Minimising Collateral Damage: Privacy-Preserving Investigative Data Acquisition Platform

Zbigniew Kwecka
School of Computing, Edinburgh Napier University, UK

….
Edinburgh Napier University, UK

…..
Edinburgh Napier University, UK

ABSTRACT

Investigators define invasion of privacy during their operations as collateral damage. Inquiries that require gathering data about potential suspects from third parties, such as banks, Internet Service Providers (ISPs) or employers are likely to impact the relation between the data subject and the data controller. In this research a novel privacy-preserving approach to mitigating collateral damage during the acquisition process is presented and Investigative Data Acquisition Platform (IDAP) is defined. IDAP is an efficient symmetric Private Information Retrieval (PIR) protocol optimised for the specific purpose of facilitating public authorities' enquiries for evidence. This research introduces a semi-trusted *proxy* into the PIR process in order to gain the acceptance of the general public for the trap-door based privacy-preserving techniques. Then the *dilution factor* is defined as a level of anonymity required in a given investigation. Defining this factor allows restricting the number of records processed, and therefore, minimising the processing time while maintaining an appropriate level of privacy. Finally, the technique allowing retrieval of records matching multiple selection criteria is described.

Keywords: Privacy Enhancing Technology; Data Mining; Data Retrieval;

I. INTRODUCTION

> *Those who would give up essential Liberty, to purchase a little temporary safety, deserve neither Liberty nor Safety. (Benjamin Franklin 11 Nov 1755)*

Since the fall of 2001 many western governments have passed laws giving public authorities wider rights to gather operational data (Home Office, 2009; Swire & Steinfeld, 2002; Young, Kathleen, Joshua, & Meredith, 2006). For a number of years the public opinion accepted privacy

intrusions as the sacrifice everybody must make to *fight the terror* (Rasmussen Reports, 2008). However, slowly the public opinion is shifting back to the state were intrusion of privacy is considered as unacceptable. This is shown by different surveys such as the one conducted by Washington Post in 2006 (Balz & Deane, 2006), where 32% of respondents agreed that they would prefer federal government to ensure that privacy rights are respected rather than to investigate possible terrorism threats. This was 11% increase from the similar survey conducted in 2003.

In the UK the public authorities such as Police, Customs, and Tax Offices need to request information from third-parties on regular basis and the data protection legislations allow for such requests even without warrants (EUROPEAN PARLIAMENT, 1995; Home Office, 2007). Depending on the way these requests are performed human and natural rights of the data-subject can be breached and/or investigation can be jeopardized (Kwecka, Buchanan, Spiers, & Saliou, 2008). A recent proposal by the UK government went further and recommended allowing the public authorities direct access to data held by Content Service Providers (CSPs), such as mobile telephony providers and Internet Service Providers (ISPs) (Home Office, 2009). According to the public consultation document there were a few major motivating factors behind this proposal, including: increasing access speeds to records; allowing for secret enquiries for antiterrorist and national security purposes; lowering collateral damage to potential suspects being investigated; and analysing patterns in the data to allow profiling of terrorists. The concerns were raised that if the proposal was implemented it would thwart privacy of the internet users around the globe, in order to increase security on the nation. This research shows that most objectives of the proposal can be achieved maintaining high level of privacy. It is shown that an investigative system can provide high level of privacy to the data subjects and preserve the confidentiality of investigations. However, both security and privacy must be built into the system from the design stage, as Swire and Steinfeld (2002) prove on the example of Health Insurance Portability and Accountability Act in the US health system.

This research gives an insight on how the Privacy Enhancing Technologies (PETs) can be used to improve current investigative data acquisition practices. The proposed Investigative Data Acquisition Platform (IDAP) is a novel, efficient approached to maintain the secrecy of an enquiry, preserving suspect's privacy and gaining public's support for the PET technologies. Section II describes requirements that IDAP must meet in order to be accepted for the intended use, later the background to this work is presented, followed by introduction of the privacy-preserving building blocks used to construct the platform. Finally advantages and disadvantages of building IDAP on an already existing Private Equijoin (PE) protocol are given, resulting in recommendations for improvements that are addressed in Section VI.

II. IDAP REQUIREMENTS

The public authorities are often required to carry out investigations based on data supplied by third parties. Such investigations may include benefit fraud enquiries from HMRC, solving a crime by Police, investigating alleged terrorism cases by Scotland Yard, or gathering health information about a patient at Accident and Emergency department. The process of obtaining third party records is usually referred to as *data acquisition*. In the UK there are a two major data acquisition legislations available to the public authorities, these are: Data Protection Act 1998

(DPA) and The Regulation of Investigatory Powers Act 2000 (RIPA), but similar legislations can be found across Europe. Depending on the nature of the investigation and the type of data required, the public authorities choose between the above legislations while preparing their data acquisition request. In order to prototype the system and test the concept of IDAP a set of initial system requirements have been gathered from publicly available literature, including the afore mentioned legislations, as well as data acquisition and protection guidelines (Home Office, 2007; Information Commissioner, 2007), and articles relating to computer forensic investigations (Association of Chief Police Officers, 2003; Palmer, 2001). These assumed requirements were as follows:

1. In some cases there is more than one suspect in a forensic inquiry.
2. Investigators need to provide justification for the acquisition requests under DPA and the dataholder can refuse providing any data without a warrant.
3. Data acquisition notices served under RIPA do not need any form of justification to the dataholder and the dataholder will face penalty if the relevant data is not provided to requesting public authority within two weeks. Still, the dataholder may choose to accept the penalty and refuse to provide any data without subpoena.
4. Any evidence collected may need to be presented in front of court of law. If such requirement arises the electronic evidence must be provided as a true image of the data gathered.
5. Under RIPA the public authorities must make a contribution towards the costs incurred by a CSP during fulfilling the data acquisition notice.

Based on the gather requirements, a protocol chosen for data gathering should allow retrieval of a number of records at the time. If this is not the case multiple sequential runs of the protocol should bear low computational and communicational overhead. The protocol must leave the dataholder in control of the data, since the data retrieval can only be performed with the dataholder's consent. Taking into consideration that a CSP has two weeks to provide the data, but other data controllers are not obliged to provide any data the computational complexity of the protocol can be reasonably large. However, data should be retrieved from the dataholder on record-by-record basis, so that if only one of many records provided as a response to an enquiry needs to be submitted to court other records can be discarded. Otherwise the public authorities can end up storing large amount of unnecessary data, and this can prove costly taking into consideration the level of security and auditing involved. Finally the cost of the solution should be low since the public authorities will have to cover the costs of running the system. If the costs were not covered by the authorities, the dataholders would transfer the costs of handling the enquiries to the end-users and such solution would be unacceptable.

The next section describes the concepts and the technologies that are the basis for the creation of IDAP.

III. BACKGROUNG AND RELATED WORK

Leaving the investigative context aside, the retrieval of information from a third-party in a private manner is a generic problem that has been researched for use in a variety of different

scenarios. Initially, Private Information Retrieval (PIR) protocols were designed with a basic requirement of acquiring an interesting data record, or just a specific data bit, from a dataholder, *sender*, in a way that this dataholder is unable to judge which record is of interest to the requestor, *chooser*. These protocols were not concerned with the secrecy of the records stored in the database, thus in its least optimised state a PIR could have been achieved by transferring the whole database from the *sender* to the *chooser*, as this would allow the *chooser* to retrieve a record in a private manner. Consequently, the main motivation behind the PIR schemes is achievement of a minimal communicational and computational complexity (Ostrovsky & William E. Skeith III, 2007). A stronger notion than PIR is *1-out-of-n* Oblivious Transfer (OT) primitive that allows the retrieval of a randomly selected record from the dataset of *n* elements held by the *sender* in a way that the *sender* cannot learn which record has been transferred, and the *chooser* cannot learn anything about other records in the dataset (Schneier, 1995). *1-out-of-n* OT protocols that allow *chooser* to actively select a record to be retrieved, and that have linear or sub-linear complexity, can be referred to as symmetric PIR (SPIR) protocols, since they protect the records of both parties during the information retrieval. These useful privacy-preserving data retrieval protocols can be employed in a variety of systems: electronic watch-lists of suspects (Frikken & Atallah, 2003); cooperative scientific computation (Du & Atallah, 2001; Goldwasser & Lindell, 2002); and on-line auctions (Cachin, 1999).

With the use of the protocols described in the above paragraph a *chooser* would be capable of privately retrieving a record from the *sender's* database, by secretly referring to its index in this database. In SPIR such index is expected to be publically available in an electronic catalogue or a directory (Aiello, Ishai, & Reingold, 2001; Bao & Deng, 2001). However, ISPs and other dataholders with large databases of private data cannot be expected to maintain such freely available indexes. Also, it is expected that an investigator would normally refer to a suspect by name, ID or phone number, etc. For this reason before the data can be received using SPIR, a search would need to be performed by the *chooser* against the records in the *sender's* database. Such a private search operation requires a protocol that allows two parties to compare their values in a private manner. The protocols that are optimised to make comparisons for equality are referred to as Private Equality Test (PEqT) protocols. PEqT protocols are often based on commutative (Frikken & Atallah, 2003; Kwecka et. al. 2008) or homomorphic cryptosystems (Bao & Deng, 2001).

An interesting record can be located in a database using a *1-out-of-n* PEqT protocol and then retrieved with help of SPIR. Often each of these protocols would have a separate computationally expensive preparation phases, such solution would not be optimal for IDAP. The exception to this rule is a range of protocols including: private intersection; private intersection size; and PE defined in (Agrawal, Evfimievski, & Srikant, 2003)[r161]. These protocols are based on commutative encryption and thanks to the use of different properties of the underlying commutative algorithms are capable of allowing for both private matching and private data retrieval. The operation of the PE is later described in Section IV, while a brief introduction of the cryptographic mechanisms used by PE and IDAP follow.

IV. BUILDING BLOCKS

This section describes PE protocol that is the basis for creation of the privacy preserving investigative platform - IDAP. The PE protocol relies on commutative cryptography, a thus some background for this is provided first.

Commutative Cryptosystems

Many cryptographic applications employ sequential encryption and decryption operations under one or more underlying cryptosystems. The reasons to sequence (cascade) different cryptographic schemes together include, strengthening the resulting ciphertext and achieving additional functionality which is impossible under any given encryption scheme on its own (Shannon, 1949; Weis, 2006). A basic cascadable cryptosystem can consist of a number of encryption stages, where the output from one stage is treated as an input to another. In such a basic cascadable cryptosystem it is necessary to decrypt in the reverse order of encryption operations. However, a special class of sequential cryptosystems - commutative cryptosystems – allows for the decryption of a ciphertext in an arbitrary order. Thus, a ciphertext $c = e_b e_a(m)$ (*c* – ciphertext, *m* – plaintext, *e* – encryption operation under keys *a* and *b*), could be decrypted as either $m = d_b d_a(c)$ or as $m = d_a d_b(c)$. The advantages of such cryptosystems were widely promoted by Shamir (1980) as used in his, Rivest's and Aldman's, now classic, game of *mental poker*, employing the Three-Pass (3Pass) secret exchange protocol.

The most commonly used commutative cryptosystem is based on the Pohling-Hellman (PH), asymmetric private key scheme (1978). This scheme first published in 1978 has never become popular since it is asymmetric, and therefore, slow in comparison to other private key systems. While the PH protocol influenced the design of Rivest-Shamir-Adleman (RSA) public key scheme (1978), the main strength of PH is that it is commutative for keys based on the same prime number and that it allows for comparing the encrypted ciphertexts. Consequently, under PH the two ciphertext $c_{ba} = e_b e_a(m)$ and $c_{ab} = e_a e_b(m)$ hiding the same plaintext *m* are equal (1), while this is not the case with ordinary encryption protocols, that satisfy (2).

$$e_a e_b(m) = e_b e_a(m) \quad (1)$$

$$e_a e_b(m) \neq e_b e_a(m) \quad (2)$$

Thanks to those properties PH can be used in the 3Pass primitive that allows two parties to exchange data without exchange of keys, as well as to perform PEqT that permits private matching of data records.

Three Pass Protocol (3Pass)

The 3Pass protocol, shown in Fig. 1, was intended to allow two parties to share a secret without exchanging any private or public key.

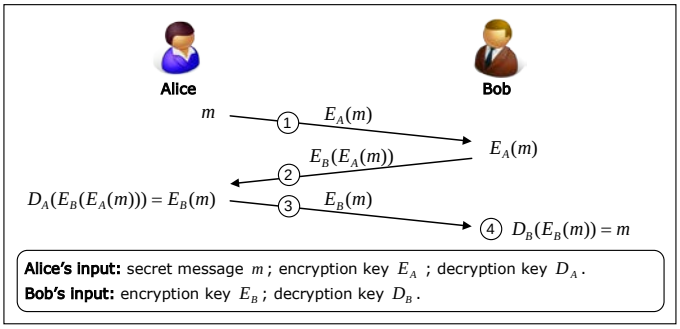

Fig. 1 Three-Pass Secret Exchange Protocol.

The protocol was aimed at providing an alternative to public-key encryption and DH-like key negotiation protocols.

The operation of the protocol can be described using the following physical analogy:

1. Alice places a secret message *m* in a box and locks it with a padlock E_A.
2. The box is sent to Bob, who adds his padlock E_B to the latch, and sends the box back.
3. Alice removes her padlock and passes the box back to Bob.
4. Bob removes his padlock, and this enables him to read the message from inside the box.

There could be more parties, or encryption stages, involved in a 3Pass-like protocol, and this property makes it ideal for locking a plaintext multiple times and then unlocking it in an arbitrary order, as long as the parties are cooperating until the execution of the protocol is completed. Such functionality is required by IDAP as described later in this paper.

Private Equality Test (PEqT)

PEqT protocols can be used to privately verify whether two secret inputs to the protocol are equal or not. Agrawal, Evfimievski and Srikant (2003) proposed one of the most scalable and flexible PEqT protocols for operations on datasets. The scheme is illustrated in Fig. 2 and can be described in the following steps:

1. Alice encrypts her input and sends it to Bob.
2. Bob encrypts the ciphertext received from Alice and sends it back.
3. Bob encrypts his secret input and sends it to Alice.
4. Alice encrypts the ciphertext containing Bob's input.
5. Alice compares the two resulting ciphertexts, if they are equal then her and Bob's inputs are equal.
6. Alice may inform Bob about the result.

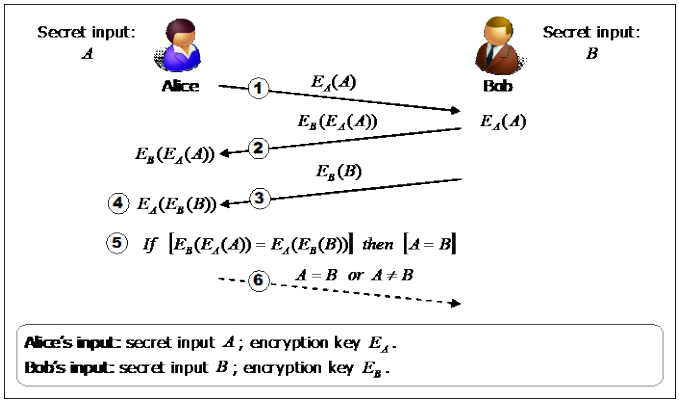

Fig. 2 Private Equality Test.

This protocol allows two parties to compare their secret inputs.

The following section describes a scheme that extends both the PEqT and 3Pass primitives to form the PE protocol that is the blueprint for IDAP.

Private Equijoin Protocol

PE protocol can enable two parties, the *chooser* and the *sender*, to privately compare their sets of unique values V_C and V_S, and allows the chooser to retrieve some extra information $ext(v)$ about records V_S, that match records V_C on a given parameter. The PE protocol involves the following steps:

1. Both parties apply hash function *h* to the elements in their sets, so that $X_C = h(V_C)$ and $X_S = h(V_S)$. *Chooser* picks a secret PH key E_C at random, and *sender* picks two PH keys E_S and E'_S, all from the same group Z_p^*.
2. *Chooser* encrypts entries in the set: $Y_C = E_C(X_C) = E_C(h(V_C))$.
3. *Chooser* sends to *sender* set Y_C, reordered lexicographically.
4. *Sender* encrypts each entry $y \in Y_C$, received from the *chooser*, with both E_S and E'_S and for each returns 3-tuple $\langle y, E_S(y), E'_S(y) \rangle$.
5. For each $h(v) \in X_S$, *sender* does the following:
   (a) Encrypts $h(v)$ with E_S for use in equality test.
   (b) Encrypts $h(v)$ with E'_S for use as a key to lock the extra information about *v*, $\kappa(v) = E'_S(h(v))$.
   (c) Encrypts the extra information $ext(v)$:
   $c(v) = K(\kappa(v), ext(v))$
   Where *K* is a symmetric encryption function and $\kappa(v)$ is the key crafted in Stage 5b.
   (d) Forms a pair $\langle E_S(h(v)), c(v) \rangle$. These pairs, containing a private match element and the encrypted extra information about record *v*, are then transferred to *chooser*.
6. *Chooser* removes her encryption E_C from all entries in the 3-tuples received in Step 4 obtaining tuples α, β, and γ such that $\langle \alpha, \beta, \gamma \rangle = \langle h(v), E_S(h(v)), E'_S(h(v)) \rangle$. Thus, α is the

hashed value $v \in V_C$, β is the hashed value v encrypted using E_S, and γ is the hashed value v encrypted using E'_S.

7. *Chooser* sets aside all pairs received in Step 5, whose first entry is equal to one of the β tuples obtained in Step 6. Then using the γ tuples as symmetric keys it decrypts the extra information contained in the second entry in the pair $\langle E_S(h(v)), c(v) \rangle$.

The above protocol can perform the basic functions required for the purpose of investigative data acquisition. Its use in investigative scenarios is described in the following section.

V. IDAP VS. PRIVATE EQUIJOIN

This section evaluates use of the PE protocol as basis for IDAP. The operations required during investigative data acquisition from a third party in general consist of:

1. Identification of the type of the information that is required. These could be h parameters, that contain answers to investigator's questions, referred to as return parameters $rp_{1\text{-}k}$, e.g. DOB, address, location of a card payment, or numbers called by a given subscriber.
2. Specification of any circumstantial request constrains, or l different input parameters, $ip_{1\text{-}l}$, with values $ip\_val_{1\text{-}l}$, e.g. time frame of the transactions being requested.

3. Specification of the relevant data subject e.g. by identifying the individual whose data is to be retrieved, or by providing the mobile phone number of the suspect, etc. This parameter is referred to as the record of the interest, *ri* with value *ri_val*.
4. Retrieval of the relevant records

Then, if we refer to the dataset as the *source*, the request for investigative data could be mapped into the following SQL query:

$$\begin{aligned} &SELECT\ rp_1, rp_2, \ldots, rp_h \\ &FROM\ source \\ &WHERE\ \mathbf{ri=ri\_val}\ AND\ ip_1=ip\_val_1\ AND\ ip_2=ip\_val_2\ AND \ldots AND\ ip_l= ip\_val_l \end{aligned} \quad (2)$$

In most cases the names of the return parameters, as well as the names of the input parameters, and values of these input parameters can be openly communicated. But the value of the interesting record, *ri_val* is used to uniquely identify the suspect and must be hidden. This can be achieved by running a database query for the return parameters of all the records that satisfy the conditions defined by the input parameters and then collecting the interesting record from the sender using a PE protocol. Consequently, the query that is actually run on the sender's database can be rewritten to:

$$\begin{aligned} &SELECT\ \mathbf{ri}, rp_1, rp_2, \ldots, rp_h \\ &FROM\ source \\ &WHERE\ ip_1=ip\_val_1\ AND\ ip_2=ip\_val_2\ AND \ldots AND\ ip_l= ip\_val_l \end{aligned} \quad (3)$$

The results of such query (3) would be an input to a PE that would enable the chooser to privately select only the record of interest that match given *ri_val*.

PE's Performance

Section III discussed different types of protocols available that could enable the *chooser* to download a record from the *sender's* database maintaining the secrecy of the record selected. It also mentioned that most available protocols could not achieve IDAP on them own, and combination of two or more protocols is required. Such combination typically results in high computational and communicational complexity, because each protocol usually requires its own preparation phase. The PE protocol described in Section IV is capable of both private matching and performing SPIR, and has a low overhead. Table I defines the computational complexity of the protocol.

TABLE I
Computational Complexity of the PE protocol

	Symmetric Crypto.	Asymmetric Crypto.	
	crypto. operation	key generation	crypto. operation
Step 1	-	$O(3)$	-
Step 2	-	-	$O(m)$
Step 4	-	-	$O(2m)$
Step 5	$O(n)$	-	$O(2n)$
Step 6	-	-	$O(2m)$
Step 7	$O(m)$	-	-
Total Complexity	$O(n+m)$	$O(3)$	$O(5m+2n)$
Cost (ms/operation)	0.33	7	30

The complexity of each of the steps in the proposed initial solution. Where n *is the number of the data rows in the source, and* m *is the number of interesting records. Cost is the measured average time in* ms *to perform given cryptographic operation from managed C# .NET code.*

In practice this particular solution based on the PH cipher and implemented in C# .NET can process thousand records a minute, on average. The following section discusses the performance in context of investigation, and discusses issues that could possibly limit the usability of the solution presented.

Advantages of PE in data acquisition process

As per the requirements the PE protocol allows for acquiring more than one interesting record at the time, and adding more records to the enquiry increases the processing time by a negligible value (~151ms) per each extra interesting record in an enquiry. Use of the PE would also satisfy the condition that the dataholder remains in full control of data, and decides which data can be disclosed. In the EP protocol each record is processed separately and there are no chances of the records being mixed up by the privacy-preserving process. Thanks to this fact unnecessary data

of non-suspects could be discarded on reception by the authorities and still the encrypted interesting records received would form valid evidence for use in a court of law. The costs involved in building and deploying PE based IDAP are anticipated to be low since it is a software system and the architecture would is based on a protocol that is in the public domain. By automating the acquisition process the costs to the CSP's would also be lowered, since the human operators would only have to allow or disallow a given request.

Currently, any time the personal data in an electronic system is accessed for any other purpose than maintenance it is considered as being *processed* (Spiers, 2009). Under the DPA *processing* of the personal data can only be done with expressed or implied consent of the *data subject*, unless it is required to satisfy legal requirements of the data controller. Therefore, if the use of the data acquisition platform would be compulsory for all data controllers that would most likely render this method of data acquisition legal. Also the government and many public institutions use Internet and physical media to transfer data between different departments and locations. On occasions such data is lost, e.g CDs are left on a train such as in the cases described in (BBC News, 2008)[r164]. This is never considered a problem by the public eye if the data is encrypted. In the proposed system some unsolicited records, i.e. the results to the query (3), are transferred from the sender to the chooser in a way that no other party but the *sender* can access, these with the exception of the investigators that can access the requested interesting records.

Disadvantages of PE in data acquisition process

The processing time required for the protocol to run is the main drawback of the PE protocol. If there is a thousand records in the database it only takes approx. one minute for the complete run of the protocol, however, the processing time is linear to the number of records in a dataset and data acquisition from a database with five million records would take three and a half days to run on an ordinary PC. During an urgent enquiry, when life of and individual is in danger, or an individual can seriously endanger others, police can currently get access to relevant location data from a mobile network operator in less than half an hour. Such a result could not be expected of PE if the database has more than thirty thousand records. Additionally, even if the data requested is relatively small in size, e.g. 100kB per record, then the results from a database of five million records would be more thank 500MB of data that would need to be transferred over the Internet. Clearly, there is a requirement for the PE to run on a subset of the sender's database rather than the whole database or another solution would need to be chosen. The first approach is described later in Section VI.

Another issue is that the PE based system allows for secure matching on a single value per record, e.g. IP address, name or a credit card number. In some scenarios it may be required to request records based on a number of secret input parameters. Consider scenario where Police has a profile of a suspect (e.g. sex, age, and ethnic origin) and would like to find individuals fitting this profile working in organizations in a neighbourhood to the crime scene, but revealing the profile to these organizations may harm the investigation and the individuals matching the profile. Currently the police would often have to delay their enquiries in order to protect the investigation, and the innocent individuals fitting their profile. For example if the case being investigated had a public tension around it, and the suspect's profile matched individuals in a local minority, an openly conducted enquiry could have serious consequences to the members of

this minority. IDAP should be able to assist the police in such a scenario, thus some modifications that need to be introduced to the protocol are proposed in the next section.

Finally, the lawyer's opinion about legality of the protocol that transfers large chunks of non-suspect data to the investigators is divided. Some consider this solution as acceptable as long as it can be proven that the public authorities are unable to decrypt any unsolicited data, while others suggest that anything that creates a privacy risk, however remote, requires the consent of the parties involved. The case law supports both of these opinions, thus, until such case is brought in front of court the matter cannot be clearly answered. Clearly there is a need for a process or a mechanism that would further eliminate the risk to the data records of non-suspect data-subjects. This is presented in Section VI together with other modifications required to PE in order to create acceptable IDAP solution.

VI. PROPOSED MODIFICATIONS

The previous section has listed the drawbacks of using the PE in the pursuit of IDAP. Here these drawbacks are addressed by three different correcting measures that modify the PE protocol for the specific purpose of investigative data acquisition.

Improvement 1 – Lowering Processing Time

Section V recommended minimising the processing time required for each run of the protocol in large databases, such as those belonging to ISPs and mobile telephony providers. Theoretically, in order to maintain privacy of the suspect, the *chooser* needs to request from the *sender* to process all the records in the database. Only this way no information about interesting record is revealed and the correctness of this scheme can be proven under the requirements of the multiparty computation (Asonov & Freytag, 2003). In its current form the system would not be capable of processing any urgent requests due to the processing time required, and this would be a major drawback. The mitigation for this could be to limit the numbers of records that needs to be processed and then sent by the *sender* per enquiry. Privacy of the alleged suspect should be protected, but if the probability of the *sender* guessing the ID of the interesting record is for example 1:1000 and not 1:*n*, and the dataholder has no other information that could help infer any knowledge as to the identity of the suspect, then this research argues that the privacy of the suspect and the investigation is maintained. The police sources suggest than on occasion during traditional, i.e. face-to-face, information gathering the officers would use a concept of *diffusion* - hiding the suspect's identity by asking open-ended questions about a larger group of individuals rather than about a single person. This is a widely accepted technique, however, in the digitalised environment it is impossible to build a system that would maintain privacy while answering such general questions. Consequently, any attempts of investigators to *cast their net wide* during electronic investigations are prohibited and treated as *fishing* for evidence. Taking in consideration that using IDAP the investigators will not get more data than required for the inquiry, limiting the set of records that is processed per enquiry should be acceptable.

The problem is to decide on the technique of narrowing down the scope in a way that ensures interesting records are among the results returned. If the list of the record identifiers is public, such as the list of the Internet Protocol (IP) addresses or telephone numbers served by a given network operator, then the *chooser* could simply selected records to be processed at random from

such directory. However, in case such list is not publicly available it would be possible to split PE protocol back into separate parts: PEqT; and OT, and an additional off-line preparation phase. This way the initial off-line phase could be run against the whole database but the information retrieval would be performed against a smaller set of records. If as a number of records requested per each interesting record is defined as the diluting factor - *o* the protocol IDAP would be defined as follows:

Phase A - Preparation

1. *Sender* applies hash function *h* to the elements in the input set V_S, so that $X_S = h(V_S)$.
2. *Sender* picks a encryption PH key E_S at random from a group Z_p^*, where *p* is a strong prime.
3. *Sender* encrypts each $h(v) \in X_S$ with the key E_S, the result is a list of encrypted identities $Y_S = E_S(X_S) = E_S(h(V_S))$

If more record needs to be added to the set these can be processes using steps 1 and 3, and then added to the list.

Phase B - PEqT

1. Following a request for data, *sender* provides *chooser* with a complete list of encrypted identities prepared during Phase A, reordered lexicographically.
2. *Chooser* applies hash function *h* to the elements in set containing the identities of the interesting records, so that $X_C = h(V_C)$.
3. *Chooser* picks a commutative cryptography key pair, encryption key E_C and decryption key D_C, at random from the same group Z_p^* that was used by *sender* in the Phase A.
4. *Chooser* encrypts entries in the set X_C, so that: $Y_C = E_C(X_C) = E_C(h(V_C))$.
5. *Chooser* sends to *sender* set Y_C, reordered lexicographically.
6. *Sender* encrypts with key E_S each entry $y \in Y_C$ received from *chooser*.
7. *Sender* returns set of pairs $\langle y, E_S(y) \rangle$ to *chooser*.
8. *Chooser* decrypts each entry in $E_S(Y_C)$, obtaining $E_S(X_C) = D_C E_S(E_C(X_C)) = D_C E_S(Y_C)$.
9. *Chooser* compares each entry in $E_S(X_C)$ to the entries of Y_S received in the Step B1 (Step 1 of Phase B). This way the interesting records can be identified.

Phase C - OT

1. After identifying the interesting records in Y_S the *chooser* selects at random $o-1$ other unique records from Y_S for each interesting record in V_C. These are the diluting records, that together with the records of interest form a shortlist for the enquiry . If the number of interesting records multiplied by *o* is greater than *n*, the size of the dataset V_S, then the complete Y_S is shortlisted.
2. Send the shortlist to *sender*.
3. *Sender* picks an encryption PH key E'_S at random from the group Z_p^*.
4. *Sender* identifies entries $h(v)$ from X_S that have been shortlisted and processes each shortlisted record in the following way:

(a) Encrypts $h(v)$ with E'_S to form the key used to lock the extra information about *v*, i.e. $ext(v)$, $\kappa(v) = E'_S(h(v))$.

(b) Encrypts the extra information using a symmetric encryption function *K* and the key $\kappa(v)$ crafted in the previous step:
$c(v) = K(\kappa(v), ext(v))$

(c) Forms a pair $\langle E_S(h(v)), c(v) \rangle$.

5. The pairs formed in C4(c), containing a private match element and the encrypted extra information about record *v*, are then transferred to *chooser*.
6. *Sender* encrypts each entry $y \in Y_C$, received from *chooser* in Step B5, with key E'_S to form set of pairs $\langle y, E'_S(y) \rangle$
7. Pairs $\langle y, E'_S(y) \rangle$ are then transferred to *chooser*.
8. *Chooser* removes the encryption E_C from all entries in the 2-tuples received in Step C7 obtaining tuples α, β such that $\langle \alpha, \beta \rangle = \langle h(v), E'_S(h(v)) \rangle$. Thus, α is the hashed value $v \in V_C$, and β is the hashed value *v* encrypted using E'_S.
9. *Chooser* sets aside all pairs received in Step C5, whose first entry is equal to one of the first entry of any two-tuples obtained in Step B9. Then uses the appropriate β tuple associated with a given interesting record as a symmetric key to decrypt the extra information contained in the second entry in the pair received in C5. This is performed for all the matching entries.

In this improved protocol the initial processing depends on the size of the dataset - *n*, but it needs to be performed only once in a given period of time, e.g. once par month, or per year. The remaining operations are less processing savvy as illustrated in Table II.

TABLE II
Computational Complexity of Improvement 1

		Symmetric Crypto.	Asymmetric Crypto	
		crypto. operation	key generation	crypto. operation
Phase A (run periodically)	Step 1	-	-	-
	Step 2	-	$O(1)$	-
	Step 3	-	-	$O(n)$
Phase B (run per enquiry)	Step 3	-	$O(1)$	-
	Step 4	-	-	$O(m)$
	Step 6	-	-	$O(m)$
	Step 8		-	$O(m)$
Phase C (run per enquiry)	Step 3	-	$O(1)$	-
	Step 4(a)	-	-	$O(m \times o)$
	Step 4(b)	$O(m \times o)$	-	-
	Step 6	-	-	$O(m)$
	Step 8	-	-	$O(m)$
	Step 9	$O(m)$	-	-
Total Complexity for k enquiries, where $n < m \times o$		$O(km(o+1))$	$O(2k+1)$	$O(km(o+5)+n)$
Cost (ms/operation)		0.33	7	30

The complexity of each of the steps in the proposed improved solution. Where n *is the number of the data rows in the* source, m *is the number of interesting records. Also the diluting factor* o, *as well as the number of the protocol runs* k *affect the processing time required by the protocol. Cost is the measured average time in* ms *to perform given cryptographic operation from managed C# .NET code.*

Fig. 3 illustrates the processes involved in this improved version of acquisition protocol. It is worth noting that there is only five communication rounds required in this protocol. This is two rounds more than in the original PE protocol, still, most of efficient SPIR protocols require considerably more rounds. This method provides significant improvements to the processing time required for enquiries if total number of records in the *sender's* database is higher than $o \times m$, i.e. higher then the number of interesting records *m* multiplied by the diluting factor *o*. This is illustrated in Fig. 4. Furthermore, the true strength of this version of the protocol is seen when multiple enquiries are run of the same database using a single encrypted catalogue of the records, compiled by the *sender* in Phase 1 (shown in Fig. 5).

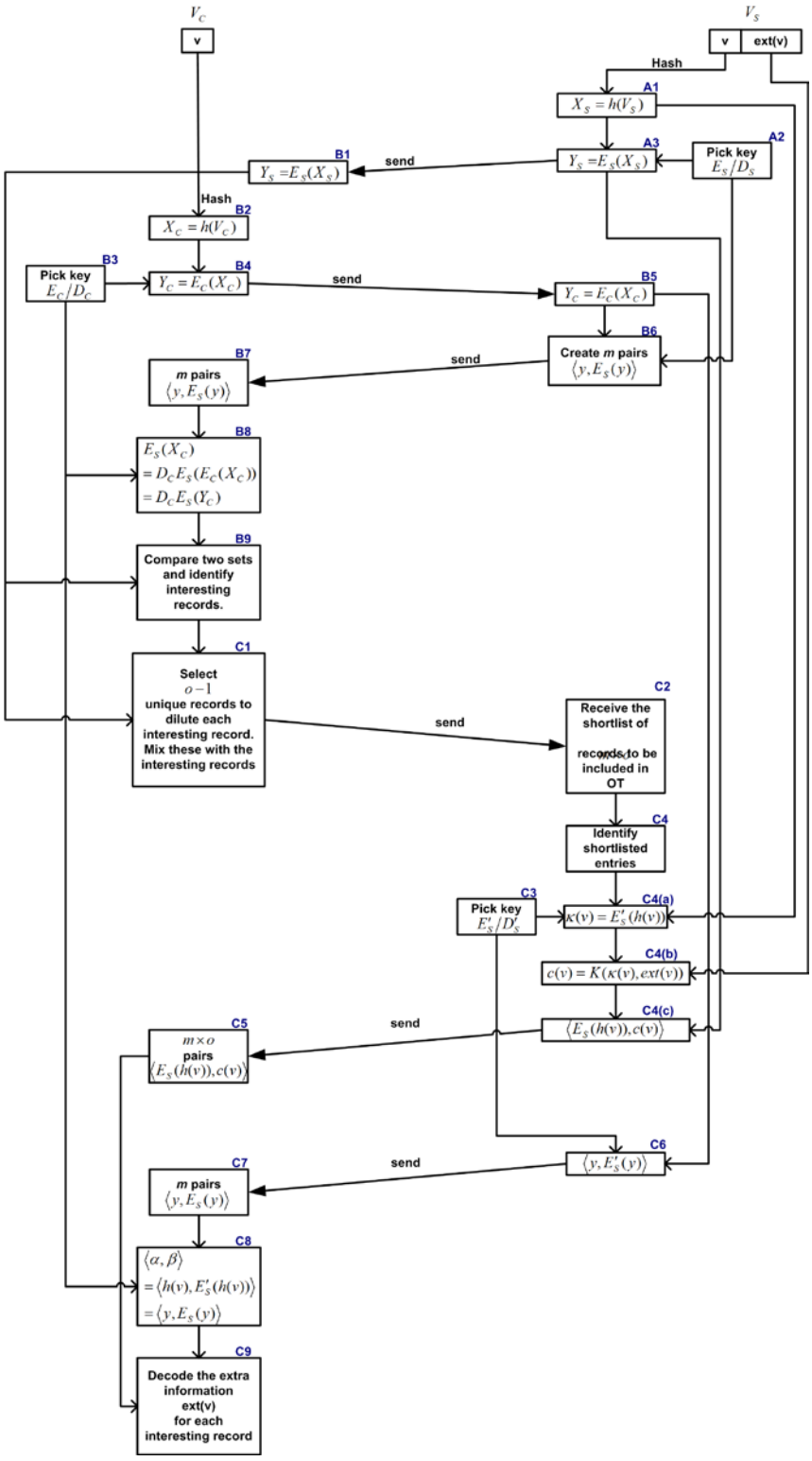

Fig. 3 IDAP Process Flow

Graphical representation of the improved IDAP

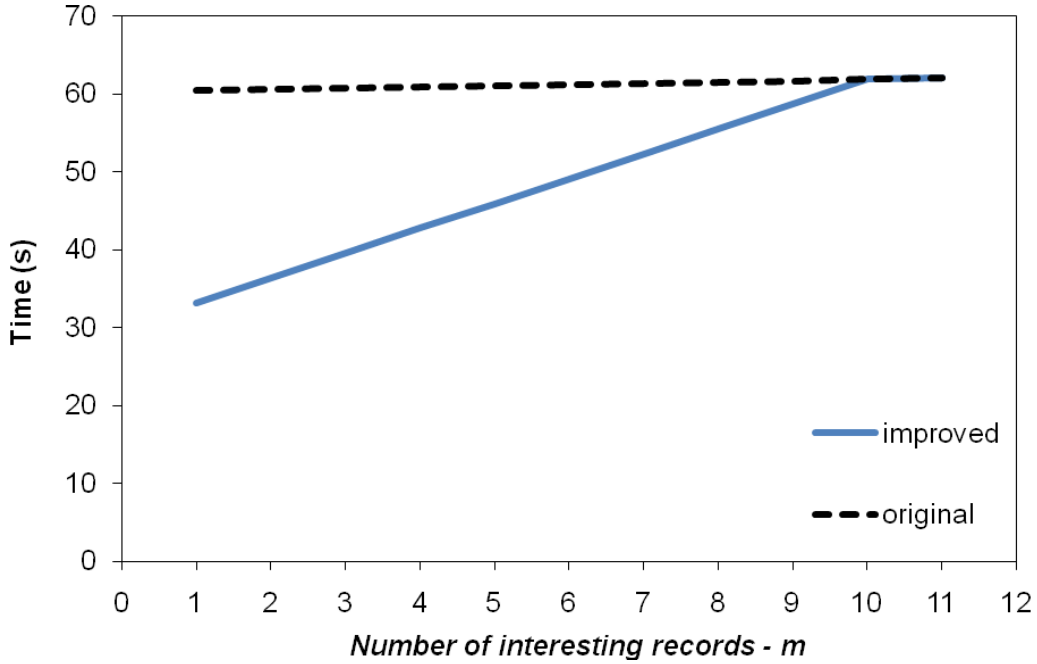

Fig. 4 Processing time per enquiry depending on the number of interesting records

This proposed modification of the protocol improves significantly the processing time required for the protocol to run for the cases where the product of the number of the interesting records m *and diluting factor* o *is smaller that the number of the records in the database* n.

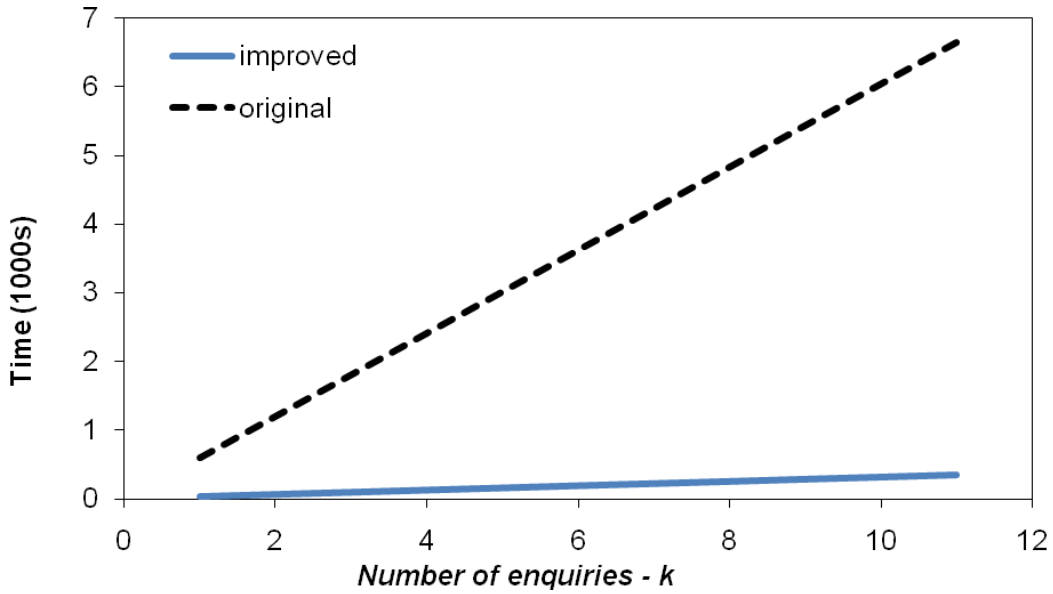

Fig. 5 Processing time depending on the number of enquires

This proposed modification improves significantly the processing time required for the protocol to run for the cases where more than one enquiry is run against the same database.

Improvement 2 – Allow multiple selection criteria

The PE protocol can be used to privately retrieve data if the data is identified by a single parameter, such as ID number, credit card number, IP address, etc. However, this is not always the case. Consequently, if IDAP is used to find a suspect based on circumstantial knowledge, or a suspect's profile the PE protocol would need to be modified. Query (4) shows the way the request (3) would be modified for such enquiry, here $sip_{1\text{-}j}$ stand for j secret input parameters:

SELECT $\mathbf{sip_1, sip_2, \ldots, sip_j}$, $rp_1, rp_2, \ldots, rp_h$
FROM source (4)
WHERE $ip_1=ip\_val_1$ AND $ip_2=ip\_val_2$ AND … AND $ip_l=ip\_val_l$

A computationally expensive solution to this problem has been published by Kwecka, Buchanan, and Spiers (2010). The authors suggest that the symmetric encryption should be used to lock the return parameters and the symmetric keys should be secured with relevant commutative encryption keys that are unique to each value of the secret input parameter returned for the given row. Despite being computationally expensive this solution has a unique benefit of allowing semi-fuzzy matching of the results if the underlying commutative protocol is ElGamal-based.

In this work a simplified approach is proposed. Since, the query (4) replaces the *ri* parameter with *j* different *sip* parameters then the list of these *j* parameters could be used as a complex *ri* in the improved IDAP protocol. Thus, in Steps B2 and A1 a list of all values of given *sip* parameters would be hashed together to form records in sets V_C and V_S. This way the security of the protocol nor its complexity is affected by this improvement.

Improvement 3 – Reassuring the Public

Sad quis custodiet ipsos custodies?
But who will watch the watchers? (Juvenal, Satires VI, 347)

It is likely that providing government agencies with encrypted records of innocent, non-suspect individuals would worry the general public. This is despite the data being encrypted in the way that would render the records unusable to the authorities i.e. secure against attacks in polynomial time. However, the public may worry that the government organisations have enough computing power to break the encryption used in IDAP. There are few actions that may reassure the public that the data is safe. First, if the technique for minimising the processing time (Improvement 1) is employed the chances that investigators will retrieve encrypted records of a particular individual that is not a suspect are small in large datasets. Thus, for a dataset with *n* records, during investigation with *m* interesting records and the diluting factor *o* the probability of this event *A* can be defined as (5)

$$P(A)=\frac{(o-1)\times m}{n-m} \quad (5)$$

Consequently for investigation with five interesting records, with diluting factor of a thousand and dataset consisting a million records, the probability of this event occurring during a single run of the protocol would be less than 0.5%. Since, the runs of the protocol are independent this probability would stay the same. This also means that the investigators would need to first break the encryption key used by the *sender* to hide identities (Phase A), before they could attempt to obtain the data about a specific individual that is not a suspect, otherwise the probability of the encrypted data being provided to them would be small. Additionally, if the identity of a data subject is never encrypted under the same key as the data records then investigators would need to successfully brute force two separate keys in order to make use of the retrieved encrypted records. Otherwise the information would be unintelligible.

The merits of the above discussion could certainly improve the perception of the system. Still most security professionals trust into a security process more than they trust in encryption. The solution proposed in this research in order to reassure the public is to introduce a semi-trusted party into the protocol. This party would be a *proxy* between the investigators and the dataholder. The following modifications to the IDAP are proposed:

1. All communication between *chooser* and *sender* goes through *proxy*.
2. Chooser provides *proxy* with the identifiers of the interesting records encrypted by *sender*, $E_S(h(v))$. This is done over a secure channel or with use of a 3Pass protocol once the parties are authenticated.
3. At the stage where data is transferred from *sender* in Step C4, *proxy* filters the response and discards the records that were not specified by *chooser's* request, i.e. the records other than the ones identified in Step 2.

The semi-trusted party should have no interest in finding out the object of the investigation or the content of the data records returned by the dataholder, for this reason it is suggested that the role of this party should be conducted by Information Commissioner's Office (ICO)[i] or its equivalent in other countries. The party that is chosen must not cooperate with the *sender* or the protocol will be broken, since simple matching exercise would reveal the identities of the suspects. A key concept is that the *proxy* has no incentives to find out the detail of the investigation, thus it is not going to purchase expensive cutting edge decryption technology to decode the data, nor it is going to cooperate with the *sender* in order to establish the identity of the suspect. On the other hand, if the need arises to verify the *chooser's* requests in front of court of law, the *proxy* and the *sender* could work together to establish the identities of the records requested by the *chooser*.

The initial design of IDAP from Section IV has shifted the balance of the privacy protection from innocent individuals towards the suspect and the secrecy of investigation. Introduction of the semi-trusted third party into this protocol restores the natural order, where the rights of the innocent are put ahead of the secrecy of the investigation. This is likely to benefit the general public's perception of IDAP.

VIII. CONCLUSION

This paper presented a platform for investigative data acquisition that preserves the privacy of the suspects and secrecy of the investigations. After a careful analysis of the related issues and research of available privacy-preserving primitives IDAP has been defined. The platform is build on PE protocol, a SPIR protocol based on commutative cryptography, that allows retrieval of extra information about the records that are common between two sets. The features of this protocol closely match those required of information retrieval platform, so only some improvements were required. These were documented in this paper.

In this research a view that in certain circumstances hiding the object of the PIR protocol by running the data retrieval protocol against only a subset of the dataset provides sufficient privacy protection is presented. This is certainly the case in the investigative data acquisition process. The number of records that is collected per every interesting record is specified by the *dilution factor o* introduced by this research. Since, this factor can be dynamically changed before each protocol run, the investigators can decide the appropriate level of protection for the given investigation, the data subject and the data controller. The protocol operates by creating a single encrypted table of identities held in the third party's database and allowing the investigators to privately match their suspects against this table. Once the investigators know the encrypted ID of the suspect a number of records is selected at random to make up a request of size *o*. Consequently the data controller can then facilitate private data retrieval operating on a small subset of the database. This way the processing time is significantly reduced and requests from large databases are feasible. Such technique could be potentially risky if the same enquiry is made against few different data controllers, since the intersection of the requested results could help the cooperating controllers to identify the suspect. However, according to the Police it is not likely that data controllers will cooperate in such matters, especially if such cooperation would be forbidden by the letter of law. In the cases that the data is being retrieved from large databases that require use of the *dilution* technique during data retrieval process, the interesting records are usually identified by a mobile phone number or an IP address. Phone numbers and IP addresses are unique to the operators and their assignment can be obtained from the call and network routing tables. This way in most cases the investigators only ask a single operator for information about a given identity. This fact makes most investigations equivalent to a single database PIR and *dilution* can be applied, with no adverse affect on the privacy of the data-subjects.

Quick solution to performing private database searches against secret selection criteria is also provided in this research. The investigators can simply create a list of values for every secret input parameter, and then this list is used in the same fashion an identifier of the interesting record would be used. This technique works, as long as the dataholder prepares the response in the same way. Consequently, neither the complexity of the protocol nor its security properties are altered in providing this additional functionality to the acquisition protocol.

Finally, this paper addresses concerns of the general public in employing encryption based PETs to handle sensitive data. People generally trust the security process more that they trust encryption. For this reason a semi-trusted third party is added to the protocol to act as a proxy. The entire communication between the investigators and the dataholders is done via this proxy and the key objective of the proxy is to filter out the records that were not requested by the investigators in their request. This protocol is secure as long as the proxy is trusted not to

cooperate with the dataholder. For this reason a party whose main concern is privacy of the individuals should held the function of the proxy. In UK Information Commissioner's Office could handle such a function. This approach ensures that the balance between the privacy of the alleged suspect and the privacy of the innocent individuals stays the same after IDAP is introduced. Such move is likely to improve the public's perception of the platform.

REFERENCES

Agrawal, R., Evfimievski, A., & Srikant, R. (2003). *Information sharing across private databases.* Paper presented at the Proceedings of the 2003 ACM SIGMOD international conference on Management of data, San Diego, California.

Aiello, B., Ishai, Y., & Reingold, O. (2001). Priced Oblivious Transfer: How to Sell Digital Goods. In B. Pfitzmann (Ed.), *Advances in Cryptology — EUROCRYPT 2001* (Vol. 2045, pp. 119-135): Springer-Verlag.

Asonov, D., & Freytag, J.-C. (2003). Almost Optimal Private Information Retrieval. In *Privacy Enhancing Technologies* (pp. 239-243).

Association of Chief Police Officers. (2003). Good Practice Guide for Computer based Electronic Evidence (version 3). Retrieved 19 December 2006, 2006, from http://www.acpo.police.uk/asp/policies/Data/gpg_computer_based_evidence_v3.pdf

Balz, D., & Deane, C. (2006, 11/1/2006). Differing Views on Terrorism. *The Washington Post*, p. A04.

Bao, F., & Deng, R. (2001). Privacy Protection for Transactions of Digital Goods. In *Information and Communications Security* (pp. 202-213).

BBC News. (2008, 15 June 2008). More secret files found on train. Retrieved 15 June 2008, from http://news.bbc.co.uk/1/hi/uk/7455084.stm

Cachin, C. (1999). *Efficient private bidding and auctions with an oblivious third party.* Paper presented at the 6th ACM conference on Computer and communications security - CCS '99, Singapore.

Du, W., & Atallah, M. J. (2001). *Privacy-Preserving Cooperative Scientific Computations.* Paper presented at the Proceedings of the 14th IEEE workshop on Computer Security Foundations.

EUROPEAN PARLIAMENT. (1995). European Data Protection Directive 95/46/EC. *Official Journal of the European Union, L*(281), 31-50.

Frikken, K. B., & Atallah, M. J. (2003). *Privacy preserving electronic surveillance.* Paper presented at the Proceedings of the 2003 ACM workshop on Privacy in the electronic society, Washington, DC.

Goldwasser, S., & Lindell, Y. (2002). *Secure Computation without Agreement.* Paper presented at the Proceedings of the 16th International Conference on Distributed Computing.

Home Office. (2007). *Acquisition and Disclosure of Communications Data - Code of Practice*. Retrieved. from http://security.homeoffice.gov.uk/ripa/publication-search/ripa-cop/acquisition-disclosure-cop.pdf?view=Binary.

Home Office. (2009). Protecting the Public in a Changing Communications Environment [Electronic Version]. Retrieved 20 April 2009 from http://www.homeoffice.gov.uk/documents/cons-2009-communications-data?view=Binary.

Information Commissioner. (2007). Data Protection Guidance Note: Privacy enhancing technologies (2.0 ed.). London: Information Commissioner's Office.

Kwecka, Z., Buchanan, W., Spiers, D., & Saliou, L. (2008, 30 June – 1 July). *Validation of 1-N OT Algorithms in Privacy-Preserving Investigations.* Paper presented at the 7th European Conference on Information Warfare and Security, University of Plymouth.

Kwecka, Z., Buchanan, W. J., & Spiers, D. (2010). *Privacy-Preserving Data Acquisition Protocol.* Paper presented at the Sibircon, Irkutsk.

Ostrovsky, R., & William E. Skeith III. (2007). A Survey of Single-Database PIR: Techniques and Applications. In O. Tatsuaki & W. Xiaoyun (Eds.), *Public Key Cryptography* (Vol. 4450, pp. 393-411). Berlin: Springer

Palmer, G. (2001a). A road map for digital forensic research.

Palmer, G. (2001b). *A road map for digital forensic research* (Technical Report). Utica, NY: DFRWS.

Pohlig, S., & Hellman, M. (1978). An improved algorithm for computing logarithms overGF(p)and its cryptographic significance. *Transactions on Information Theory, IEEE, 24*(1), 106-110.

Rasmussen Reports. (2008). 51% Say Security More Important than Privacy. Retrieved 01/09/2009, from http://www.rasmussenreports.com/public_content/politics/current_events/general_current_events/51_say_security_more_important_than_privacy

Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM, 21*(2), 120-126.

Schneier, B. (1995). *Applied Cryptography: Protocols, Algorithms, and Source Code in C*: John Wiley & Sons, Inc.

Shamir, A. (1980). *On the Power of Commutativity in Cryptography.* Paper presented at the Proceedings of the 7th Colloquium on Automata, Languages and Programming.

Shannon, C. (1949). Communication theory of secrecy systems. *Bell System Technical Journal, 28*.

Spiers, D. (2009). *Intellectual Property Law Essentials*. Dundee: Dundee University Press.

Swire, P., & Steinfeld, L. (2002). *Security and privacy after September 11: the health care example.* Paper presented at the Proceedings of the 12th annual conference on Computers, freedom and privacy, San Francisco, California.

Weis, S. A. (2006). *New Foundations for Efficient Authentication, Commutative Cryptography, and Private Disjointness Testing.* Unpublished PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Young, B. C., Kathleen, E. C., Joshua, S. K., & Meredith, M. S. (2006). Challenges Associated with Privacy in Health Care Industry: Implementation of HIPAA and the Security Rules. *J. Med. Syst., 30*(1), 57-64.

[i] ICO in the United Kingdom, is a non-departmental public body which reports directly to Parliament and is sponsored by the Ministry of Justice. It is the independent regulatory office dealing with the Data Protection Act 1998 and the Privacy and Electronic Communications (EC Directive) Regulations 2003 across the UK.