100.00**	0.00	**100.00**		0.00	
	Both	336	58	16	84.07	15.93	84.07		15.93	

Figure A.7: *Z* Dataset Results - BRD of 91% Highlighted

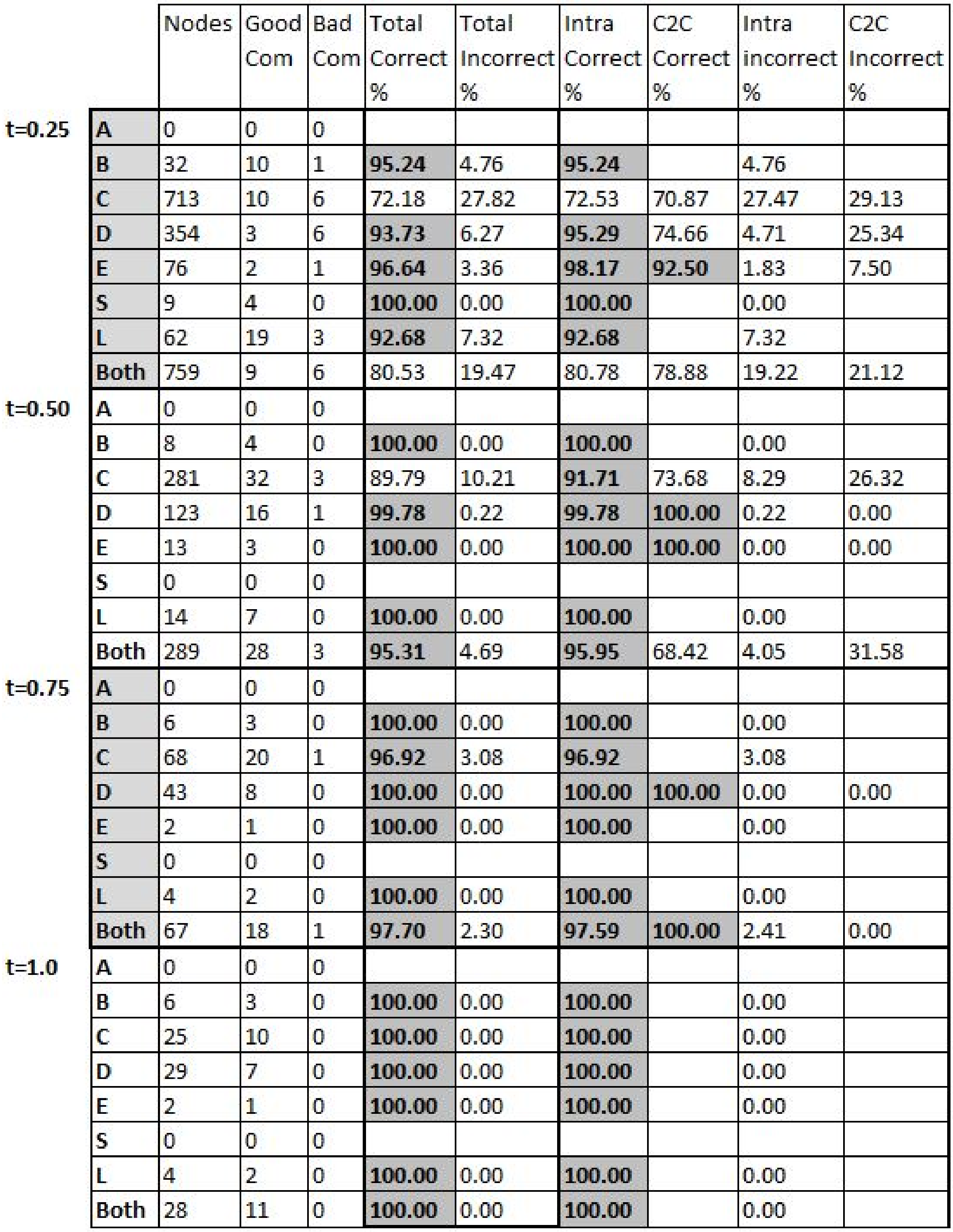
***S-val* S=2**



		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	2	0	1	0.00	**100.00**	0.00		**100.00**	
	B	57	12	5	87.80	12.20	87.80		12.20	
	C	767	13	9	68.04	31.96	66.22	75.17	33.78	24.83
	D	374	3	4	**93.71**	6.29	**95.05**	77.48	4.95	22.52
	E	98	2	1	**99.37**	0.63	**99.62**	**98.87**	0.38	1.13
	S	23	7	1	**94.12**	5.88	**94.12**		5.88	
	L	82	21	5	**92.19**	7.81	**92.19**		7.81	
	Both	804	6	12	73.17	26.83	73.76	70.36	26.24	29.64
t=0.50	A	2	0	1	0.00	**100.00**	0.00		**100.00**	
	B	30	10	3	83.33	16.67	83.33		16.67	
	C	547	35	9	78.69	21.31	78.71	78.60	21.29	21.40
	D	268	20	3	**97.30**	2.70	**97.41**	**95.83**	2.59	4.17
	E	90	4	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	15	5	1	88.89	11.11	88.89		11.11	
	L	30	11	2	88.89	11.11	88.89		11.11	
	Both	562	32	8	79.32	20.68	83.30	56.03	16.70	43.97
t=0.75	A	0	0	0						
	B	7	3	0	**100.00**	0.00	**100.00**		0.00	
	C	151	36	4	**93.64**	6.36	**93.64**		6.36	
	D	71	19	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	16	3	0	**100.00**	0.00	**100.00**		0.00	
	S	4	2	0	**100.00**	0.00	**100.00**		0.00	
	L	6	3	0	**100.00**	0.00	**100.00**		0.00	
	Both	132	33	3	**97.08**	2.92	**97.08**		2.92	
t=1.0	A	0	0	0						
	B	7	3	0	**100.00**	0.00	**100.00**		0.00	
	C	108	26	3	**93.36**	6.64	**93.36**		6.64	
	D	57	18	0	**100.00**	0.00	**100.00**		0.00	
	E	16	3	0	**100.00**	0.00	**100.00**		0.00	
	S	4	2	0	**100.00**	0.00	**100.00**		0.00	
	L	6	3	0	**100.00**	0.00	**100.00**		0.00	
	Both	93	27	2	**97.19**	2.81	**97.19**		2.81	

Figure A.8: *Z* Dataset Results - BRD of 91% Highlighted

***S-val* S=3**

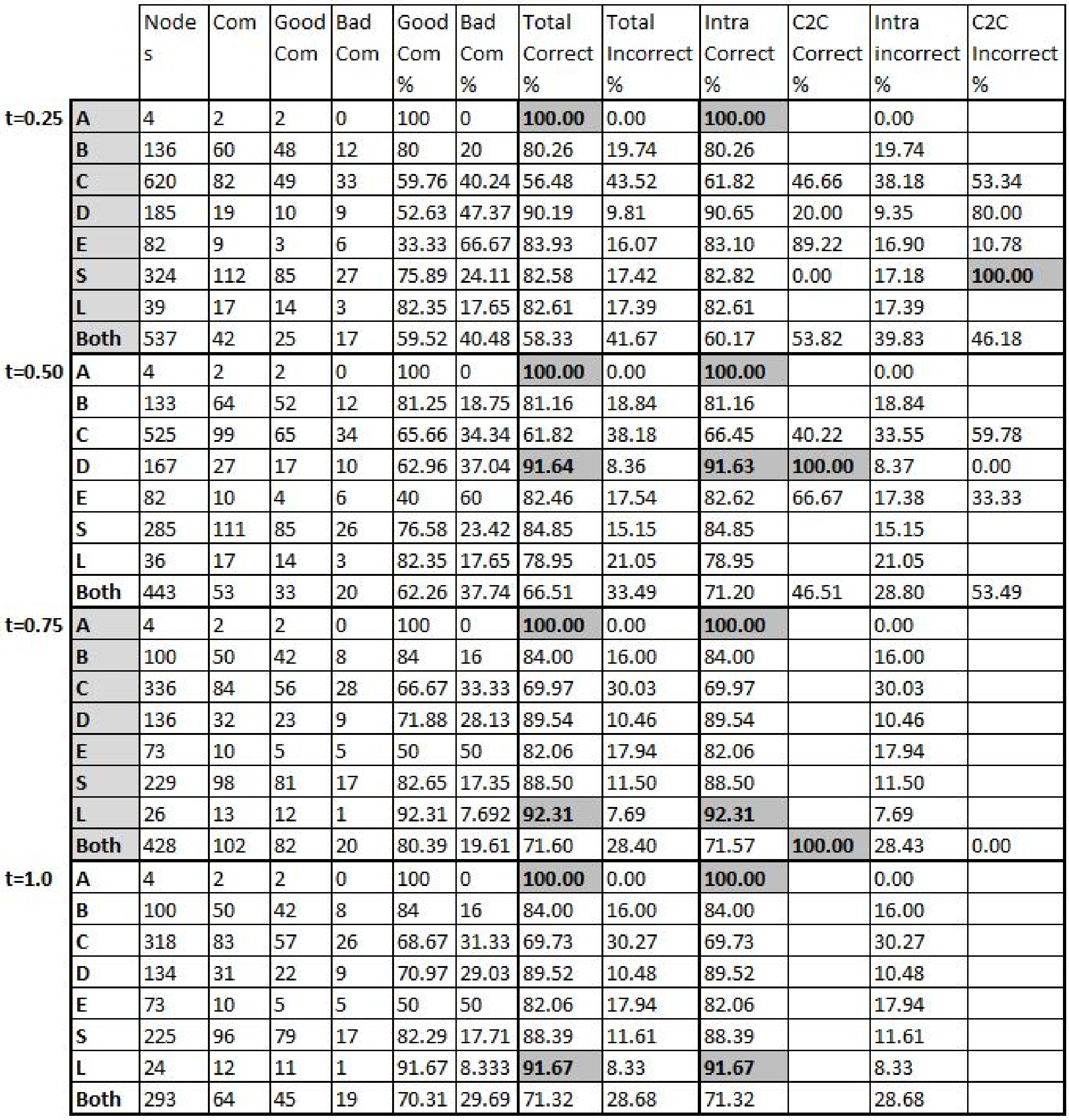


		Nodes	Good Com	Bad Com	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0						
	B	32	10	1	**95.24**	4.76	**95.24**		4.76	
	C	713	10	6	72.18	27.82	72.53	70.87	27.47	29.13
	D	354	3	6	**93.73**	6.27	**95.29**	74.66	4.71	25.34
	E	76	2	1	**96.64**	3.36	**98.17**	**92.50**	1.83	7.50
	S	9	4	0	**100.00**	0.00	**100.00**		0.00	
	L	62	19	3	**92.68**	7.32	**92.68**		7.32	
	Both	759	9	6	80.53	19.47	80.78	78.88	19.22	21.12
t=0.50	A	0	0	0						
	B	8	4	0	**100.00**	0.00	**100.00**		0.00	
	C	281	32	3	89.79	10.21	**91.71**	73.68	8.29	26.32
	D	123	16	1	**99.78**	0.22	**99.78**	**100.00**	0.22	0.00
	E	13	3	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	0	0	0						
	L	14	7	0	**100.00**	0.00	**100.00**		0.00	
	Both	289	28	3	**95.31**	4.69	**95.95**	68.42	4.05	31.58
t=0.75	A	0	0	0						
	B	6	3	0	**100.00**	0.00	**100.00**		0.00	
	C	68	20	1	**96.92**	3.08	**96.92**		3.08	
	D	43	8	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	2	1	0	**100.00**	0.00	**100.00**		0.00	
	S	0	0	0						
	L	4	2	0	**100.00**	0.00	**100.00**		0.00	
	Both	67	18	1	**97.70**	2.30	**97.59**	**100.00**	2.41	0.00
t=1.0	A	0	0	0						
	B	6	3	0	**100.00**	0.00	**100.00**		0.00	
	C	25	10	0	**100.00**	0.00	**100.00**		0.00	
	D	29	7	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	0	**100.00**	0.00	**100.00**		0.00	
	S	0	0	0						
	L	4	2	0	**100.00**	0.00	**100.00**		0.00	
	Both	28	11	0	**100.00**	0.00	**100.00**		0.00	

Figure A.9: *Z* Dataset Results - BRD of 91% Highlighted

A.0.4 *Y* DATASET UNWEIGHTED, RESOLUTION 1.0

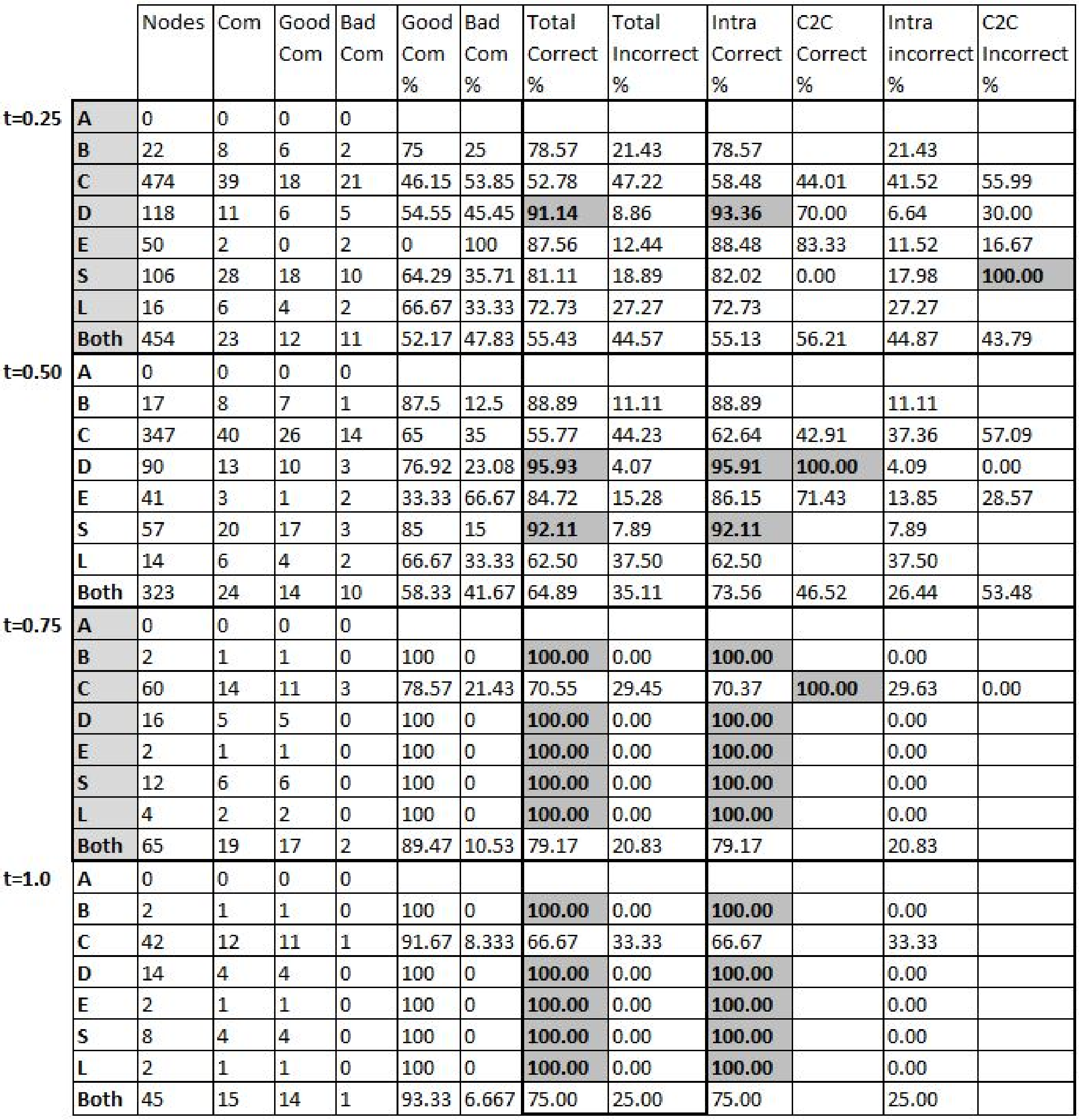
S-val **S=1**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	136	60	48	12	80	20	80.26	19.74	80.26		19.74	
	C	620	82	49	33	59.76	40.24	56.48	43.52	61.82	46.66	38.18	53.34
	D	185	19	10	9	52.63	47.37	90.19	9.81	90.65	20.00	9.35	80.00
	E	82	9	3	6	33.33	66.67	83.93	16.07	83.10	89.22	16.90	10.78
	S	324	112	85	27	75.89	24.11	82.58	17.42	82.82	0.00	17.18	**100.00**
	L	39	17	14	3	82.35	17.65	82.61	17.39	82.61		17.39	
	Both	537	42	25	17	59.52	40.48	58.33	41.67	60.17	53.82	39.83	46.18
t=0.50	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	133	64	52	12	81.25	18.75	81.16	18.84	81.16		18.84	
	C	525	99	65	34	65.66	34.34	61.82	38.18	66.45	40.22	33.55	59.78
	D	167	27	17	10	62.96	37.04	**91.64**	8.36	**91.63**	**100.00**	8.37	0.00
	E	82	10	4	6	40	60	82.46	17.54	82.62	66.67	17.38	33.33
	S	285	111	85	26	76.58	23.42	84.85	15.15	84.85		15.15	
	L	36	17	14	3	82.35	17.65	78.95	21.05	78.95		21.05	
	Both	443	53	33	20	62.26	37.74	66.51	33.49	71.20	46.51	28.80	53.49
t=0.75	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	336	84	56	28	66.67	33.33	69.97	30.03	69.97		30.03	
	D	136	32	23	9	71.88	28.13	89.54	10.46	89.54		10.46	
	E	73	10	5	5	50	50	82.06	17.94	82.06		17.94	
	S	229	98	81	17	82.65	17.35	88.50	11.50	88.50		11.50	
	L	26	13	12	1	92.31	7.692	**92.31**	7.69	**92.31**		7.69	
	Both	428	102	82	20	80.39	19.61	71.60	28.40	71.57	**100.00**	28.43	0.00
t=1.0	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	318	83	57	26	68.67	31.33	69.73	30.27	69.73		30.27	
	D	134	31	22	9	70.97	29.03	89.52	10.48	89.52		10.48	
	E	73	10	5	5	50	50	82.06	17.94	82.06		17.94	
	S	225	96	79	17	82.29	17.71	88.39	11.61	88.39		11.61	
	L	24	12	11	1	91.67	8.333	**91.67**	8.33	**91.67**		8.33	
	Both	293	64	45	19	70.31	29.69	71.32	28.68	71.32		28.68	

Figure A.10: *Y* Unweighted Dataset Resolution 1.0 - BRD of 91% Highlighted

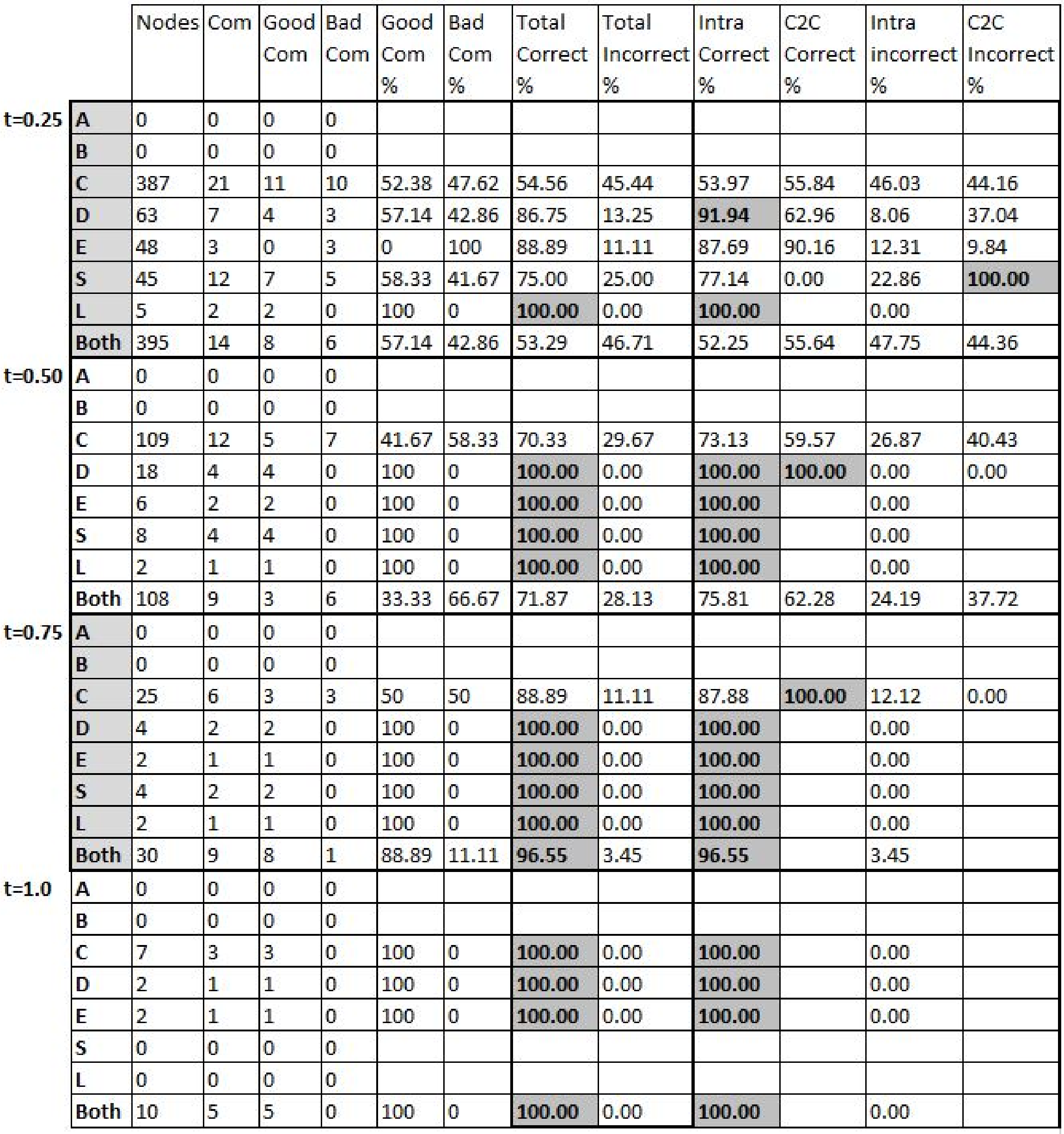
S-val S=2



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	22	8	6	2	75	25	78.57	21.43	78.57		21.43	
	C	474	39	18	21	46.15	53.85	52.78	47.22	58.48	44.01	41.52	55.99
	D	118	11	6	5	54.55	45.45	**91.14**	8.86	**93.36**	70.00	6.64	30.00
	E	50	2	0	2	0	100	87.56	12.44	88.48	83.33	11.52	16.67
	S	106	28	18	10	64.29	35.71	81.11	18.89	82.02	0.00	17.98	**100.00**
	L	16	6	4	2	66.67	33.33	72.73	27.27	72.73		27.27	
	Both	454	23	12	11	52.17	47.83	55.43	44.57	55.13	56.21	44.87	43.79
t=0.50	A	0	0	0	0								
	B	17	8	7	1	87.5	12.5	88.89	11.11	88.89		11.11	
	C	347	40	26	14	65	35	55.77	44.23	62.64	42.91	37.36	57.09
	D	90	13	10	3	76.92	23.08	**95.93**	4.07	**95.91**	**100.00**	4.09	0.00
	E	41	3	1	2	33.33	66.67	84.72	15.28	86.15	71.43	13.85	28.57
	S	57	20	17	3	85	15	**92.11**	7.89	**92.11**		7.89	
	L	14	6	4	2	66.67	33.33	62.50	37.50	62.50		37.50	
	Both	323	24	14	10	58.33	41.67	64.89	35.11	73.56	46.52	26.44	53.48
t=0.75	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	60	14	11	3	78.57	21.43	70.55	29.45	70.37	**100.00**	29.63	0.00
	D	16	5	5	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	12	6	6	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	65	19	17	2	89.47	10.53	79.17	20.83	79.17		20.83	
t=1.0	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	42	12	11	1	91.67	8.333	66.67	33.33	66.67		33.33	
	D	14	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	8	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	45	15	14	1	93.33	6.667	75.00	25.00	75.00		25.00	

Figure A.11: *Y* Unweighted Dataset Resolution 1.0 - BRD of 91% Highlighted

***S-val* S=3**

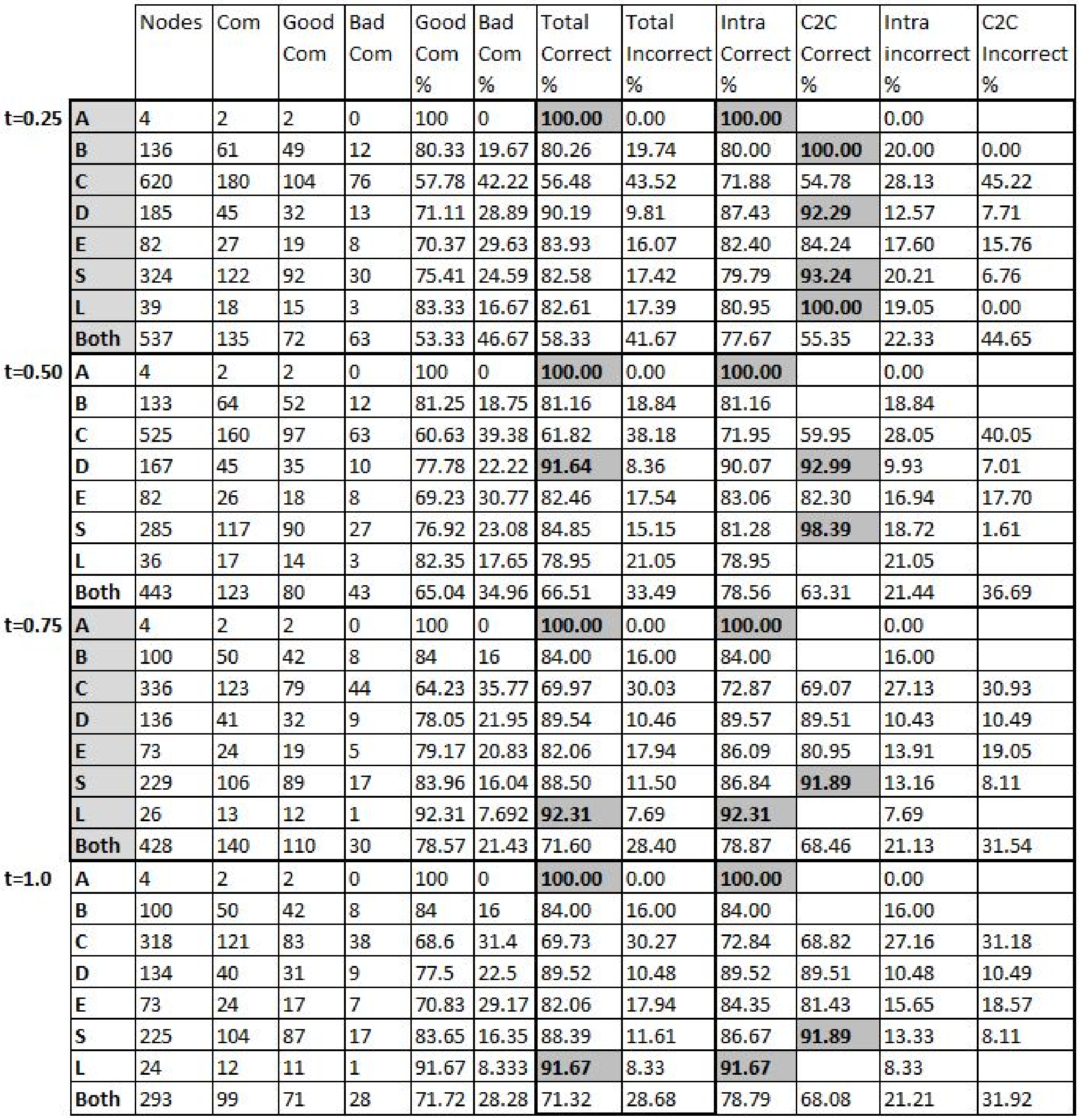


		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	0	0	0	0								
	C	387	21	11	10	52.38	47.62	54.56	45.44	53.97	55.84	46.03	44.16
	D	63	7	4	3	57.14	42.86	86.75	13.25	**91.94**	62.96	8.06	37.04
	E	48	3	0	3	0	100	88.89	11.11	87.69	90.16	12.31	9.84
	S	45	12	7	5	58.33	41.67	75.00	25.00	77.14	0.00	22.86	**100.00**
	L	5	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	395	14	8	6	57.14	42.86	53.29	46.71	52.25	55.64	47.75	44.36
t=0.50	A	0	0	0	0								
	B	0	0	0	0								
	C	109	12	5	7	41.67	58.33	70.33	29.67	73.13	59.57	26.87	40.43
	D	18	4	4	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	6	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	8	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	108	9	3	6	33.33	66.67	71.87	28.13	75.81	62.28	24.19	37.72
t=0.75	A	0	0	0	0								
	B	0	0	0	0								
	C	25	6	3	3	50	50	88.89	11.11	87.88	**100.00**	12.12	0.00
	D	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	30	9	8	1	88.89	11.11	**96.55**	3.45	**96.55**		3.45	
t=1.0	A	0	0	0	0								
	B	0	0	0	0								
	C	7	3	3	0	100	0	**100.00**	0.00	**100.00**		0.00	
	D	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	0	0	0	0								
	L	0	0	0	0								
	Both	10	5	5	0	100	0	**100.00**	0.00	**100.00**		0.00	

Figure A.12: *Y* Unweighted Dataset Resolution 1.0 - BRD of 91% Highlighted

A.0.5 *Y* DATASET UNWEIGHTED, RESOLUTION 0.1

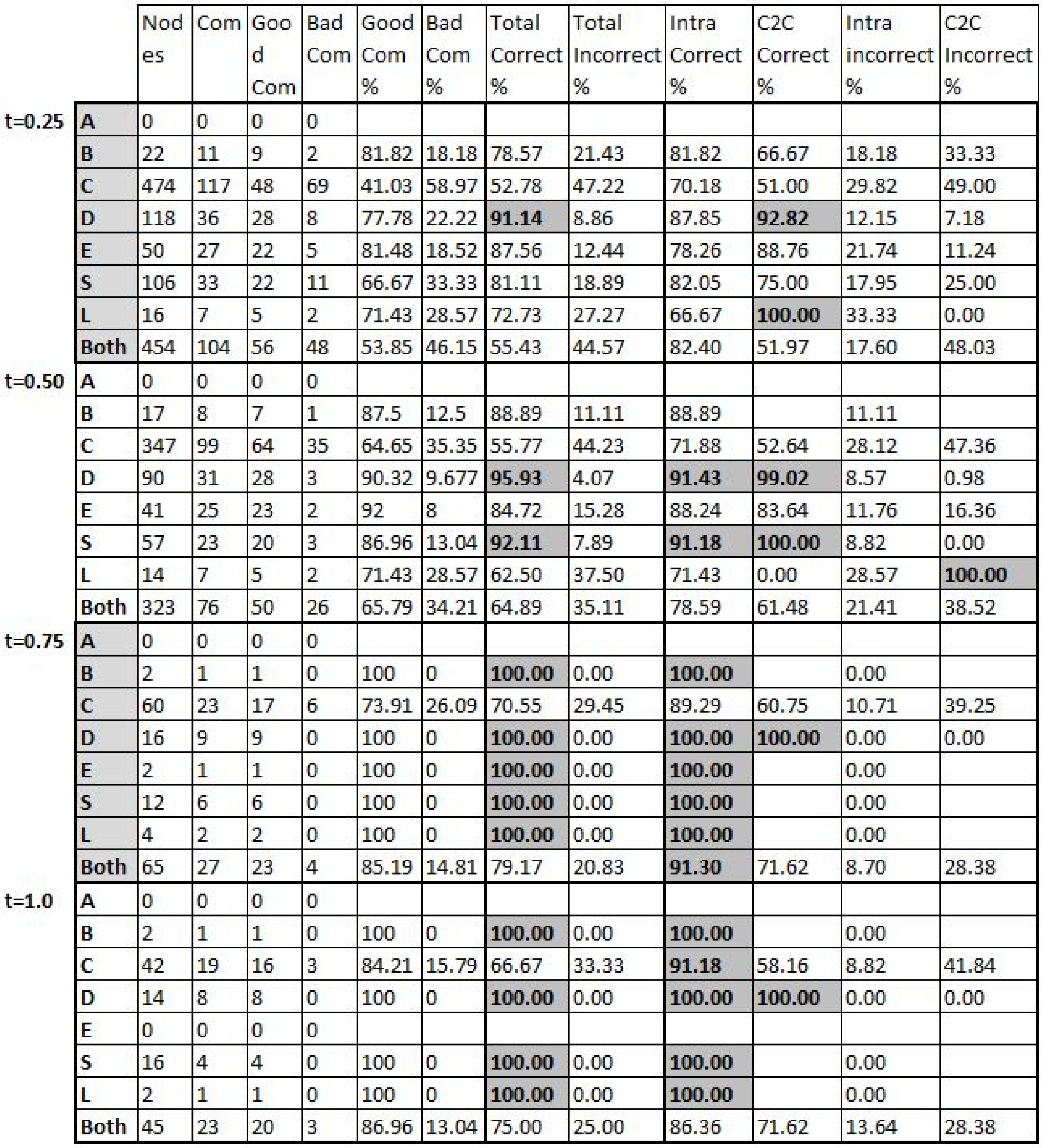
S-val **S=1**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	136	61	49	12	80.33	19.67	80.26	19.74	80.00	**100.00**	20.00	0.00
	C	620	180	104	76	57.78	42.22	56.48	43.52	71.88	54.78	28.13	45.22
	D	185	45	32	13	71.11	28.89	90.19	9.81	87.43	**92.29**	12.57	7.71
	E	82	27	19	8	70.37	29.63	83.93	16.07	82.40	84.24	17.60	15.76
	S	324	122	92	30	75.41	24.59	82.58	17.42	79.79	**93.24**	20.21	6.76
	L	39	18	15	3	83.33	16.67	82.61	17.39	80.95	**100.00**	19.05	0.00
	Both	537	135	72	63	53.33	46.67	58.33	41.67	77.67	55.35	22.33	44.65
t=0.50	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	133	64	52	12	81.25	18.75	81.16	18.84	81.16		18.84	
	C	525	160	97	63	60.63	39.38	61.82	38.18	71.95	59.95	28.05	40.05
	D	167	45	35	10	77.78	22.22	**91.64**	8.36	90.07	**92.99**	9.93	7.01
	E	82	26	18	8	69.23	30.77	82.46	17.54	83.06	82.30	16.94	17.70
	S	285	117	90	27	76.92	23.08	84.85	15.15	81.28	**98.39**	18.72	1.61
	L	36	17	14	3	82.35	17.65	78.95	21.05	78.95		21.05	
	Both	443	123	80	43	65.04	34.96	66.51	33.49	78.56	63.31	21.44	36.69
t=0.75	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	336	123	79	44	64.23	35.77	69.97	30.03	72.87	69.07	27.13	30.93
	D	136	41	32	9	78.05	21.95	89.54	10.46	89.57	89.51	10.43	10.49
	E	73	24	19	5	79.17	20.83	82.06	17.94	86.09	80.95	13.91	19.05
	S	229	106	89	17	83.96	16.04	88.50	11.50	86.84	**91.89**	13.16	8.11
	L	26	13	12	1	92.31	7.692	**92.31**	7.69	**92.31**		7.69	
	Both	428	140	110	30	78.57	21.43	71.60	28.40	78.87	68.46	21.13	31.54
t=1.0	**A**	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	318	121	83	38	68.6	31.4	69.73	30.27	72.84	68.82	27.16	31.18
	D	134	40	31	9	77.5	22.5	89.52	10.48	89.52	89.51	10.48	10.49
	E	73	24	17	7	70.83	29.17	82.06	17.94	84.35	81.43	15.65	18.57
	S	225	104	87	17	83.65	16.35	88.39	11.61	86.67	**91.89**	13.33	8.11
	L	24	12	11	1	91.67	8.333	**91.67**	8.33	**91.67**		8.33	
	Both	293	99	71	28	71.72	28.28	71.32	28.68	78.79	68.08	21.21	31.92

Figure A.13: *Y* Unweighted Dataset Resolution 0.1 - BRD of 91% Highlighted

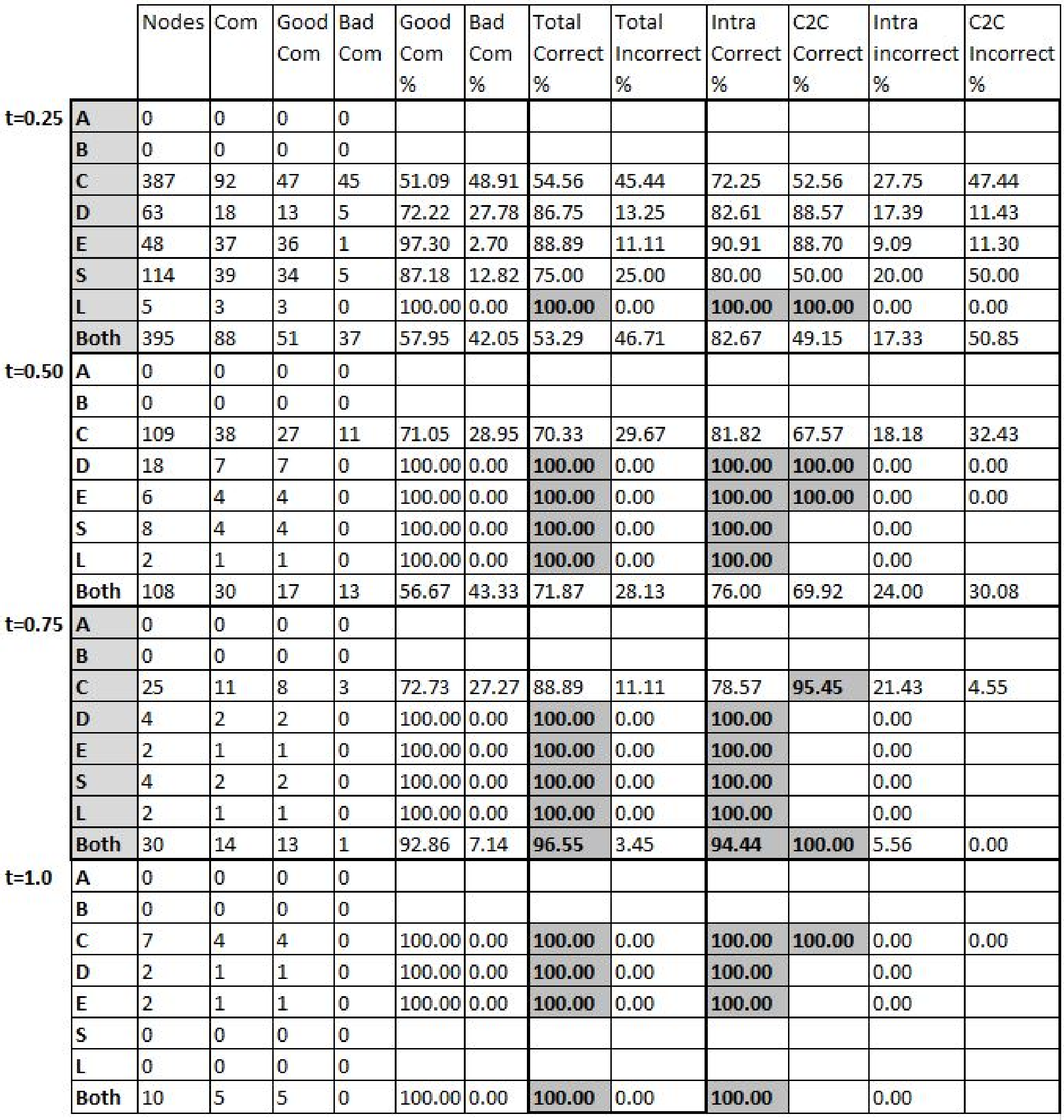
***S-val* S=2**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	22	11	9	2	81.82	18.18	78.57	21.43	81.82	66.67	18.18	33.33
	C	474	117	48	69	41.03	58.97	52.78	47.22	70.18	51.00	29.82	49.00
	D	118	36	28	8	77.78	22.22	**91.14**	8.86	87.85	**92.82**	12.15	7.18
	E	50	27	22	5	81.48	18.52	87.56	12.44	78.26	88.76	21.74	11.24
	S	106	33	22	11	66.67	33.33	81.11	18.89	82.05	75.00	17.95	25.00
	L	16	7	5	2	71.43	28.57	72.73	27.27	66.67	**100.00**	33.33	0.00
	Both	454	104	56	48	53.85	46.15	55.43	44.57	82.40	51.97	17.60	48.03
t=0.50	A	0	0	0	0								
	B	17	8	7	1	87.5	12.5	88.89	11.11	88.89		11.11	
	C	347	99	64	35	64.65	35.35	55.77	44.23	71.88	52.64	28.12	47.36
	D	90	31	28	3	90.32	9.677	**95.93**	4.07	**91.43**	**99.02**	8.57	0.98
	E	41	25	23	2	92	8	84.72	15.28	88.24	83.64	11.76	16.36
	S	57	23	20	3	86.96	13.04	**92.11**	7.89	**91.18**	**100.00**	8.82	0.00
	L	14	7	5	2	71.43	28.57	62.50	37.50	71.43	0.00	28.57	**100.00**
	Both	323	76	50	26	65.79	34.21	64.89	35.11	78.59	61.48	21.41	38.52
t=0.75	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	60	23	17	6	73.91	26.09	70.55	29.45	89.29	60.75	10.71	39.25
	D	16	9	9	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	12	6	6	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	65	27	23	4	85.19	14.81	79.17	20.83	**91.30**	71.62	8.70	28.38
t=1.0	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	42	19	16	3	84.21	15.79	66.67	33.33	**91.18**	58.16	8.82	41.84
	D	14	8	8	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	0	0	0	0								
	S	16	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	45	23	20	3	86.96	13.04	75.00	25.00	86.36	71.62	13.64	28.38

Figure A.14: *Y* Unweighted Dataset Resolution 0.1 - BRD of 91% Highlighted

S-val **S=3**

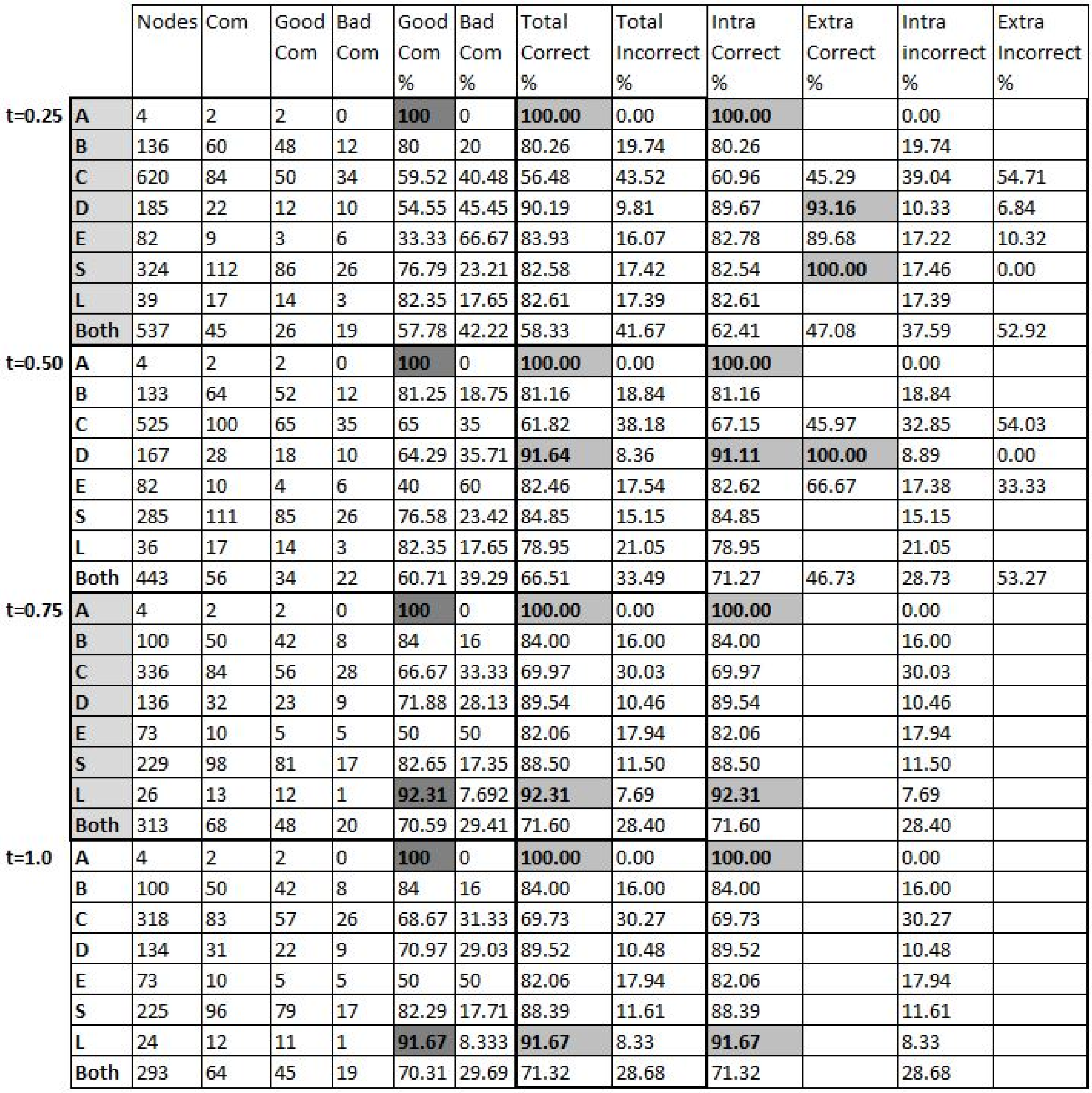


		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	0	0	0	0								
	C	387	92	47	45	51.09	48.91	54.56	45.44	72.25	52.56	27.75	47.44
	D	63	18	13	5	72.22	27.78	86.75	13.25	82.61	88.57	17.39	11.43
	E	48	37	36	1	97.30	2.70	88.89	11.11	90.91	88.70	9.09	11.30
	S	114	39	34	5	87.18	12.82	75.00	25.00	80.00	50.00	20.00	50.00
	L	5	3	3	0	100.00	0.00	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	Both	395	88	51	37	57.95	42.05	53.29	46.71	82.67	49.15	17.33	50.85
t=0.50	A	0	0	0	0								
	B	0	0	0	0								
	C	109	38	27	11	71.05	28.95	70.33	29.67	81.82	67.57	18.18	32.43
	D	18	7	7	0	100.00	0.00	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	6	4	4	0	100.00	0.00	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	8	4	4	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	Both	108	30	17	13	56.67	43.33	71.87	28.13	76.00	69.92	24.00	30.08
t=0.75	A	0	0	0	0								
	B	0	0	0	0								
	C	25	11	8	3	72.73	27.27	88.89	11.11	78.57	**95.45**	21.43	4.55
	D	4	2	2	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	S	4	2	2	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	Both	30	14	13	1	92.86	7.14	**96.55**	3.45	**94.44**	**100.00**	5.56	0.00
t=1.0	A	0	0	0	0								
	B	0	0	0	0								
	C	7	4	4	0	100.00	0.00	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	D	2	1	1	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	
	S	0	0	0	0								
	L	0	0	0	0								
	Both	10	5	5	0	100.00	0.00	**100.00**	0.00	**100.00**		0.00	

Figure A.15: *Y* Unweighted Dataset Resolution 0.1 - BRD of 91% Highlighted

A.0.6 *Y* DATASET WEIGHTED, RESOLUTION 1.0

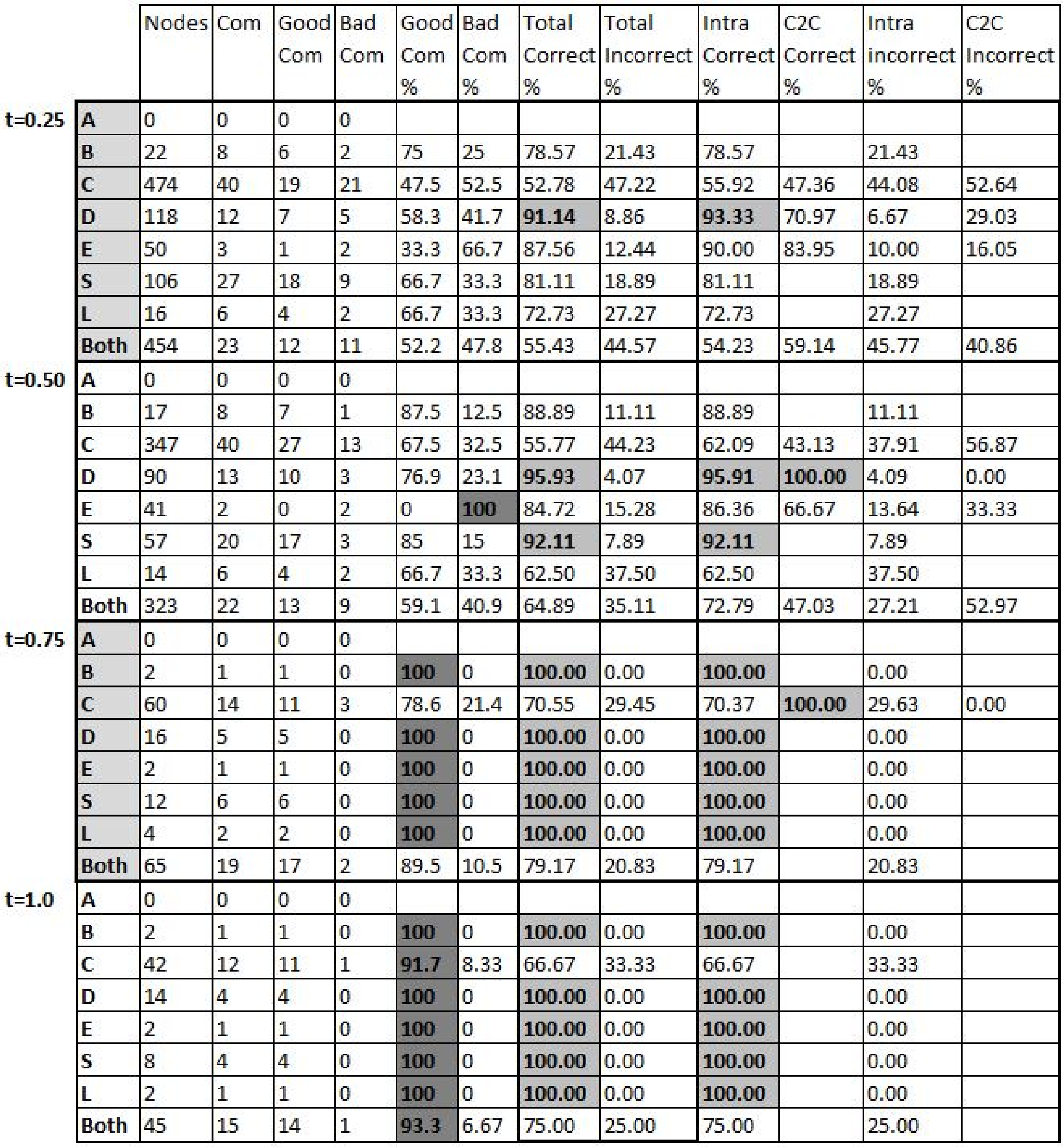
S-val **S=1**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	Extra Correct %	Intra incorrect %	Extra Incorrect %
t=0.25	A	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	B	136	60	48	12	80	20	80.26	19.74	80.26		19.74	
	C	620	84	50	34	59.52	40.48	56.48	43.52	60.96	45.29	39.04	54.71
	D	185	22	12	10	54.55	45.45	90.19	9.81	89.67	**93.16**	10.33	6.84
	E	82	9	3	6	33.33	66.67	83.93	16.07	82.78	89.68	17.22	10.32
	S	324	112	86	26	76.79	23.21	82.58	17.42	82.54	**100.00**	17.46	0.00
	L	39	17	14	3	82.35	17.65	82.61	17.39	82.61		17.39	
	Both	537	45	26	19	57.78	42.22	58.33	41.67	62.41	47.08	37.59	52.92
t=0.50	A	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	B	133	64	52	12	81.25	18.75	81.16	18.84	81.16		18.84	
	C	525	100	65	35	65	35	61.82	38.18	67.15	45.97	32.85	54.03
	D	167	28	18	10	64.29	35.71	**91.64**	8.36	**91.11**	**100.00**	8.89	0.00
	E	82	10	4	6	40	60	82.46	17.54	82.62	66.67	17.38	33.33
	S	285	111	85	26	76.58	23.42	84.85	15.15	84.85		15.15	
	L	36	17	14	3	82.35	17.65	78.95	21.05	78.95		21.05	
	Both	443	56	34	22	60.71	39.29	66.51	33.49	71.27	46.73	28.73	53.27
t=0.75	A	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	336	84	56	28	66.67	33.33	69.97	30.03	69.97		30.03	
	D	136	32	23	9	71.88	28.13	89.54	10.46	89.54		10.46	
	E	73	10	5	5	50	50	82.06	17.94	82.06		17.94	
	S	229	98	81	17	82.65	17.35	88.50	11.50	88.50		11.50	
	L	26	13	12	1	**92.31**	7.692	**92.31**	7.69	**92.31**		7.69	
	Both	313	68	48	20	70.59	29.41	71.60	28.40	71.60		28.40	
t=1.0	A	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	318	83	57	26	68.67	31.33	69.73	30.27	69.73		30.27	
	D	134	31	22	9	70.97	29.03	89.52	10.48	89.52		10.48	
	E	73	10	5	5	50	50	82.06	17.94	82.06		17.94	
	S	225	96	79	17	82.29	17.71	88.39	11.61	88.39		11.61	
	L	24	12	11	1	**91.67**	8.333	**91.67**	8.33	**91.67**		8.33	
	Both	293	64	45	19	70.31	29.69	71.32	28.68	71.32		28.68	

Figure A.16: *Y* Weighted Dataset Resolution 1.0 - BRD of 91% Highlighted

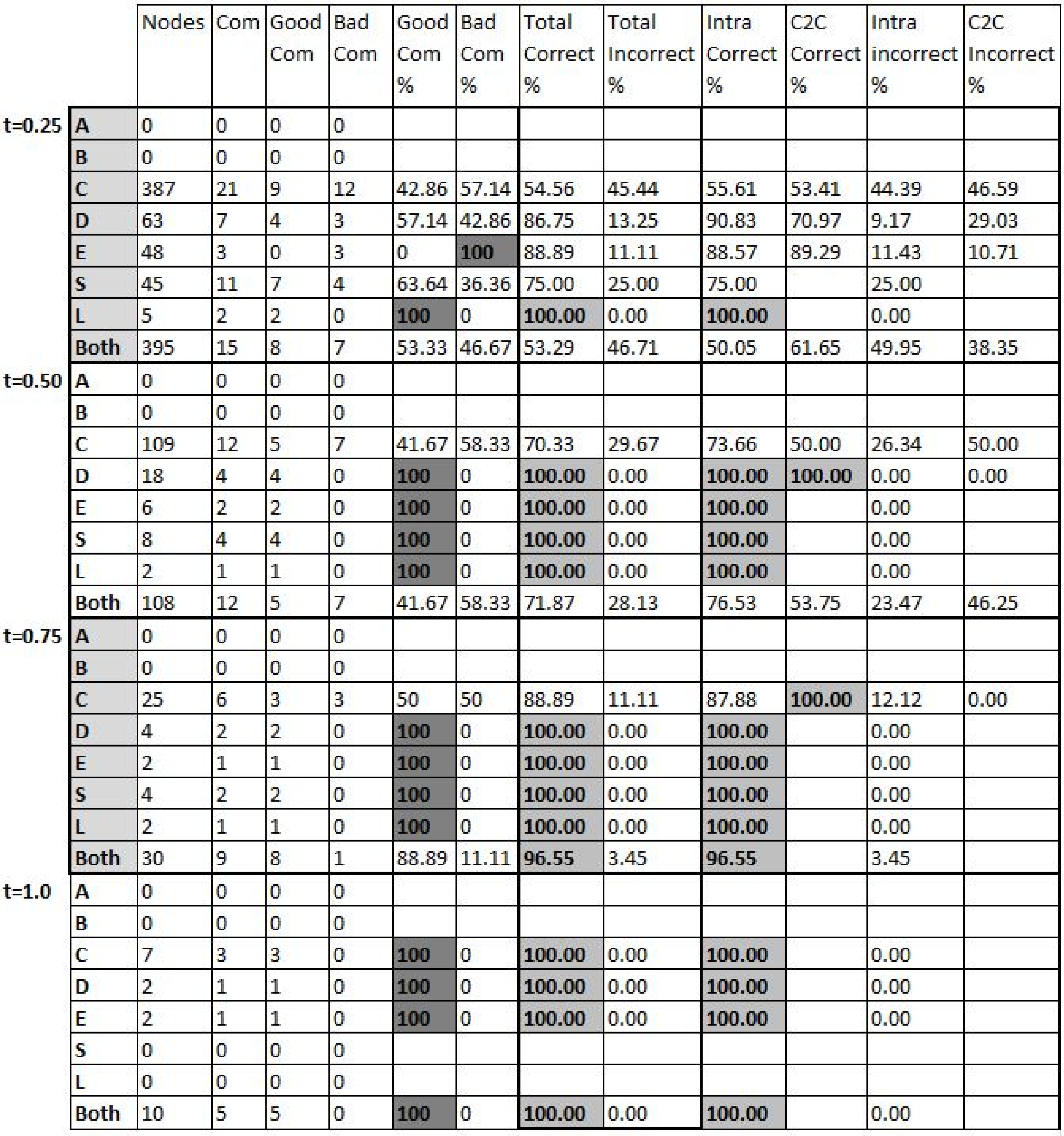
***S-val* S=2**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	22	8	6	2	75	25	78.57	21.43	78.57		21.43	
	C	474	40	19	21	47.5	52.5	52.78	47.22	55.92	47.36	44.08	52.64
	D	118	12	7	5	58.3	41.7	**91.14**	8.86	**93.33**	70.97	6.67	29.03
	E	50	3	1	2	33.3	66.7	87.56	12.44	90.00	83.95	10.00	16.05
	S	106	27	18	9	66.7	33.3	81.11	18.89	81.11		18.89	
	L	16	6	4	2	66.7	33.3	72.73	27.27	72.73		27.27	
	Both	454	23	12	11	52.2	47.8	55.43	44.57	54.23	59.14	45.77	40.86
t=0.50	A	0	0	0	0								
	B	17	8	7	1	87.5	12.5	88.89	11.11	88.89		11.11	
	C	347	40	27	13	67.5	32.5	55.77	44.23	62.09	43.13	37.91	56.87
	D	90	13	10	3	76.9	23.1	**95.93**	4.07	**95.91**	**100.00**	4.09	0.00
	E	41	2	0	2	0	**100**	84.72	15.28	86.36	66.67	13.64	33.33
	S	57	20	17	3	85	15	**92.11**	7.89	**92.11**		7.89	
	L	14	6	4	2	66.7	33.3	62.50	37.50	62.50		37.50	
	Both	323	22	13	9	59.1	40.9	64.89	35.11	72.79	47.03	27.21	52.97
t=0.75	A	0	0	0	0								
	B	2	1	1	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	C	60	14	11	3	78.6	21.4	70.55	29.45	70.37	**100.00**	29.63	0.00
	D	16	5	5	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	S	12	6	6	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	L	4	2	2	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	Both	65	19	17	2	89.5	10.5	79.17	20.83	79.17		20.83	
t=1.0	A	0	0	0	0								
	B	2	1	1	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	C	42	12	11	1	**91.7**	8.33	66.67	33.33	66.67		33.33	
	D	14	4	4	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	S	8	4	4	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	**100**	0	**100.00**	0.00	**100.00**		0.00	
	Both	45	15	14	1	**93.3**	6.67	75.00	25.00	75.00		25.00	

Figure A.17: Y Weighted Dataset Resolution 1.0 - BRD of 91% Highlighted

S-val **S=3**

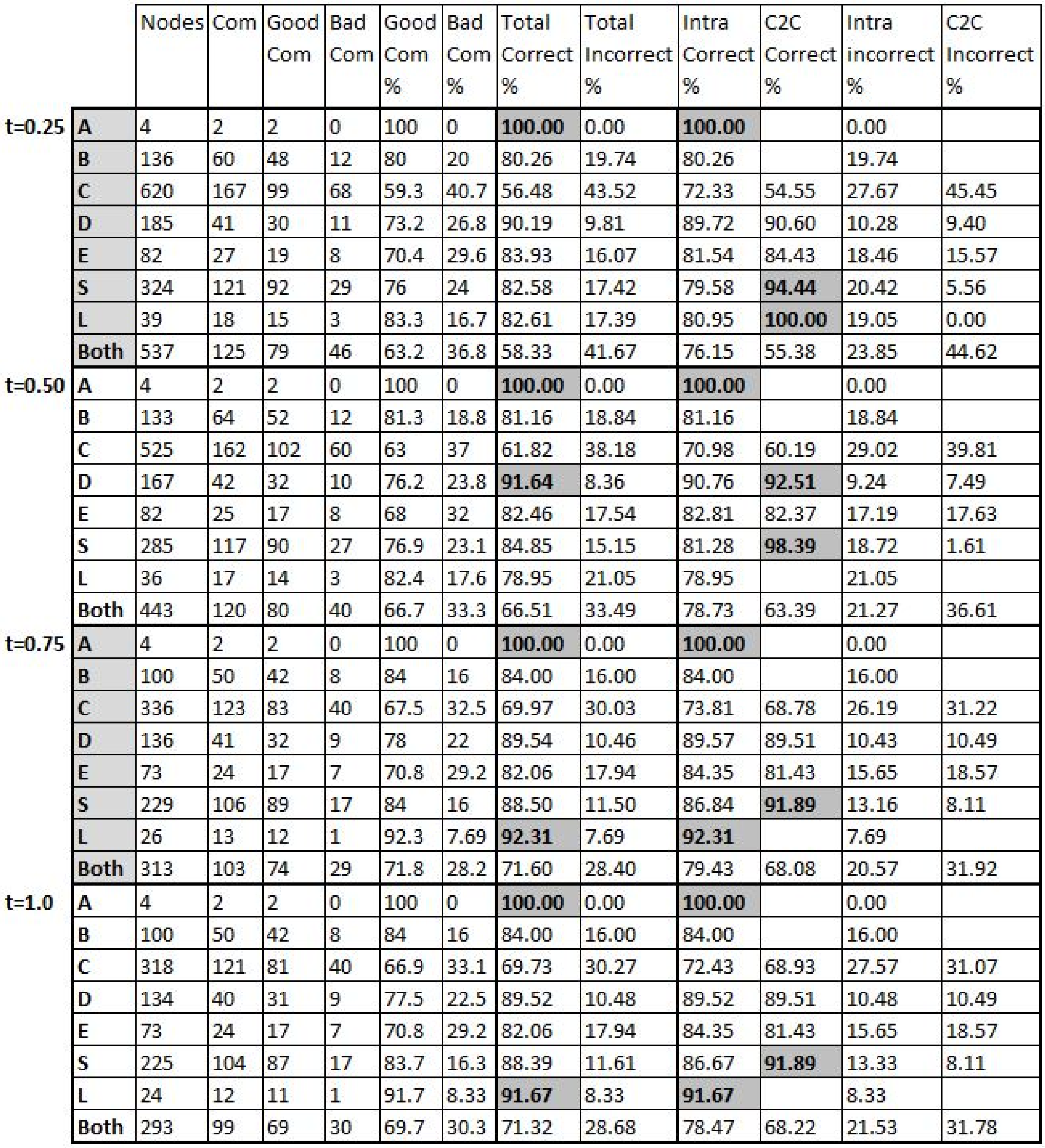


		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	0	0	0	0								
	C	387	21	9	12	42.86	57.14	54.56	45.44	55.61	53.41	44.39	46.59
	D	63	7	4	3	57.14	42.86	86.75	13.25	90.83	70.97	9.17	29.03
	E	48	3	0	3	0	100	88.89	11.11	88.57	89.29	11.43	10.71
	S	45	11	7	4	63.64	36.36	75.00	25.00	75.00		25.00	
	L	5	2	2	0	100	0	100.00	0.00	100.00		0.00	
	Both	395	15	8	7	53.33	46.67	53.29	46.71	50.05	61.65	49.95	38.35
t=0.50	A	0	0	0	0								
	B	0	0	0	0								
	C	109	12	5	7	41.67	58.33	70.33	29.67	73.66	50.00	26.34	50.00
	D	18	4	4	0	100	0	100.00	0.00	100.00	100.00	0.00	0.00
	E	6	2	2	0	100	0	100.00	0.00	100.00		0.00	
	S	8	4	4	0	100	0	100.00	0.00	100.00		0.00	
	L	2	1	1	0	100	0	100.00	0.00	100.00		0.00	
	Both	108	12	5	7	41.67	58.33	71.87	28.13	76.53	53.75	23.47	46.25
t=0.75	A	0	0	0	0								
	B	0	0	0	0								
	C	25	6	3	3	50	50	88.89	11.11	87.88	100.00	12.12	0.00
	D	4	2	2	0	100	0	100.00	0.00	100.00		0.00	
	E	2	1	1	0	100	0	100.00	0.00	100.00		0.00	
	S	4	2	2	0	100	0	100.00	0.00	100.00		0.00	
	L	2	1	1	0	100	0	100.00	0.00	100.00		0.00	
	Both	30	9	8	1	88.89	11.11	96.55	3.45	96.55		3.45	
t=1.0	A	0	0	0	0								
	B	0	0	0	0								
	C	7	3	3	0	100	0	100.00	0.00	100.00		0.00	
	D	2	1	1	0	100	0	100.00	0.00	100.00		0.00	
	E	2	1	1	0	100	0	100.00	0.00	100.00		0.00	
	S	0	0	0	0								
	L	0	0	0	0								
	Both	10	5	5	0	100	0	100.00	0.00	100.00		0.00	

Figure A.18: *Y* Weighted Dataset Resolution 1.0 - BRD of 91% Highlighted

A.0.7 *Y* DATASET WEIGHTED, RESOLUTION 0.1

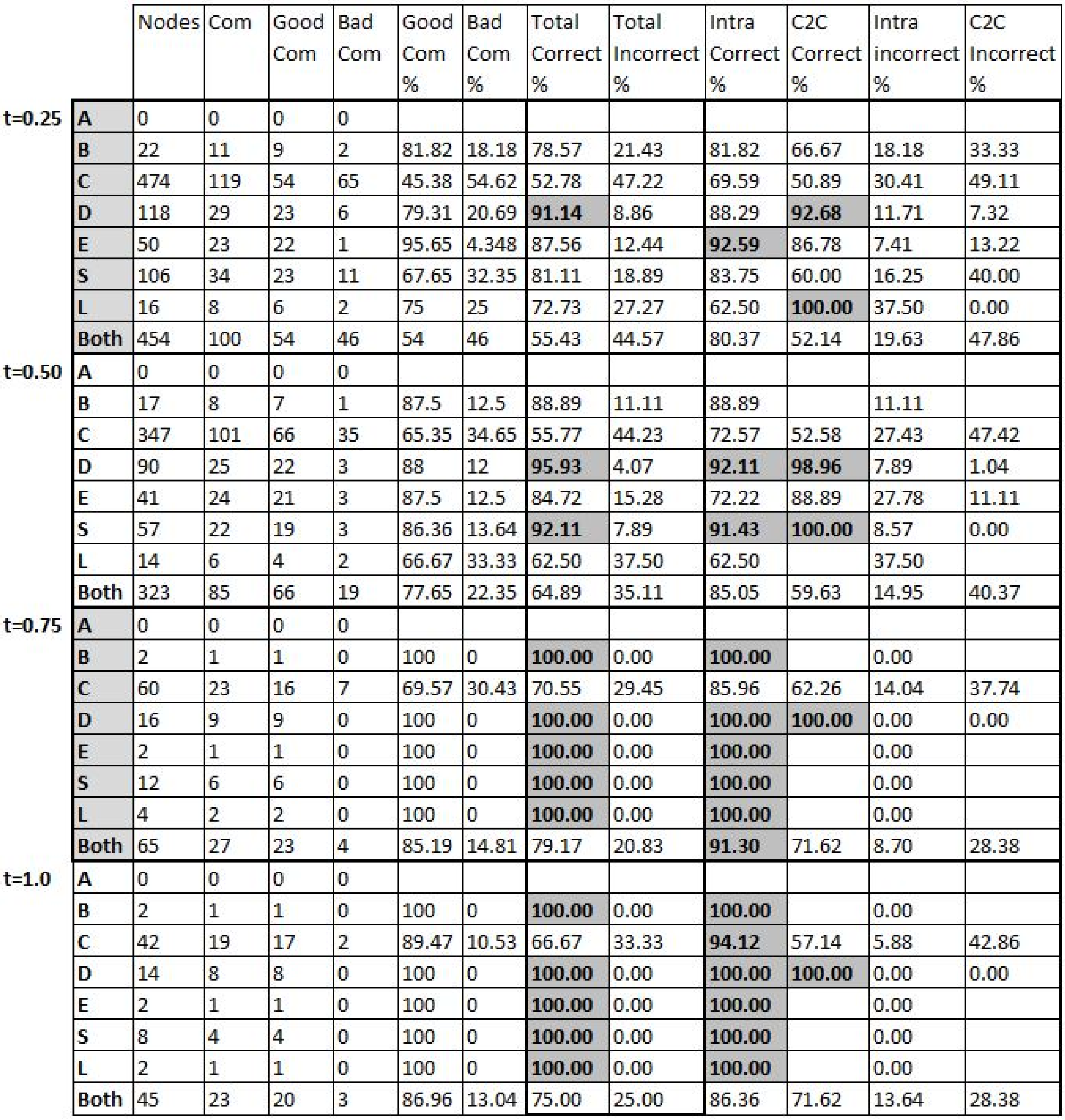
S-val S=1



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	136	60	48	12	80	20	80.26	19.74	80.26		19.74	
	C	620	167	99	68	59.3	40.7	56.48	43.52	72.33	54.55	27.67	45.45
	D	185	41	30	11	73.2	26.8	90.19	9.81	89.72	90.60	10.28	9.40
	E	82	27	19	8	70.4	29.6	83.93	16.07	81.54	84.43	18.46	15.57
	S	324	121	92	29	76	24	82.58	17.42	79.58	**94.44**	20.42	5.56
	L	39	18	15	3	83.3	16.7	82.61	17.39	80.95	**100.00**	19.05	0.00
	Both	537	125	79	46	63.2	36.8	58.33	41.67	76.15	55.38	23.85	44.62
t=0.50	A	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	133	64	52	12	81.3	18.8	81.16	18.84	81.16		18.84	
	C	525	162	102	60	63	37	61.82	38.18	70.98	60.19	29.02	39.81
	D	167	42	32	10	76.2	23.8	**91.64**	8.36	90.76	**92.51**	9.24	7.49
	E	82	25	17	8	68	32	82.46	17.54	82.81	82.37	17.19	17.63
	S	285	117	90	27	76.9	23.1	84.85	15.15	81.28	**98.39**	18.72	1.61
	L	36	17	14	3	82.4	17.6	78.95	21.05	78.95		21.05	
	Both	443	120	80	40	66.7	33.3	66.51	33.49	78.73	63.39	21.27	36.61
t=0.75	A	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	336	123	83	40	67.5	32.5	69.97	30.03	73.81	68.78	26.19	31.22
	D	136	41	32	9	78	22	89.54	10.46	89.57	89.51	10.43	10.49
	E	73	24	17	7	70.8	29.2	82.06	17.94	84.35	81.43	15.65	18.57
	S	229	106	89	17	84	16	88.50	11.50	86.84	**91.89**	13.16	8.11
	L	26	13	12	1	92.3	7.69	**92.31**	7.69	**92.31**		7.69	
	Both	313	103	74	29	71.8	28.2	71.60	28.40	79.43	68.08	20.57	31.92
t=1.0	A	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	B	100	50	42	8	84	16	84.00	16.00	84.00		16.00	
	C	318	121	81	40	66.9	33.1	69.73	30.27	72.43	68.93	27.57	31.07
	D	134	40	31	9	77.5	22.5	89.52	10.48	89.52	89.51	10.48	10.49
	E	73	24	17	7	70.8	29.2	82.06	17.94	84.35	81.43	15.65	18.57
	S	225	104	87	17	83.7	16.3	88.39	11.61	86.67	**91.89**	13.33	8.11
	L	24	12	11	1	91.7	8.33	**91.67**	8.33	**91.67**		8.33	
	Both	293	99	69	30	69.7	30.3	71.32	28.68	78.47	68.22	21.53	31.78

Figure A.19: *Y* Weighted Dataset Resolution 0.1 - BRD of 91% Highlighted

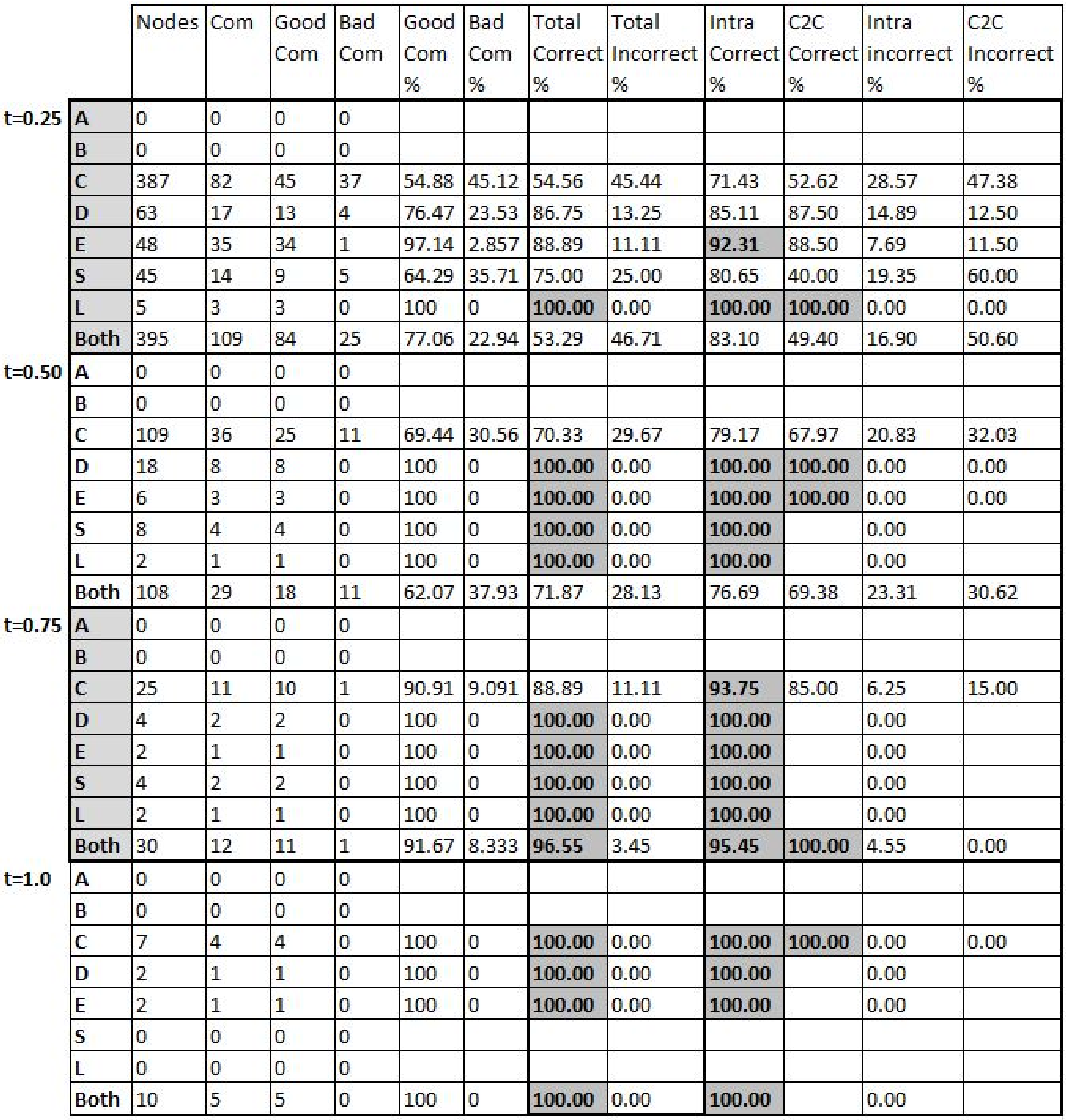
S-val **S=2**



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	A	0	0	0	0								
	B	22	11	9	2	81.82	18.18	78.57	21.43	81.82	66.67	18.18	33.33
	C	474	119	54	65	45.38	54.62	52.78	47.22	69.59	50.89	30.41	49.11
	D	118	29	23	6	79.31	20.69	**91.14**	8.86	88.29	**92.68**	11.71	7.32
	E	50	23	22	1	95.65	4.348	87.56	12.44	**92.59**	86.78	7.41	13.22
	S	106	34	23	11	67.65	32.35	81.11	18.89	83.75	60.00	16.25	40.00
	L	16	8	6	2	75	25	72.73	27.27	62.50	**100.00**	37.50	0.00
	Both	454	100	54	46	54	46	55.43	44.57	80.37	52.14	19.63	47.86
t=0.50	A	0	0	0	0								
	B	17	8	7	1	87.5	12.5	88.89	11.11	88.89		11.11	
	C	347	101	66	35	65.35	34.65	55.77	44.23	72.57	52.58	27.43	47.42
	D	90	25	22	3	88	12	**95.93**	4.07	**92.11**	**98.96**	7.89	1.04
	E	41	24	21	3	87.5	12.5	84.72	15.28	72.22	88.89	27.78	11.11
	S	57	22	19	3	86.36	13.64	**92.11**	7.89	**91.43**	**100.00**	8.57	0.00
	L	14	6	4	2	66.67	33.33	62.50	37.50	62.50		37.50	
	Both	323	85	66	19	77.65	22.35	64.89	35.11	85.05	59.63	14.95	40.37
t=0.75	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	60	23	16	7	69.57	30.43	70.55	29.45	85.96	62.26	14.04	37.74
	D	16	9	9	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	12	6	6	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	65	27	23	4	85.19	14.81	79.17	20.83	**91.30**	71.62	8.70	28.38
t=1.0	A	0	0	0	0								
	B	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	C	42	19	17	2	89.47	10.53	66.67	33.33	**94.12**	57.14	5.88	42.86
	D	14	8	8	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	8	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	45	23	20	3	86.96	13.04	75.00	25.00	86.36	71.62	13.64	28.38

Figure A.20: *Y* Weighted Dataset Resolution 0.1 - BRD of 91% Highlighted

S-val S=3



		Nodes	Com	Good Com	Bad Com	Good Com %	Bad Com %	Total Correct %	Total Incorrect %	Intra Correct %	C2C Correct %	Intra incorrect %	C2C Incorrect %
t=0.25	**A**	0	0	0	0								
	B	0	0	0	0								
	C	387	82	45	37	54.88	45.12	54.56	45.44	71.43	52.62	28.57	47.38
	D	63	17	13	4	76.47	23.53	86.75	13.25	85.11	87.50	14.89	12.50
	E	48	35	34	1	97.14	2.857	88.89	11.11	**92.31**	88.50	7.69	11.50
	S	45	14	9	5	64.29	35.71	75.00	25.00	80.65	40.00	19.35	60.00
	L	5	3	3	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	Both	395	109	84	25	77.06	22.94	53.29	46.71	83.10	49.40	16.90	50.60
t=0.50	**A**	0	0	0	0								
	B	0	0	0	0								
	C	109	36	25	11	69.44	30.56	70.33	29.67	79.17	67.97	20.83	32.03
	D	18	8	8	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	E	6	3	3	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	S	8	4	4	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	108	29	18	11	62.07	37.93	71.87	28.13	76.69	69.38	23.31	30.62
t=0.75	**A**	0	0	0	0								
	B	0	0	0	0								
	C	25	11	10	1	90.91	9.091	88.89	11.11	**93.75**	85.00	6.25	15.00
	D	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	4	2	2	0	100	0	**100.00**	0.00	**100.00**		0.00	
	L	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	Both	30	12	11	1	91.67	8.333	**96.55**	3.45	**95.45**	**100.00**	4.55	0.00
t=1.0	**A**	0	0	0	0								
	B	0	0	0	0								
	C	7	4	4	0	100	0	**100.00**	0.00	**100.00**	**100.00**	0.00	0.00
	D	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	E	2	1	1	0	100	0	**100.00**	0.00	**100.00**		0.00	
	S	0	0	0	0								
	L	0	0	0	0								
	Both	10	5	5	0	100	0	**100.00**	0.00	**100.00**		0.00	

Figure A.21: *Y* Weighted Dataset Resolution 0.1 - BRD of 91% Highlighted