Digital Investigation 9 (2012) S99–S107

Contents lists available at SciVerse ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

The use of random sampling in investigations involving child abuse material

Brian Jones [a], Syd Pleno [b], Michael Wilkinson [c,*]

[a] State Electronic Evidence Branch, NSW Police Force, Sydney, NSW, Australia
[b] Australian Federal Police, Sydney, NSW, Australia
[c] Champlain College, Burlington, VT, USA

Keywords:
Child abuse material
Sampling
Random sampling
Efficiency
Encase
Scripting
Automated digital forensics

A B S T R A C T

This paper presents a methodology which has been used to address two ubiquitous problems of practising digital forensics in law enforcement, the ever increasing data volume and staff exposure to disturbing material. It discusses how the New South Wales Police Force, State Electronic Evidence Branch (SEEB) has implemented a "Discovery Process". Using random sampling of files and applying statistical estimation to the results, the branch has been able to reduce backlogs from three months to 24 h. The process has the added advantage of reducing staff exposure to child abuse material and providing the courts with an easily interpreted report. The software portion of the Discovery process is contained within the framework of Guidance software's forensic tool, EnCase©. This is then further customised for the Discovery process by using the EnCase EnScript© language.

1. Introduction

Since 2003 the State Electronic Evidence Branch (SEEB) of the Australian New South Wales (NSW) Police Force has been the sole provider of digital forensic analysis for the State's 16,000 police officers. Due to ever increasing demand for digital forensic support SEEB has been required to develop and implement a range of processes to maximise its ability to provide timely support to serious major crime investigations. One area of significant demand is investigations relating to the possession of child abuse material[1] (CAM). Historically it was the responsibility of the SEEB analyst to identify all pictures, documents and videos depicting CAM. This approach had numerous failings; the most significant of these were the burnout of staff and long delays due to the small number of analysts available. This has been further exacerbated by the increasing volume of CAM encountered as Internet speeds and hard drive sizes (Stewart and Black, 2012) have made the sharing and storage of material faster and easier. In order to address this challenge SEEB has implemented a method of random sampling of potential CAM that has resulted in a reduction in response time from 3 months to 24 h and reduced the exposure of SEEB analysts and investigating officers to CAM. In addition to reducing backlog the reports generated by this process have been received extremely favourably by the NSW courts.

1.1. The role of SEEB in CAM investigations

As the provider of digital forensic analysis for the NSW Police Force, SEEB receives approximately 1200 requests for assistance a year. Since 2005 between 17% and 23% have related to the charge "possession of child abuse material"(CRIMES ACT, 1900). Other types of investigations the branch assists with are shown in Fig. 1.

From 2005 to 2012 less than 24% of the investigations involving SEEB related to the possession of CAM. However

* Corresponding author. Tel.: +1 802 865 6460.
E-mail address: wilkinson@champlain.edu (M. Wilkinson).

[1] The term Child Abuse Material is used throughout this document as it is the term used in NSW legislation. At a national level the term Child Exploitation Material (CEM) is interchangeable with Child Abuse Material.

1742-2876/$ – see front matter
doi:10.1016/j.diin.2012.05.011

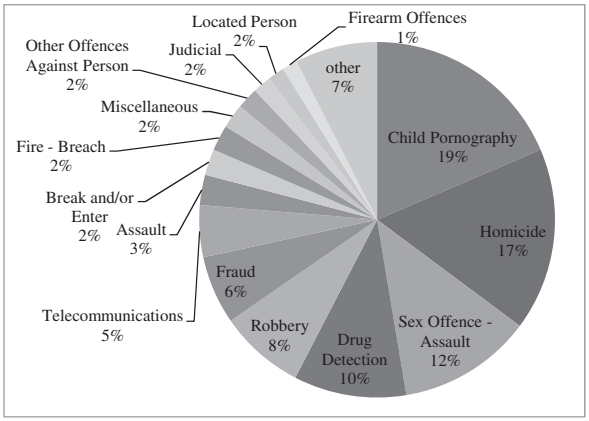

Fig. 1. SEEB investigations July 2010 to June 2011.

these investigations pose a disproportionate health risk to the analysts and investigating officers involved with them. CAM is disturbing by nature; it involves graphic pictures, videos and descriptions of sexual assault, torture and mutilation of children. It is widely recognized that reducing exposure to this type of material is an important step in improving workplace safety (Perez et al., 2010; Holt and Blevins, 2011; Wolak and Mitchell, 2009; Jewkes and Andrews, 2005) and avoiding mental health problems for those exposed to it.

1.2. Legal framework

This process has been developed specifically to meet the needs of the NSW Police Force, prosecuting crimes committed in New South Wales, Australia. It is designed to meet the requirements of New South Wales and Australian law, as such some of the procedures described here may not meet, or may exceed legal requirements in other jurisdictions. The primary focus of this paper is the scientific robustness of the process, within the confines of the NSW legal system. The authors are not aware of any similar implementations of sampling of digital evidence in criminal investigations.

In NSW offences relating to CAM may fall under both State and Federal Legalisation. The NSW Crimes Act (CRIMES ACT, 1900) specifies crimes relating to the possession and distribution of CAM, Federal law addresses offences relating to the access and distribution of CAM via a telecommunications device. This process has been developed to address offences of possession under the NSW legalisation. In particular determining the absolute and relative volumes of CAM on a given storage device. It should be noted that this information is used to assist the court in determining appropriate sentences for convicted offenders, it is not used for the purposes of determining guilt.

In NSW CAM is defined in Sect 91FB of the CRIMES ACT (1900), as "material that depicts or describes, in a way that reasonable persons would regard as being, in all the circumstances, offensive"... "a person who is, appears to be or is implied to be, a child". Thus there is no requirement to identify the individual portrayed in a picture, video or document. In fact it is possible for created pictures such as cartoons to be considered child abuse material (Whiley v R, 2010).

In DPP v Kear (Director of Public Prosecutions v Kear, 2006) it was found that in order to prove an offence of possession the prosecution must demonstrate that the accused has knowledge of the CAM on the computer. In this case it was found that a computer user cannot be reasonably expected to know that a web browser caches files on the local hard drive. Therefore being in possession of a computer containing CAM located in a web browser cache directory is not an offence. The same reasoning has been applied to CAM in deleted or unallocated space.

When CAM cases proceed to court an important factor in determination of guilt and sentencing is the quantity of CAM material found in possession of the defendant (Puhakka v R, 2009; R v Elliott, 2008; R v Gent, 2005; R v Leonard, 2008; Sivell v R, 2009). Furthermore a potential defence is that the CAM came into the possession of the accused inadvertently as a result of collecting adult pornography, as seen in Gosling v D.P.P. (Gosling v DPP, 2009). In either case in order to meet the expectations of the court it is necessary to identify the volume of CAM on any storage device and the ratio of CAM to other (legal) pornographic material. In order to address these requirements it has been necessary for investigators or SEEB analysts to view all picture, video and document files on all seized digital storage devices. As Internet speeds and usage have increased so too have the size of pornography and CAM collections.

1.3. Effects of exposure to child abuse material

During the past 8 years, SEEB staff has observed varying degrees of stress that investigators and SEEB staff experience when viewing disturbing pictures, videos and documents. The immediate reactions to this material is consistently negative and range from slightly perturbed to the inability to continue the process – with the majority in between these two extremes. In one case two investigating officers were on sick leave for two months following a day at SEEB identifying and classifying CAM. On other occasions investigators have demonstrated physical reactions such as crying and vomiting. SEEB has also had a number of trained digital forensic analysts resign as a result of exposure to this distressing material.

It has been shown in several studies that constant and prolonged exposure to disturbing material has a negative psychological effect on examiners and case officers alike (Perez et al., 2010; Holt and Blevins, 2011; Wolak and Mitchell, 2009; Krause, 2009). The SEEB experience is consistant with that of Able et al. in that:

It is to be noted that although a large percentage of the files found involve images of the exploitation of children, images were not restricted to child abuse. The tendency for suspects who collect child pornographic images to also collect images of extreme violence, lethal violence, and unusual sexual behaviour... (Abel et al., 1988).

The only way to protect investigators and analysts alike is to reduce exposure to this material. However achieving this when the courts are requesting totals of material found and ratios of illegal to legal pornography is difficult. One approach is to use random sampling of a small subset to estimate quantities within the whole set.

2. Existing uses of sampling in forensics

Sampling is defined in ISO/IEC 17,025 as:

"*a defined procedure whereby a part of a substance, material or product is taken to provide for testing or calibration of a representative sample of the whole.*"

Sampling is a well established scientific technique and a fundamental principle of many research methodologies. It is introduced in many science textbooks, in fields ranging from sociology (Browne, 2006) to physics (Bohm and Zech, 2010). Sampling is also used within other forensic disciplines for example toxicology (Levine, 2003), drug analysis (United Nations Office on Drugs and Crime, 2009, pp. 7–27) and biology (Budowle et al., 2006). However the purposes for which sampling is used may be quite different.

In the physical forensic sciences sampling is most commonly used to establish the likelihood or probability that the accused can be linked to the crime scene (Saferstein, 2007, p. 346). One of the most frequent and powerful uses of this is with Deoxyribo Nucleic Acid (DNA). There are a range of tests that may be used with DNA, with each test providing a probability that a sample belongs to that of a single person, it is a powerful evidentry tool (Saferstein, 2007, p. 346; Aytugrul v R, 2010).

The other use of sampling and statistics in forensics is to determine the volume or extent of a crime. For example when a large quantity of drugs are seized a sample of them will be analysed and deemed to be representative of the whole (Fraser, 2010). The use of this type of sampling has also been seen in forensic accounting (MBIA Ins. Corp. v. Countrywide Home Loans, Inc., 2010). In MBIA Insurance Corporation v. Countrywide Home Loans, INC. (MBIA Ins. Corp. v. Countrywide Home Loans, Inc., 2010) it was argued that statistical sampling did not provide an accurate representation of the population. The court found that "*Statistical sampling is not novel*" and "*statistical sampling is generally accepted in the scientific community*".

The sampling methodology discussed in this paper is used within a similar context to the manner it is used in drug sampling and (MBIA Ins. Corp. v. Countrywide Home Loans, Inc., 2010).

3. Theoretical framework

The goal of any statistical methodology is to be both reliable and valid. For the purposes of CAM sampling the objective is to view a sample of the files on a storage device and determine the proportions of CAM to other files. In order to be valid the sample must be representative of the population. The reliability of the results are to a large part dependent upon the individual selecting the material. This and the following section will discuss the validity of the sampling methodology and associated confidence levels and error rates used to project observations of the sample to the entire population. Methods of ensuring reliability will be discussed in section five.

One significant advantage to sampling a digital environment is the high level of control available over that environment and source population. With modern digital forensic software it is possible to identify file types of interest on a storage device. This provides us with a known population size. Thus enabling sample size to be set to achieve a predetermined confidence level.

3.1. Sampling methodologies

A sample is a representative subset of a population. The sample is examined with the expectation of gaining, in our case, quantitative data about that population – or an estimate. "*A sample is representative if the statistics computed from it accurately reflect the corresponding population parameters.*" (DeVeaux et al., 2009).

To allow for the above estimation the population must be randomly sorted and a minimum recommended number of items are to be presented – the sample set. In statistics this can be modelled as "simple random sampling".

"With simple random sampling, each member of the sampling frame has an equal chance of selection and each possible sample of a given size has an equal chance of being selected. Every member of the sampling frame is numbered sequentially and a random selection process is applied to the numbers." (McLennan, 1999).

For the purposes of determining the ratio of files containing CAM to files not containing CAM on a digital storage device simple random sampling is relativity straight forward. By assigning a random number to each file the files can then be sorted and the sample selected from the first n files.

3.2. Sample size

This is the number of items in a sample taken from the population. The recommended sample size to facilitate a confidence level of 99% and margin of error less than 5% can be calculated before the process begins. This is achieved by using Yamane's formula (Yamane, 1967):

Due to the law of large numbers and the Central Limit Theorem in order to achieve a confidence level of 99%, a maximum of approximately 10,000 files (when using the Discovery statistical constraints) have to be viewed – irrespective of the population size.

3.3. Sampling error

When an estimate is derived from a sample there are certain errors that occur, they are:

- Non-sampling errors which can occur due to human biases and errors.
- Sampling errors:"*Sampling error reflects the difference between an estimate derived from a survey and the 'true value' that would be obtained if the whole target population were included. If sampling principles are applied*

carefully, sampling error can be kept to a minimum." (McLennan, 1999).

In this case we are only concerned with sampling errors, as non-sampling errors are minimal due to the controlled environment and uniformity of the population.

3.4. Confidence level

This value indicates the reliability of the estimate. If the confidence level is 95%, this implies that if 100 samples were conducted, 95 of them would fall between the required confidence interval (Stewart, 2011).

3.5. Confidence interval

The confidence interval is the required precision of the estimate. For example if the confidence interval is ±1%, the sample is 10,000, the selection is 1000 and the population is 100,000 the estimated number of pictures on the hard disk would fall between the ranges of 9000 to 11,000. The required confidence interval and the resulting confidence interval may differ due to the actual sample size examined and the number of items selected.

Most survey samples in health, government and industry use a confidence level of at least 95% and a margin of error of less than 5%. Because of the seriousness of cases law enforcement deals with, it is desirable to achieve a 99% confidence level and a target confidence interval of less than 5% (Tang et al., 2009).

At this point it is important to reiterate that these statistics are not used for establishment of guilt, rather for the purpose of determining an appropriate sentence once guilt is proven. For the purposes of proving possession the existence of a single file containing CAM may be sufficient. The purpose of establishing an estimate of the quantity of material is to aid the court in determining the seriousness of the offence and thus an appropriate sentence. We can achieve such a high confidence level because of the controlled environment, the absence of bias and non-response and ease of access to large sample sets.

3.6. Confidence interval calculation

Once the recommended sample set has been viewed, and the selections made an estimate must be calculated. The resulting confidence interval and confidence level will probably not be equal to the required values used to calculate the recommended sample size. This is due to several variables that come in to play when calculating the confidence interval, such as:

- The actual sample viewed may be less than the required sample.
- The actual sample viewed may be more than the required sample. This will have the effect of improving the estimate.
- The number of items selected, or successes. This value does affect the final estimate, but only when the selection is less than ~0.05% of the population. The relative standard error is also calculated and indicates whether the estimate is acceptable. The recommendation for the values of the relative standard error are
  - <=25% – an acceptable estimate
  - >25% <= 50% – it is an acceptable estimate, but it should come with a warning.
  - >50% the estimate should not be considered reliable (Kazimer, 1996).

3.7. The standard (Wald) interval

To calculate the estimate, you must first calculate the interval estimation of the proportion, or the confidence interval. The standard formula for calculating the confidence interval is known in most introductory statistics textbooks as the Wald interval.

3.8. The Adjusted Wald (Agresti–Coull) interval

As the standard interval has been proven to be inconsistent and unreliable in many circumstances, it is recommended by several sources that the Adjusted Wald is a more reliable interval calculation (Australian Bureau of Statistics, 2010; Agresti and Coull, 1998; Brown et al., 2001; Sauro and Lewis, 2005).

The Adjusted Wald interval is defined as:

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\tilde{p}\left(1-\tilde{p}\right)} \times \frac{1}{\sqrt{\tilde{n}}}$$

Where:

z = confidence level (for 99% ~ 2.58)
n = sample size
s = selection

$$\tilde{p} = \frac{s + \frac{z^2}{2}}{n + z^2}$$

$$\tilde{n} = n + z^2$$

We have also conducted simulations and testing using the Adjusted Wald, Wilson & Jeffreys Bayesian interval and found that the Adjusted Wald was the most reliable and gave acceptable conservative estimates – it is also one of the easiest interval calculations to represent. Tests have also shown that it also has the best coverage or coverage probability in relation to the confidence level.

3.9. Proportion

The percentage of items in the population that are expected to have the qualities being evaluated (Stewart, 2011). If the proportion is unknown it should be set to 0.5. This results in the most conservative estimate and the largest sample size.

3.10. Finite population correction

If the population is known and the sample is greater than 5% of the population, the finite population correction factor can be used. The FPC is included into the estimate calculation and usually results in a narrower margin for the estimate, without affecting the reliability. The formula for the FPC is:

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

3.11. Point estimates

The point estimate is a ratio of the selection in relation to the sample or s/n. It is a value that is used in the final estimate calculation. There are 4 main Binomial point estimators: The Maximum likelihood estimate (MLE), Laplace, Wilson and Jeffreys (Böhning and Viwatwongkasem, 2005). The point estimates usually result in similar outcomes unless there is an unusually small selection in comparison to the population.

4. Testing

An Encase EnScript© was written to automate and simulate the process of creating a random sample of pictures, sorting the sample by random number and simulating the selection of those pictures – as would be done in a CAM investigation. The actual picture count was known beforehand for comparisons and coverage results. Several real cases have also been tested, by manually going through all available pictures and selecting pictures of interest – the results from these tests improved upon the automated test results. Two mock test cases were used for the automated processing. This was appropriate as the test is a statistical function and does not consider the type of data to be tested.

The individual CETS (Royal Canadian Mounted Police, 2008) scale for each image was not considered for any of these tests as it is purely, at this stage, a quantitative test. It may be a future option to provide estimates for each CETS scale for selected items. The final report does give the percentages for each CETS scale of the selected items – but no estimates over the total population are given.

The test constraints were:

- Picture populations between 5488 and 520,610.
- The recommended sample or a selection of 300 (If this came before the recommended sample and appropriate).
- Actual picture of interest counts between 3 and 47,000.
- Actual percentage of pictures of interest compared to population between 0.0057% and 50%.
- Estimate calculations using The Adjusted Wald, Jeffreys and Wilson interval methods.
- Using the finite population correction factor or not.
- Using the Maximum likelihood estimate, Laplace, Jeffreys and Wilson point estimators.

For the different combination of constraints above, 1000–10,000 iterations were done for each. These tests then demonstrated the average coverage probability and an average actual margin of error. For an example of an individual test with constraints and 10, 000 iterations – See Table 1.

From a total of 256 individual tests (As in the examples) with 1000 or 10,000 iterations it was found that the Laplace point estimator using the finite population correction factor produced the best overall results for each interval calculation – the averaging of the results are demonstrated in Table 2.

As can be seen from Table 2 the Adjusted Wald and Wilson are the most reliable and the Jeffreys has a smaller margin of error and narrower margins for the estimate. These are averaged out from all tests. The differences in the interval methods are minimal when the selection is above ~0.05% of the population and the recommended sample size is adhered to, however when this is not the case the differences can be magnified.

As the Adjusted Wald is the most reliable for all values (Its coverage probability is between 98% and 100%) it was selected for use in the SEEB sampling process. As testing continues the Jeffreys interval calculation may become an option – all are acceptable interval calculations in theory and, as a result of our testing, also in practical terms.

5. Implementation of sampling at SEEB

In order to implement the use of sampling for determining the quantity and ratio of CAM on a device, SEEB

Table 1
Test results. Restraints: CL – 99%, CI < 5%, population – 52,061, sample – 8388, actual items of interest – 1527.

Adjusted Wald				Jeffreys				Wilson				Point estimator	FPC
Coverage (%)	Average estimate	Average margin	Average margin of error (+−%)	Coverage	Average estimate	Average margin	Average margin of error (+−%)	Coverage	Average estimate	Average margin	Average margin of error (+−%)		
99.57	1298–1795	497	0.47715	99.46	1296–1787	491	0.46091	99.52	1299–1795	495	0.4757	Wilson	No
99.54	1287–1781	495	0.47522	99.55	1295–1789	493	0.48917	99.54	1286–1782	495	0.47579	Laplace	No
99.42	1283–1777	494	0.47464	99.48	1294–1788	494	0.49512	99.42	1283–1778	495	0.47566	Jeffreys	No
99.5	1283–1777	494	0.47457	99.5	1294–1787	494	0.49504	99.5	1282–1777	495	0.47559	MLE	No
99.13	1319–1774	455	0.43694	99.15	1313–1762	449	0.41497	99.13	1319–1773	454	0.43561	Wilson	Yes
98.98	1306–1759	453	0.43499	99.04	1311–1763	451	0.44281	99.12	1305–1759	453	0.43552	Laplace	Yes
98.99	1303–1755	452	0.43457	99.1	1311–1763	452	0.44883	98.99	1302–1756	453	0.4355	Jeffreys	Yes
98.86	1305–1758	453	0.43489	98.85	1313–1765	452	0.44916	98.86	1304–1758	454	0.43582	MLE	Yes

developed the "SEEB Discovery process". The SEEB Discovery process was designed to take advantage of statistical random sampling to limit the amount of data to be reviewed by the analyst or investigator, automating most of the pre- and post-processing of electronic data and providing "zero-skill" tools that investigators can use to review the content of their exhibits with minimal training.

In the Discovery process, the SEEB digital forensic analyst has limited involvement in the overall examination. Their role is to manage the flow of investigators and exhibits through the Discovery room, prepare exhibits for forensic examination (using "write-blocking" software and/or hardware) and provide technical support where necessary. Due to their reduced involvement in the forensic process, a single analyst can efficiently service multiple concurrent investigations. In the current implementation at SEEB, a single analyst can be responsible for up to six investigators working on multiple exhibits on any single day.

The sampling process has been developed using the *EnScript* feature in the EnCase (Guidance Software Encase Forensic v7 [WWW Document], 2012) application and has been tested on version 6.19 running on Microsoft Windows XP SP3, Vista and Windows 7 (32-bit and 64-bit). Additionally, it requires Microsoft Word/Excel 2003 (Microsoft, 2012) (or above) for generating reports and ImgBurn (ImgBurn, 2012) for creating optical media backups.

The Discovery process is a relatively linear process that repeats when there are multiple related exhibits in an investigation. A high level flowchart of this process is presented in Fig. 2. The "Add Exhibit" phase encompasses the preparing and adding of an exhibit to EnCase and is performed by the SEEB analyst. This includes following standard digital forensic practice in exhibit handling and "write protecting" storage devices. For laptops and computers that contain multiple hard disk drives (HDDs), it is recommended for simplicity and speed that *LinEn* is used to allow all internal HDDs to be examined at once. A Linux-based forensic boot CD that automatically launches a pre-configured copy of *LinEn* to communicate with EnCase has also been developed as part of the process.

The "Initialise Case" phase incorporates the automated pre-processing of data and generation of the sample set for review. There is scope for customising the level of processing based on the total amount of data that requires examination. As a baseline, operating system information, user account details and mounted devices are extracted from registry files. In order to obtain the minimum sample set, the total number of files or population needs to be ascertained.

A signature analysis is conducted to produce a more accurate count of pictures, video and document files. Files that are deleted, in unallocated space or located in Internet cache are excluded from the total file population. This is because they are not considered admissible evidence for the purposes of possession due to NSW Case Law (Director of Public Prosecutions v Kear, 2006). As part of the Initialise case, Internet cache, deleted files, non-standard images (RAW camera images) are bookmarked separately and available for viewing – but are not part of the sample. The analyst can also make an assessment on whether to process files embedded in archives (ZIP, RAR, etc.) and other compound files.

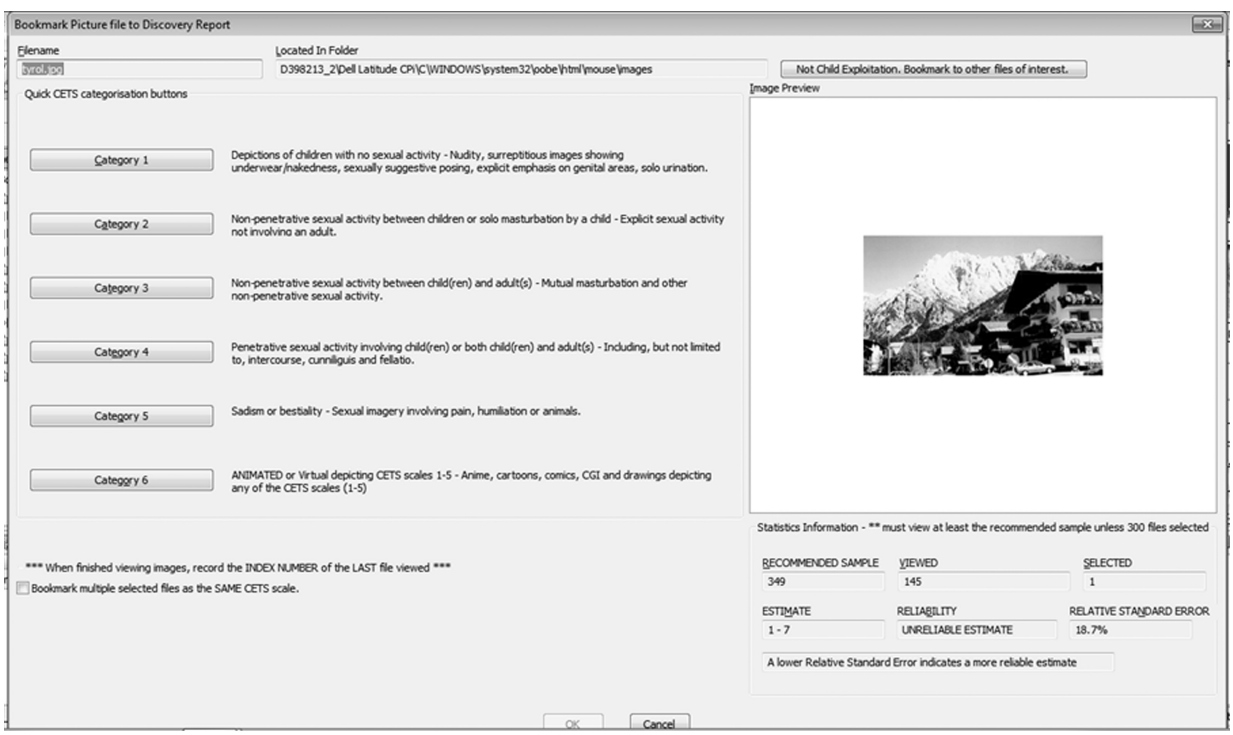

Table 2
Average test results.

	Adjusted Wald			Jeffreys			Wilson		
	Coverage (%)	Average margin	Average margin of error %	Coverage (%)	Average margin	Average margin of error %	Coverage (%)	Average margin	Average margin of error %
Average	99.50	418.48	0.43	98.69	402.22	0.42	99.59	425.41	0.43
Min	98.29	34.00	0.04	95.97	26.00	0.03	98.60	37.00	0.04
Max	100.00	2671.00	4.15	99.77	2671.00	4.18	100.00	2671.00	4.15

Once the parameters of the statistical sample are set (confidence interval, margin of error, etc.), the recommended number of files that need to be viewed is calculated using formulae presented in this paper. A random sample is then obtained from files identified on the system and presented to the investigator for review.

The "*Child Exploitation Tracking System*" (CETS) scale (Table 3) is used for categorising the CAM by the level of its severity. This scale is used in both State and Federal jurisdictions of Australia. It is also anticipated that by using the CETS scale it will streamline future integration with the *Australian National Victim Image Library* (ANVIL). Prior to commencing classification, investigators review the CETS scale with the SEEB analyst to help to improve the reliability of the classification process.

Usability is a key factor in the "Bookmark Files" stage to ensure speed and consistency when categorising files. The EnCase gallery view is used to allow investigators the ability to quickly scan the image sample set. When a file of interest or several files of interest are identified, investigators can use numerous methods to bookmark and categorise those files. The primary method is through a bookmarking EnScript, and is accessed through a toolbar button or keyboard shortcut.

For image files, the bookmarking EnScript presents a dialogue box containing a preview of the file and list of buttons that correspond to the different CETS levels. Descriptions of each of the CETS levels are included next to these buttons to minimise any confusion as to the correct level for classification. For video and document files, the investigator is also required to provide a synopsis of the viewed content into a text box.

Once the investigator has completed their review of the sample set, an EnScript allows them to undertake a final review of the classified files and the percentage ratios of files in each CETS level. The classified files are then exported into a folder structure on the forensic workstation along with comprehensive file meta-data in an XML format. If there are no further exhibits to be reviewed, the Discovery process then automates the post-processing phase. This includes generating a court report, a statement outlining the discovery process and a results disc. Additionally, a backup of files generated in the discovery process is also created and retained by SEEB in the event the case needs to be revisited at a later time.

The court report contains detailed information relating to the case/exhibit (operating system information, storage devices, and user accounts), the classified files with associated meta-data (filenames, full path, MAC timestamps, MD5 hash value) and an overview of the statistical analysis. The statistical analysis provides a breakdown of the total number of files located in each CETS level, a percentage value of CAM within the sample set, an estimate of the total number of CAM on the exhibit and the estimated error rates.

In the event a low number or no CAM files are identified the SEEB analyst will conduct a preliminary examination of the system. This will seek to determine if files may be hidden on the system, or if it has been used to access CAM. A decision will then be made on whether additional analysis is required. The methodology of both the preliminary examination and further analysis are beyond the scope of this paper.

6. Results and outcomes

The SEEB Discovery process was first implemented in 2009 and has undergone considerable refinement since that time. A dedicated room containing secure exhibit storage, six examination workstations, dedicated write blocking equipment and detailed operating procedures has been established. This makes it possible for a single analyst to support up to four investigations at a time.

As a result of the streamlined process and the ability to run investigations in parallel client waiting times have reduced from 3 months to 24 h. On average a computer can be processed in under two hours, compared to an average processing time approaching five hours previously.

While SEEB staff are exposed to the CAM they are not required to examine it in-depth. This is a significant improvement over the previous methodology where investigators and SEEB staff may be exposed to hundreds of thousands of pictures, videos and documents of distressing material.

The clear reports generated by the process have been received favourably in the courts. As a result of this work and efforts by the SEEB senior management the NSW Criminal Procedures Act has been amended to accept the results of random sampling as evidence in CAM cases (Criminal procedure act 1986 – sect 289b, 2012).

The fast response time has also benefited the suspects and those falsely accessed. In the past once a computer was seized the suspect may have to wait for an extensive period of time before the existence of CAM was established. This would have a significant detrimental effect on the suspects' family and work life. Especially in cases where the suspect was innocent this would impose a significant burden upon them.

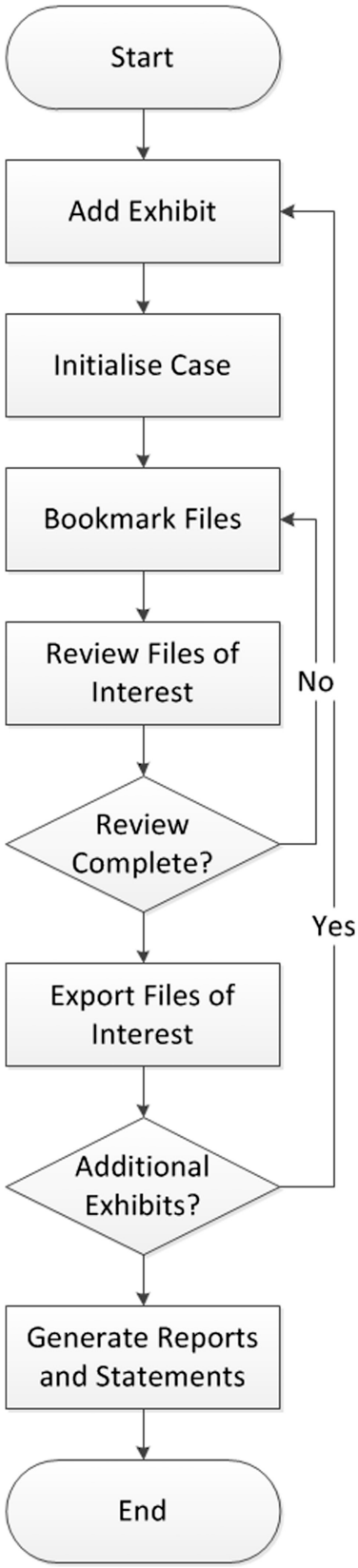

Fig. 2. SEEB discovery workflow.

Table 3
CETS CAM/CEM scale.

Category	Description	Guide
1	No sexual activity	Depictions of children with no sexual activity – Nudity, surreptitious images showing underwear nakedness, sexually suggestive posing, explicit emphasis on genital areas, solo urination.
2	Child non-penetrate	Non-penetrative sexual activity between children or solo masturbation by a child.
3	Adult non-penetrate	Non-penetrative sexual activity between child(ren) and adult(s). Mutual masturbation and other non-penetrative sexual activity.
4	Child/adult penetrate	Penetrative sexual activity between child(ren) or between child(ren) and adult(s) – Including, but not limited to, intercourse, cunnilingus and fellatio.
5	Sadism/ bestiality/ child abuse	Sadism, bestiality or humiliation (urination, defecation, vomit, bondage etc.) or child abuse as per CCA 199.
6	Animated or virtual	Anime, cartoons, comics and drawings depicting children engaged in sexual poses or activity.

7. Other considerations and future work

It must be stressed that the statistics obtain from the SEEB Discovery process are only used for determining the quantity and type of CAM on a storage device. In investigations where indications of the suspect having access to a child, distribution of CAM or data hiding further digital forensic analysis is performed.

Looking forward there is the potential to apply this methodology to other types of investigations, including other CAM offences such as misuse of carriage service and unrelated offences including drug and fraud investigations.

8. Conclusion

By using random sampling to search for CAM files SEEB has been able to significantly reduce the exposure of its staff and police investigators to disturbing child abuse material. It has also significantly reduced backlogs, enables investigators to establish the extent of their investigation in a short timeframe and provides the courts with a clear record of the quantity and severity of child abuse material on a device.

References

Abel G, Becker J, Cunningham-Rathner J, Mittleman M, Rouleau J. Multiple paraphilic diagnoses among sex offenders. The Bulletin of the American Academy of Psychiatry and the Law 1988;16:153–68.

Agresti Alan, Coull Brent A. Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician 1998; 52(2):119–26.

Australian Bureau of Statistics. What is a standard error and relative standard error, reliability of estimates for labour force data. Sydney, NSW: Australian Bureau of Statistics; 2010.

Aytugrul v R. NSWCCA 272 (3 December 2010). Supreme Court of New South Wales – Court of Criminal Appeal; 2010.

Bohm Gerhard, Zech Günter. Introduction to statistics and data. Hamburg: Deutsches Elektronen-Synchrotron; 2010.
Böhning, Viwatwongkasem. Revisiting proportion estimators; 2005. Berlin.
Brown D Lawrence, Cai Tony T, DasGupta Anriban. Interval estimation for a binomial proportion. Statistical Science 2001;16(2):101–33.
Browne Ken. An introduction to sociology. Malden: Polity Press; 2006.
Budowle B, Schutzer SE, Burans JP, Beecher DJ, Cebula TA, Chakraborty R, et al. Quality sample collection, handling, and preservation for an effective microbial forensics program. Applied and Environmental Microbiology 2006;72:6431.
CRIMES ACT. Sect 91H. New South Wales; 1900.
Criminal procedure act 1986 – sect 289Þ; 2012. New South Wales.
DeVeaux, Vellman, Bock. Intro Stats. 3rd ed. Boston: Pearson Education; 2009.
Director of Public Prosecutions v Kear. NSWSC 1145; 2006.
Fraser Jim. Forensic science: a very short introduction. Oxford University Press; 2010.
Gosling A v DPP. NSWDC 93 (23 January 2009); 2009.
Guidance software encase forensic v7 [WWW Document]. URL: http://www.digitalintelligence.com/software/guidancesoftware/encase7/; 2012.
Holt Thomas, Blevins Kristie. Examining job stress and satisfaction among digital forensic examiners. Journal of Contemporary Criminal Justice 2011;27(2):230–50.
ImgBurn. The official ImgBurn website [WWW Document]. URL: http://www.imgburn.com/; 2012.
Jewkes Y, Andrews C. Policing the filth: the problems of investigating online child pornography in England and Wales. Policing and Society 2005;15:42–62.
Kazimer Leonard J. Business statistics. 3rd ed. McGraw-Hill; 1996.
Krause Meredith. Identifying and managing stress in child pornography and child exploitation investigators. Journal of Police and Criminal Psychology; 2009:22–9.
Levine Barry, editor. Principles of forensic toxicology. USA: American Association for Clinical Chemistry; 2003.
MBIA Ins. Corp. v. Countrywide Home Loans, Inc. NY Slip OP 52239 – NY: Supreme Court; 2010.
McLennan W. An introduction to sample surveys: a users guide. Australian Statistician, Australian Bureau of Statistics; 1999.
Microsoft. Office – office.com [WWW Document]. URL: http://office.microsoft.com/en-us/; 2012.
Perez Lisa, Jones Jeremy, Englert David, Sachau Daniel. Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media. Journal of Police and Criminal Psychology; 2010:113–24.
Puhakka v R. NSWCCA 290 (10 December 2009); 2009.
R v Elliott DB. NSWDC 238 (24 October 2008); 2008.
R v Gent. NSWCCA 370; 162 A CRIM R 29 (4 November 2005); 2005.
R v Leonard WG (NO 3). NSWDC 211 (5 September 2008); 2008.
Royal Canadian Mounted Police. Child exploitation tracking system [WWW Document]. URL: http://www.rcmp-grc.gc.ca/ncecc-cncee/cets-eng.htm; 2008.
Saferstein. Criminalistics. 9th ed. Upper Saddle River, NJ: Pearson; 2007.
Sauro J, Lewis J., 2005. Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. In: Proceedings of the human factors and ergonomics society 49th annual meeting (Denver 2005), 2100–2104.
Sivell AJ v R. NSWCCA 286 (3 December 2009); 2009.
Stewart D. Application of Simple Random Sampling(SRS) in eDiscovery. In: Manuscript submitted to the organizing committee of the fourth DESI workshop on setting standards for electronically stored information in eDiscovery proceedings on April 20, 2011, Daegis, 6; April 20, 2011.
Stewart J, Black G., 2012. Forensic clusters: advanced processing with open source software. In: DOD CyberCrime conference (Atlanta 2012).
Tang Man-Lai, Ling Man-Ho, Linga Leevan, Tian Guoliang. Confidence intervals for a difference between proportions based on paired data. Statistics in Medicine; 2009.
United Nations Office on Drugs and Crime. Guidelines on representative drug sampling. New York: United Nations; 2009.
Whiley v R. NSWCCA 53; 2010.
Wolak Janis, Mitchell Kimberly. Work exposure to child pornography in ICAC task forces and affiliates. Durham; 2009.
Yamane Taro. Statistics, an introductory analysis. 2nd ed. New York: Harper and Row; 1967.

Syd Pleno is an electronic evidence specialist/software engineer with almost nine years of computer forensic experience at state and federal law enforcement agencies. He spent over seven and a half years as a founding member at the State Electronic Evidence Branch (SEEB) establishing computer forensic capability for the NSW Police Force. During his time at SEEB, he spent a significant portion of it in the Research and Development section, holding the position of Team Leader (Research and Development). His focus was on developing cutting edge solutions for digital forensics, reverse engineering file formats and working on in-depth forensic examinations. His last major project at SEEB was the design, implementation, testing and deployment of the Discovery Process for child exploitation investigations.

In 2011, he took up employment with the Australian Federal Police as a Senior Computer Forensic Examiner where he works on serious, major and organised criminal investigations. Syd is a Certified Forensic Computer Examiner (CFCE) and an instructor/coach for the International Association of Computer Investigative Specialists. His tertiary qualifications include a Graduate Diploma in Science (Information Assurance) from the University of South Australia and a Bachelor of Computer Science and Technology from the University of Sydney.

Brian Jones is a software engineer in the research and development area of the State Electronic Evidence Branch (SEEB) of the New South Wales Police force, Australia. This role entails the provision of high level forensic analysis of digital data associated with the commission of serious, major and organised crime and providing tools to assist with the presentation, processing and analysis of electronic evidence. Prior to this he was an electronic evidence specialist at SEEB for 3 years and successfully completed hundreds of forensic examinations on electronic evidence, including Computers, mobile phones, PDA's, external storage devices, cameras and CCTV systems. He was also a serving member of the New South Wales Police force for 8 years, attaining the level of senior constable.

During his time at SEEB he has designed and implemented numerous scripts and software tools to assist with the presentation, processing and analysis of electronic evidence for SEEB operatives, New South Wales Police and several other agencies. He also played a major role in the implementation and testing of the statistical portion of the SEEB Discovery process.

Brian has a Bachelor of Computer Science, from the University of Newcastle, A Diploma of Policing Practice from Charles Sturt University and is currently undertaking a Graduate Certificate in Computer Forensics at Macquarie University, Sydney.

Michael Wilkinson is the Program Director of both the M.S. Digital Forensic Management and M.S. Digital Forensic Science Programs at Champlain College. He teaches both undergraduate and graduate courses and is a Co-Director of the Senator Leahy Center for Digital Investigation which provides investigative and research services to Law Enforcement, Government agencies and the general public. Prior to joining Champlain College Michael was a coordinator of the New South Wales Police Force, State Electronic Evidence Branch, responsible for the management and oversight of all forensic analysis work performed by the Branch. In his time with the NSW Police Force Michael has conducted hundreds of examinations and regularly presented evidence in court. Michael has been actively involved in the development of the Digital Forensics profession through his involvement with the National Institute of Forensic Science. Where he was involved in the creation of national competency and validation standards. He is also a technical assessor with NATA the Australian national body for ISO17025 forensic laboratory accreditation.

Michael has been involved in education since 2001 when he was appointed to a Lecturer position at the Australian Catholic University, where he taught programming, data communications and information system security. While with the NSW Police Force he assisted Macquarie University and the University of New South Wales in developing postgraduate programs in digital forensics and taught Network Forensics at the University of South Australia, where he is also pursuing a PhD.