# Returning Control of Data to Users with a Personal Information Crunch - A Position Paper

Mark A. Will*, Jeffery Garae†, Yu Shyang Tan†, Craig Scoon† and Ryan K L Ko*
Cyber Security Lab
The University of Waikato
Hamilton, New Zealand
Email: *{mark.will, ryan.ko}@waikato.ac.nz, †{jg147, yst1, cs145}@students.waikato.ac.nz

***Abstract*—With the data universe expanding to uncontrollable limits, we are losing control of our personal information. From online purchases to movie streaming, we are giving vendors more and more information, such that our privacy is at stake. Hackers and third-parties can gain access to this information, putting us at risk to a number of attacks. The current model where every online vendor has personal information, such as name, addresses and date of birth should be reconsidered. A user needs to have full or at least more control over their personal data, and who has access to it. This paper presents alternatives to vendors having all of a users personal information and raises many concerns about the current state of play. A simple model is proposed where personal information is stored on the users mobile device, and requested by vendors when needed. Information can then be given in either a private or trusted manor, and encrypted responses can be cached by a relay service. Vendors should only use the data inflight, and never store personal information. This provides the user with data provenance and access control, while providing the vendor with accountability and enhanced security.**

***Index Terms*—Cloud Computing; Personal Information; Privacy-Preserving; Provenance; Returning Control; Security;**

## I. Introduction

A serious deficiency we are observing with the current state of the cyber security industry is *the inability for data owners to control their data* [1]. In the event of a data breech, such as the recent Ashley Madison hack [2] or the Yahoo hack of a billion accounts [3], personal information is being leaked, often without the user being notified. This leads to issues such as identity theft or credit card fraud. Personal information is not just at risk from malicious attackers, but also people employed by these companies. In 2010, an engineer at Google abused his privileged administrator rights to spy on teenagers using the GTalk service [4]. Only after the teenagers parents reported the administrator was Google made aware. The sole reliance on the trust and reputation of a cloud service provider and their employees is neither a strong nor sustainable way forward for the cloud computing industry [1].

The risk of personal information being stolen or misused is not the only issue with the collection of personal information. Information about a user is constantly being sold to advertising companies [5][6]. For example, an online search on a car may results in targeted advertisements for the same brand of car through AdSense. Information is also leaked through other means such as cookies and third-party sites (for example ads within the page) [7][8].

This paper is targeting the issue of returning control of personal information back to the user, while proposing a different mindset to a user needing to share their personal information with vendors and websites. The issue of leakage through cookies etc. has been discussed in detail in other work [7][8]. But solving the problem of lack of control over data has not been addressed [1]. Other motivation for returning control of data to users is the introduction of the General Data Protection Regulation, discussed in Section II.

The paper defines personal information in Section III, limiting the scope to information such as name, addresses and date of birth. The system model is proposed in Section IV, which stops the spread of personal information, by reducing it to a single entity where vendors can request information when needed. How personal information can be protected using the system model, or other means is discussed in Sections V and VI. The current challenges and limitations of protecting users privacy, and related work addressing some issues is given in Sections VII and VIII respectively.

## II. Law Requirements and Limitations

With the Internet's increasing reach, data can be processed across multiple borders. This complicates the legal jurisdiction and control of user data. Each country enacts their own legislation relating to data privacy, and wording will differ in each of these countries. Such legislations dictate a range of requirements, and more importantly, regulations surrounding the treatment of user data (for example access, amending and erasure). There are some generally agreed upon principles across the countries, including that information is collected for a legitimate and lawful purpose, which needs to be outlined to the data subject before it is collected. The data subject also needs to be informed on the details of who is collecting the data. Data needs to be collected directly from the data subject. Often these details are given in unreadable terms and conditions, and must be accepted before use.

### A. Access

There are some general legislative requirements for anyone who provides their personal information. These include access to information to ensure it is accurate, relevant and complete. These may seem like fundamental rights, however they vary

between countries. Some countries will charge the data subject a small fee to access their data, and may take days to retrieve it.

### B. Storage

A big limitation is the storage of data. Legislation is always playing catch up with technology and therefore not much legislation defines how data should be stored. Most countries legislation have a similar section around data storage which says reasonable security measures must be taken to secure the data. The problem is the subjective term *'reasonable'*. This could mean completely different requirements for different organisations depending on the data they are storing. Encryption is not mentioned in legislation but again could be seen as a reasonable security measure.

### C. General Data Protection Regulation

Currently in the European Union (EU) there are numerous directives in place which aim to protect personal data. The EU Data Protection Directive, which is the main document within the EU for data protection, regulates how data can be processed within the EU. In addition there are two other directives which compliment the Data Protection Directive: (1) the 2009 E-Privacy Directive, and (2) the Data Retention Directive.

The General Data Protection Regulation (GDPR) is a new regulation from the EU that will come into force from 25 May 2018, replacing the existing EU Data Protection Directive. The GDPR will help to strengthen and unify data protection for individuals who reside within the EU, by introducing or further defining the following areas [9]: (1) increased territorial scope, (2) tougher sanctions, (3) consent, (4) breach notification, (5) right to access, (6) right to be forgotten, (7) data portability, (8) privacy by design, and (9) data protection officers.

Once in force, the GDPR will be legally binding on all member states of the EU. This will also extend the scope to *all* organisations who may operate within the EU or process data of EU citizens, whether they are headquartered there or not [10][11]. Much discussion has taken place around the effects that the GDPR will have on the data privacy landscape. The general consensus is that the GDPR will have a positive effect. Therefore the GDPR should become the new standard for data privacy. The new principles in the GDPR aim to give back the control of data to users, with the proposed model presenting in Section IV strengthening this control. Then with vendors in the EU needing to abide by the GDPR, removing personal information from their systems into a centralised model will be beneficial to all parties.

## III. Personal Information

### A. Cyber or Digital Identity

This paper defines 'Personal Information' as pieces of information identifying an individual, such as name or date of birth [12]. Furthermore, the state where personal information is entered online, tampered with or involved in any fraudulent related events is referred to as a person's cyber-identity or digital-identity [13]. Pieces of cyber-identity information often can be attributed back to a specific person. Other personal data such as medical history, images or current location are not regarded in the scope of this paper.

The reason and validity of disclosing personal information to vendors is often unclear and seemly unnecessary in many cases. Users surfing the Internet day-to-day are often oblivious to the security and privacy of their personal information being stored online. However cyber identity thefts, selling of data to third-parties, phishing, and social engineering attacks suggest the need to develop solutions to protect personal information [14][15].

Common sites and vendors require the following information before a user can use their service. Note that by requiring accurate information, users creating accounts with fake details (for example Bob entering his first name as Alice) to protect themselves can be regarded as a criminal offence. Where the proposed model in Section IV still links to accurate personal information, even if the vendor sees anonymised data. Often online shops require two more sets of information, which are also shown below.

**Registration:**

- Email Address
- First or Family Name (or both)
- Phone Contact
- Date of Birth (or Age Range)
- Gender

**Payment/Billing:**

- Name on Card
- Card Number
- Security Code
- Card Type
- Expiry Date (mm/yyyy)

**Delivery:**

- Name (Company or Personal)
- Street Address and Suburb
- City and State (or Province)
- Post Code
- Country
- Contact Number

### B. Social Media

This paper will not address in detail the issues of social media leaking personal information, due to the very nature of social media [16]. Given the amount of money surrounding social media today, seeing a change in how personal information is abused may be difficult. However more user-centric security measures are required. Combined with an easier understanding of the information actually being leaked and the dangers of improper security settings. Finally by users having control of their personal information on other sites, could be a catalyst for change in social media as well.

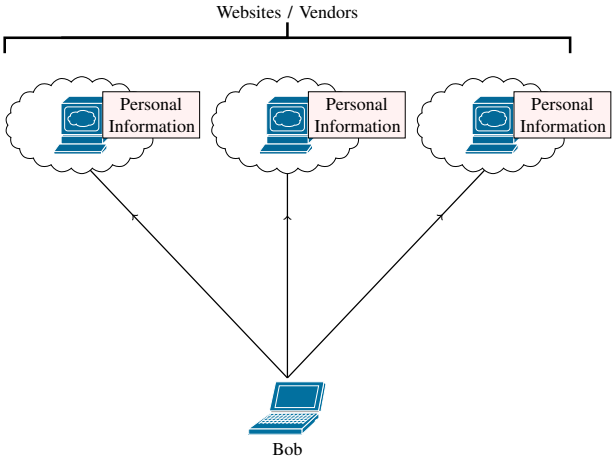


Fig. 1. System model of current state where the spread of personal information is expanding.

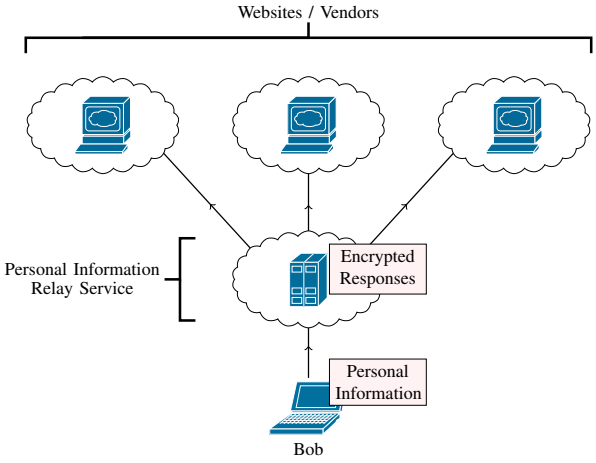


Fig. 2. System model with personal information reduced to a single point, and a relay service storing encrypted responses.

### C. Modern Gold

Personal information has become a commodity for marketers, online advertising companies and other businesses [17]. With the purpose of improving service delivery, companies buy user information and behaviour to better understand preferences and choices when shopping online. For example products and services are tailored towards what are thought of as a users preferences and needs. Whether names and email addresses are in plain-text, anonymised or removed before sale, there still exists the ability to join information together on a single user. One example is hashing an email address, the same hash function without salting will produce the same hash value. Whereas on the black market all stolen personal information will be in plain-text.

### D. Personal Information Management

Personal information privacy is a process of mutual concessions and compromises between two entities, in particular an end user and service provider. Identity management has been proposed as a solution to reducing overheads in managing personal information. For example the UNIQuE framework focuses on security and trust [18], but is not primarily focused on privacy and anonymity, while personal information is visible to the management service. Personal information management provides measures on how to handle, process and manage any information attributing to someone, for example its life cycle. The personal information life cycle (PILC) (collecting, distributing, storing and manipulating) has users being cautious over how their PILC is being handled/processed [19]. Current guidelines, disclosures and regulations have been introduced to control the use of personal information in the entire PILC. However, different business and organisations have different ways of taking care of the PILC process. This allows the potential threat to personal information.

### E. Threats to Personal Information

The very first threat to personal information is the act of invading and exploiting someones privacy with the use of his/her personal information at hand [12]. Social engineering fraud refers to scams which manipulate or lure people into giving out their confidential or personal information, resulting in financial gains for the criminals involved [13]. Similarly, cyber identity theft shows one in twenty five grown-ups in western countries each year are victims to fraudulent activities [20]. An act of crime on personal information or accidental use of personal information either from an insider threat or poor security implementation on systems may lead to identity theft. As a result, financial losses, pharmaceutical record losses and often users become victims to public scrutiny of their online personas and risk physical safety by revealing excessive personal information.

### F. Personal Information and Law Enforcement Approaches

While threats to personal information has been discussed, law enforcement organisations have taken a step ahead by looking at processing, securing and managing personal information [21]. The implementation of data breach disclosure laws, consumer privacy acts, policies and guidelines have a reasonable level of protection for users as part of their mandatory information security policy awareness [12][22]. The use of personal information through law enforcement information security awareness programs have empowered users to be vigilant on how personal information is used on a day-to-day basis [23]. One very important step which was highlighted is the ‘Information Custodian’. At all times, information custodians should manage the information asset life cycle (from creation until destruction) which in most sense minimise the risk of exploiting personal information.

## IV. System Model

Today the mindset of vendors and online services or websites is to require personal information on account creation

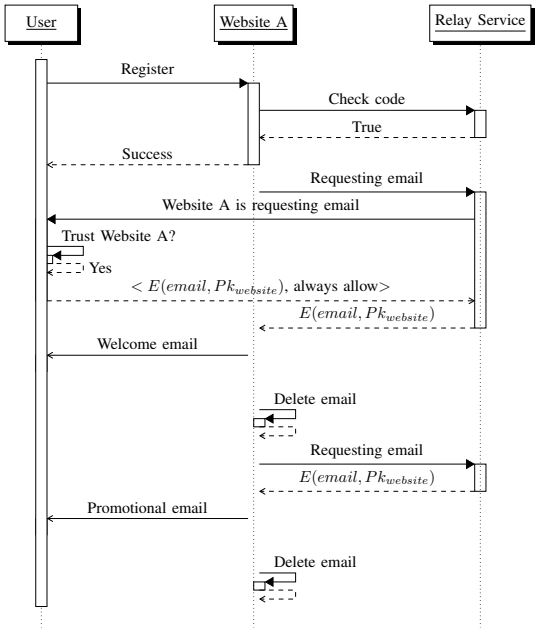


Fig. 3. Trusted Website Signup Flow.

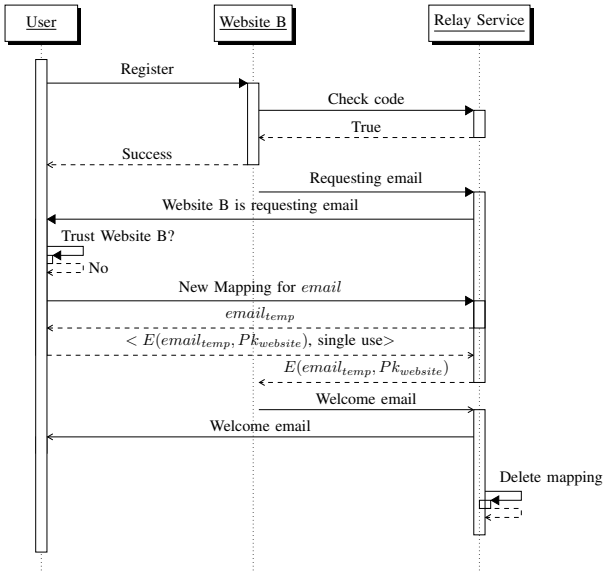


Fig. 4. Untrusted Website Signup Flow.

as discussed in Section III. This results in a users personal information being spread across many locations as shown in Figure 1. Each registration requires the same information, but have no real requirement for it. Without a change of mindset, this problem will worsen.

A centralised model, where vendors do not store personal information but merely request it, giving the user more control over their data is proposed. Figure 2 shows an example model where all permanent storage of personal information is on the users device. A relay service is used to (1) hide the users device, (2) cache encrypted responses, (3) authorise vendors, (4) filter unwanted requests, and (5) provide features automatically like anonymous email.

An example of a trusted vendor is shown in Figure 3 for a user registering, then the website sending a welcome email. On registration the user provides the website with an unique code (different for each site and user), which the website checks before allowing a successfully registration. The website then asks the relay service for the email associated with the unique code. This request is forwarded to the users device prompting them if they trust the website (through an mobile application for example). Since this is a trusted website, the users email address is automatically encrypted with the websites public key, and sent to the relay which forwards it on. After the website has sent an email, the address is deleted from their system (never at rest). By the user always allowing the address to be used by the site (can be disabled again), next time the relay service can quickly reply with the encrypted response. Each request for information can be logged by the relay service to provide the user with useful statistics about their information.

This model does not stop vendors from caching users data without their permission. But for services such as banks and government services, there may be requirements to have full access to the information. For example, if a user changes their postal address in the proposed system, their bank can see they have moved. The bank then can ask for confirmation of the move before updating its records. This prevents statements or replacement cards from accidentally going to the wrong address.

Untrusted vendors or vendors who do not need the users personal information is the same as shown in Figure 3, however the information given is automatically made anonymous. Figure 4 shows the same example as Figure 3 however the email address given is anonymous. This address is directed to the relay service, which maps it to the real address. Note this means the relay service has access to the users real email address, but for this model the relay service is trusted. An untrusted relay service could conduct a man-in-the-middle attack with the encryption keys, therefore for this paper the relay is trusted.

The user only needs to interface with an application, such as an iPhone app. Any notifications can easily be pushed to the application for user input, for example Figure 5. Online account registration is also easy for the user. An example is shown in Figure 6 where a longer registration form can be replaced with a shorter one. How a random username can be generated automatically is discussed in Section VIII-C. The user code can be generated in the application and copied to the input box, or browser plugins could aid this process.

Details of the user code are implementation specific, and are not discussed in detail. The purpose of the user code in this paper is a means of linking an anonymous account to the real account via the relay service. Further authentication or security measures may need to be applied. Also note that there is no discussion around public key exchange as again this is an implementation problem. However the relay service

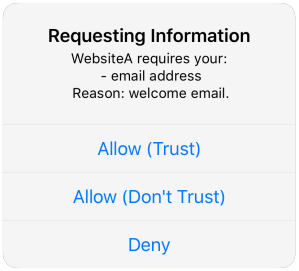


Fig. 5. Sample notification showed to the user.

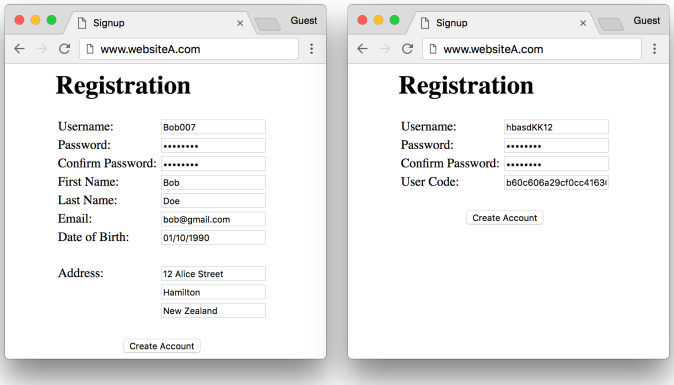


Fig. 6. Change of registration page, where the left shows the current mindset, and the right shows the proposed model.

could act as a verifier or holder of public keys so the user knows the encryption key for the vendor is correct.

## V. Protecting Private Information

### A. Customers Name

Companies often use a customers name in letters and emails to make them feel more personal. But it should not be a necessity to divulge private information for this purpose. To maintain this personal connection, an email merely needs to use a placeholder or tag, for example *$user.firstname$*. When an email client is displaying this message, it can replace the placeholder on the local machine, keeping their name private.

If the company requires the customers name for warranty purposes, then the name only needs to be revealed if the customers has a problem with their product. This can be achieved by the company receiving an encryption of the name where the customer has the decrypt key. If the customer wants to return the product, then they can give the company the decryption key or password. This keeps the user in control of their data.

### B. Email Address

An email address is similar to an IP address, where it provides an endpoint for the message. IP addresses can be kept private from senders by hiding behind proxies. This concept can be applied to email addresses where an anonymous address is used. When the mail server receives the email, it can replace the destination address to the real users address and resend it [24]. The messages should also be encrypted to further improve privacy. Therefore instead of a website gaining access to their real email address, an anonymous one can be automatically generated in conjugation with a public key.

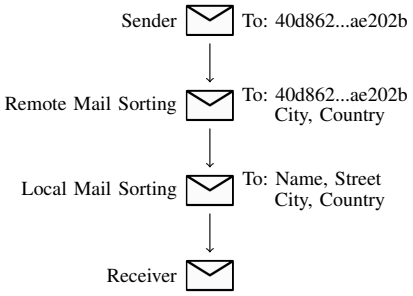


Fig. 7. Privacy Preserving Mail Delivery Flow Diagram

### C. Postal Address

The only entities that need to see a users postal address are the postal companies. A vendor only needs to know how much to charge for postage. In fact some postal services charge for delivery upon receiving the package, for example the New Zealand Post YouShop service [25]. With either method of payment, the vendor just needs to address the item with a unique identifier which the postal services can use to access the real postal address.

Figure 7 shows and example of an envelope being delivered. The sender addresses the envelope to a random value. Upon receiving the envelope, the first postal service has permissions to see the city and country of the address. Once the envelope reaches the local mail service, they have permission to request the full postal address. Then the envelope can be delivered to the current address. Therefore only the local mail service requires the full address, and the remote mail services only need to get the envelope to the local mail service.

### D. Credit Card

When purchasing an item online, pages often ask if a user would like to save this card for future use. Even when declining this option, the credit card details may have been visible and therefore retrievable in the remote system. Information from this card could be used to identify the user, and used to illegally purchase other goods. Solutions exist to protect privacy and fraud, such as using a prepaid virtual credit card, Bitcoin, or trying to use PayPal anonymously [26]. These solutions do not provide a user-centric experience as they can be difficult to setup, slow down payments due to the need to transfer money or incur extra fees.

Instead banks and credit card services should provide a means of generating random single-use codes. Note that the Issuer Identification Number (IIN) would still be required so that the remaining random digits can be mapped to the real account. This prevents the code from being used again for malicious purposes, and can protect the users privacy. Another

possibility would be a predefined sequence of security codes. On verification of the card, the bank can verify the security code is the next one in the queue.

The protection of credit card information can also be achieved by encrypting the details with the banks public keys. This means the store receives the card details, but cannot make sense of them as there are encrypted. However the banks need to verify the card and user hence the details are encrypted with a public key managed by the bank.

### E. Date of Birth

Companies rarely require a customers Date of Birth (DoB), other than validating they are of a certain age, for example in New Zealand one must be over 18 to purchase alcohol. With our proposed system model, the site only needs to query if the user is older than 18. Other uses of the DoB are for identifying birthdays for discounts, but again this can be achieved by querying if they have a birthday this month. A few website registration pages asked for an age range for content recommendations, which could also be queried for.

### F. Phone Numbers

With the increasing popularity of the Voice Over Internet Protocol (VoIP) [27], one time or temporary contact numbers are becoming a reality for anonymous calls [28]. Therefore instead of a user giving out their real phone number, an anonymous one can be used. Then in the case of receiving spam calls, for example someone claiming to be from Microsoft [29], the number can simply be discarded.

### G. Warranty Claims

In New Zealand, under the Consumer Guarantees Act 1993 consumers have a right to a one-year manufacturer warranty. To claim a warranty, the consumer must have proof of purchase, for example a receipt. The vendor does not require any personal information to be able to honour a warranty.

## VI. Features and Possibilities

### A. Single Point for Updating Information

A major problem with the current model given in Figure 1 is the ability to update information easily. Currently a user has to remember and login to each account, where the proposed model they only need to update the information once. Even though this is not a security or privacy issue, it is relevant to GDPR (described in Section II-C).

### B. Disabling Access to Information

Removing information from accounts and vendors is sometimes difficult. For example, unsubscribing from an email mailing list. Even when successful there is no guarantee the address was deleted from their servers. By giving each site/company a different anonymous email address, the mapping can be simply disabled to prevent any further emails. This is important for privacy and also the GDPR (described in Section II-C).

### C. Querying over Explicit Requests

Instead of asking for information directly, often a vague query is sufficient. This can be seen in examples of registration pages shown in Section III-A, where an age range is requested in place of the date of birth. Querying however does have limitations if a change of response is encountered. For example, if a user was unable to satisfy the age requirement for the purchase of alcohol in a particular month but able to do so in the next month, the user's age can be inferred. This is still better than the site having the users full date of birth.

### D. Verification

Websites such as Trade Me allow a user to have their address verified to increase trading trust [30]. With the proposed model, the relay service could verify users, allowing Trade Me to query the result. However, if the relay service does not store any plain-text information, the user cannot be verified. A possible solution would be to treat the relay service as a trusted party and allow it to request for user information. For the Trade Me example, the relay service could request the users address, and send a verification code in the post (similar to Trade Me [30]).

### E. Two-Factor Authentication

Even though two-factor authentication can be viewed as ineffective against modern attacks [31], there are still benefits. Mobile devices make a good token [32] as a large percentage of people carry them around wherever they go. SMS is a simple solution for two-factor authentication however the delivery media is not guaranteed to be secure [33]. For the proposed model, an application is required on the users device, which could be used to provide two-factor authentication similar to the Authy application [34]. Because many modern mobile devices now come with biometric fingerprint sensors, they could also be used to further improve security [35]. The difference from the likes of Authy is the user is still anonymous as all requests go through the relay service.

### F. Email Address Protection

There are a few approaches to providing an email relay for mapping an anonymous email address to a legitimate one. The easiest solution is to setup a mail server which contains a lookup table for translating the addresses before resending the email, and a means of creating and deleting entries. Services exist which offer this today, where most also provide an anonymous inbox for the user to manage. However these offer insufficient privacy protection, and therefore are not addressing the problem of returning control of data to users. This is due to the service having access to the emails being received, and the real email address of the user. Other limitations are the size of the lookup table, and preventing email loops.

Removing the ability to easily delete address mappings provides another possibility for email service providers to implement. The maximum length of an email address is 320 characters (2560 bits), where 64 characters are allowed before the "@" and 255 characters after [36]. Therefore there is

enough bits to encrypted the first 32 bits of an email address with a 2048-bit RSA public key (allowing all 8-bit characters $0-255$ in the address). An example of an @gmail.com address is shown below, where a byte is encrypted with a public key generated by Google.

$$E(b_x) \rightarrow \text{<64 bytes>@<192 bytes>.encrypted.gmail.com}$$

When the remaining bytes are encrypted, gives 16 cipher email addresses. Sending an email now requires the first cipher email to be in the *to* field, and the rest in the *cc* field. Note that mail clients will need to be configured to send 1 email, not send 16. When the encrypted.gmail.com servers receive the message, they can decrypt the parts and form the original email address. Because this would be implemented by Google, if a user replies the address can be hidden again, with one address in the *from* field, with the rest in the *cc* field.

### G. Data Provenance

Traceability of the data can also be augmented to the proposed model from the client and vendor side using data provenance. Data provenance refers to the metadata that captures the derivation history of a piece of data [37] and is conceptualised as a graph [38]. Through the data provenance, the *who*, *when* and *how* regarding interactions with the data can be deduced.

On the client side, provenance of the user's information can be created by recording and modelling vendor requests into a data provenance graph. The graph would capture *when* and *what* information was requested by the different vendors. With just information from the client side, the data provenance graph would resemble a timeline showing the various requests by different vendors across time. However, the graph can be augmented with information from the vendor side to show how the data was being used.

Third-parties provide little transparency and auditability of users data while in transit or rest within their systems [1]. Provenance capturing tools such as Progger [39] and Pegasus [40] can be deployed at the vendor side to capture the provenance showing how the requested user information is being used. Upon request by the clients, vendors can choose to send the relevant portions of the tamper-evident provenance graph to the clients. When augmented with the client side data provenance graph, it would show various requests across time and how each vendor used the information requested, to the users via a visual representation [41].

Other options for vendors to provide more trust to the users is to meet certain standards and certifications. International standards are now emerging for cloud service providers, with ISO/IEC 27018:2014 (with ISO/IEC 27002 as one of its normative references) being the first International code of practice that focuses on protection of Personally Identifiable Information (PII) in the cloud.

### H. Visualisation

A potential method of returning control to users with the use of their personal information is to visualise the personal information life cycle (PILC) and see how it has been created, used, updated and stored. In addition, as a user, having the ability to be in the loop with personal information and the integration of all personal information attributes when used can be shown using visualization. Visualization can provide to users the ability to carry out analysis on the PILC and dissemination of personal information [42].

## VII. Challenges

### A. Advertising

Many people would prefer a world without advertising, however it is essential to many *'free'* services on offer today. A survey in 1999 on Internet advertising [43] showed that over half of the people sampled did not mind online advertising. Therefore unless users start paying for more services (for example Facebook), online advertisements are not going to disappear.

Other possibilities for targeting advertisements is on the client side. A browser stores all of a user's recent history, which combined with personal information such as age, gives more powerful advertising capabilities. The device could receive a diverse range of advertisements [44], and the local device can decide which one the user would probably prefer. Another option would be for vendors to request from the user types of advertisements they would like to see using the proposed model.

This same technique could be applied to businesses buying personal information for brand awareness and product development. Where the users are surveyed on what they would like to see. Therefore instead of personal information being collected and sold with no control from the user, companies could get this information through other means. Surveys for example allow the user to have control over what information they are giving away.

### B. Law Enforcement

The proposed model is not going to stop law enforcement agencies from tracking a users activity. However it will be more difficult as there are more companies they need to request data from. In order for the relay service to provide the account details for a user by request of a law enforcement agency, it needs a way of knowing the name of the user. One method would be to store a hash of each users name. Then if a user needs to be found, their name could be hashed before trying to find a match. This could provide false-positives, but could be checked by the relay service requesting the users name.

An alternative would be for the relay service to securely store the personal information of each user for the sole purpose of complying with law enforcement. This should be done in a manor that decryption can only be computed offline. Thus the decryption key should not be accessible automatically, to prevent a data breach in the event of an attack.

### C. Password Reset for Relay Service

The relay service is going to need some personal information, for example mobile number and email address in

order for password reset. Section VII-B details a few methods for complying with law enforcement agencies, which could help for password reset capabilities while protecting user information. However if the relay service is acting as an email relay, the email address we need to be visible regardless.

### D. Online Tracking and Information Linkage

The structure of an IP address is such that your location on the Internet can be found. Therefore they can be approximately mapped to a real location. Techniques exist to hide your IP address such as proxies [45] and onion routing [46][47]. The Tor software [48][49] is an example of onion routing and can be simple to setup on a personal computer, however not easy for mobile devices and other applications. Tor claims to secure, by encrypting data throughout the route taken protecting privacy. However the security has been compromised [50][51]. Proxies also have issue, for example not always allowing secure traffic and can act as a malicious man-in-the-middle. Therefore currently for the regular Internet user IP address anonymisation is still a challenge. The introduction of 5G communications could lead to more challenges or possible improvements for this problem [52].

Another challenge for user privacy and protection of personal information is the online tracking between sites. For example every link on Google actually points to back Google, before redirecting to the intended address. This allows Google to track the links and sites a user is visiting [44]. In 2016, Google Chrome had $\approx 50\%$ market share on desktop computers [53]. The issue here is Google Chrome can be linked to a users Google account, allowing Google to learn even more about a user.

To help protect personal information from being linked from different vendors, the proposed model has the unique user code. Then if the user does not trust the vendor, they only receive anonymous data. This prevents any linkage between the accounts using personal information or account details. However, information can also be linked together using a user's browser fingerprint [54], IP address or cookies [55]. Hiding IP addresses have been discussed earlier in this section, where fingerprinting and cookies are more problematic. These issues probably need to be addressed by browser vendors [56].

## VIII. Related Works and Technologies

### A. End-to-End Encryption

PGP (Pretty Good Privacy) was a program developed in the mid 1990s which could encrypt a users email so that only the intended recipient could read it [57]. However today protecting privacy by sending emails with end-to-end encryption is still not easy for the average user, and both Google and Yahoo have projects to try and solve this problem. "Engineers from Google, Yahoo, and the open source community continue to work together on the End-To-End Mail extension project. It remains a work in progress" [58]. In a Google Transparency Report on Safer Emails, they state that an email will be encrypted in transit (if the other email provider supports TLS encryption), but this does not encrypt data at rest [59]. Therefore still a simple email can expose personal information and risk a users privacy. This could be a potential issue with using an email relay like in Figure 4.

### B. Secure Data Processing

Homomorphic encryption [60][61][62] and other secure processing techniques [63][64] allow data to remain private, while still being able to be processed. For example encrypting some data, computing the average over the cipher texts, and after decryption the result is revealed. These concepts can be applied to cloud services to protect user privacy, for example voting [65]. This could also allow the relay service in the proposed model to keep statistics in a secure manor.

### C. Password Management Tools

Password reuse can allow an attacker to gain access to multiple accounts after a single leak [66]. Tools which allow a user to automatically generate random passwords without the need to remember them are very convenient [67]. Examples are 1Password [68], Passpet [69], and Apple's built in iCloud Keychain [70]. However these still offer a single point of failure, and in most cases a compromised password management account is more serious than a few online accounts using the same password. The significances to user privacy of such tools is the ability to also create random usernames, as well as random passwords. For example, the flow diagrams in Figures 3 and 4 for registration could use such tools, as well as the sample website interface in Figure 6.

### D. Other Privacy Models

Many proposed privacy models try to address the challenge of anonymousing personal information [71][72][73][74]. However many fail to address usability for user accounts, traceability for law enforcement, verification for payments, anonymous postal addresses, and do not tackle the issue of returning control of data to users.

*1) $k$-anonymity and $\ell$-diversity:* Protecting privacy for data sets, $k$-anonymity states that a user cannot be distinguished from at least $k-1$ users in the same set [71][72]. Suggested as an improvement over $k$-anonymity, $\ell$-diversity hardens against some of the known attack vectors on $k$-anonymity [73]. These work for data sets released by trustworthy vendors, such as healthcare releasing data sets for researchers. However vendors selling data online cannot always be trusted, and attackers releasing personal information on the black market are not going to protect privacy. The proposed model allows a user to choose if they trust a vendor or not, meaning anonymous information can be forced. Where $k$-anonymity cannot help protect a single user for account details. The model also states that no personal information should be saved within a vendors ecosystem, limiting attackers ability to access and release personal information.

*2) Privacy-Preserving Linkage:* Data linkage on personal information (such as names and addresses) in areas such as Health, census, and social security, help improve data processing and analysis [75]. This is made possible using several

privacy-preserving data linkage and extraction protocols across separate sources (databases). Enabling the sharing of data without the requirement of identifying any information to a specific person or data sources [76]. Other data privacy-preserving techniques are associated with geocoding whereby addresses are matched to their geographic locations [77]. As a result, these protocols and techniques brings in paramount importance to privacy and confidentiality to private (personal) information. Challenges include miss-matching specific information to incorrect identities, or having to check which data attributes are associated with the correct person [77].

## IX. Concluding Remarks

The current mindset of distributing personal information is flawed and gives users little confidence or control over their data. It results in an expanding data universe, but can be reduced with a personal data crunch. In this paper, a model where personal information is stored on a single entity is proposed. Through such an entity, appropriate responses can be returned to vendors requiring user information, based on their trust status. Most privacy models (described in Section VIII-D) focus on published datasets, not personal information for user accounts. While other frameworks [18] are not designed for privacy and anonymity, and store personal information in plain-text. The other research gap covered in this paper is returning control of data to users, empowering them to govern their own security and not be forced into handing out personal information. In future work, an implementation of the proposed model will be developed in order to fully evaluate both its privacy and security properties. Combined with other research such as secure data processing and provenance loggers, could lead to more privacy, trust and security within the Internet.

## Acknowledgements

This research is supported by STRATUS (Security Technologies Returning Accountability, Trust and User-Centric Services in the Cloud) (https://stratus.org.nz), a science investment project funded by the New Zealand Ministry of Business, Innovation and Employment (MBIE).

## References

[1] R. K. Ko, G. Russello, R. Nelson, S. Pang, A. Cheang, G. Dobbie, A. Sarrafzadeh, S. Chaisiri, M. R. Asghar, and G. Holmes, "STRATUS: Towards Returning Data Control to Cloud Users," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2015, pp. 57–70.

[2] C. Isidore and D. Goldman, "Ashley Madison hackers post millions of customer names," Online. http://money.cnn.com/2015/08/18/technology/ashley-madison-data-dump/ (Accessed 16/01/17), August 2015.

[3] T. Fox-Brewster, "Yahoo: Hackers Stole Data On Another Billion Accounts," Online. http://www.forbes.com/sites/thomasbrewster/2016/12/14/yahoo-admits-another-billion-user-accounts-were-leaked-in-2013 (Accessed 18/01/17), December 2016.

[4] A. Chen, "GCreep: Google Engineer Stalked Teens, Spied on Chats," Online. http://gawker.com/5637234/gcreep-google-engineer-stalked-teens-spied-on-chats (Accessed 16/01/17), September 2010.

[5] S. Lilley, F. S. Grodzinsky, and A. Gumbus, "Revealing the commercialized and compliant facebook user," *Journal of Information, Communication and Ethics in Society*, vol. 10, no. 2, pp. 82–92, 2012.

[6] V. Bolotaeva and T. Cata, "Marketing opportunities with social networks," *Journal of Internet Social Networking and Virtual Communities*, vol. 2010, pp. 1–8, 2010.

[7] B. Krishnamurthy, K. Naryshkin, and C. Wills, "Privacy leakage vs. protection measures: the growing disconnect," in *Proceedings of the Web*, vol. 2, 2011, pp. 1–10.

[8] D. Malandrino, A. Petta, V. Scarano, L. Serra, R. Spinelli, and B. Krishnamurthy, "Privacy awareness about information leakage: Who knows what about me?" in *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, ser. WPES '13. New York, NY, USA: ACM, 2013, pp. 279–284. [Online]. Available: http://doi.acm.org/10.1145/2517840.2517868

[9] EUGDPR.org, "GDPR Key Changes," Online. http://www.eugdpr.org/key-changes.html (Accessed 13/12/16).

[10] A. Macrae, "GDPR – The Good, the Bad and the Ugly," Online. https://www.tripwire.com/state-of-security/security-awareness/gdpr-the-good-the-bad-and-the-ugly/ (Accessed 12/01/17), February 2016.

[11] A. Olshanskaya, "Why the GDPR is good for business," Online. https://iapp.org/news/a/why-the-gdpr-is-good-for-businesses/ (Accessed 12/01/17), December 2016.

[12] R. Beckwith and S. Mainwaring, "Privacy: Personal information, threats, and technologies," in *Technology and Society, 2005. Weapons and Wires: Prevention and Safety in a Time of Fear. ISTAS 2005. Proceedings. 2005 International Symposium on*. IEEE, 2005, pp. 9–16.

[13] E. Berki and M. Jäkälä, "Cyber-identities and social life in cyberspace," *Social Computing: Concepts, Methodologies, Tools, and Applications*, pp. 92–104, 2009.

[14] L. Roberts, "Cyber identity theft," 2009.

[15] N. Chou, R. Ledesma, Y. Teraguchi, J. C. Mitchell *et al.*, "Client-side defense against web-based identity theft." in *NDSS*, 2004.

[16] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 2009, pp. 7–12.

[17] J. Sovern, "Opting in, opting out, or no options at all: The fight for control of personal information," *Wash. L. Rev.*, vol. 74, p. 1033, 1999.

[18] J. Altmann and R. Sampath, "Unique: A user-centric framework for network identity management," in *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*. IEEE, 2006, pp. 495–506.

[19] P. A. NORBERG, D. R. HORNE, and D. A. HORNE, "The privacy paradox: Personal information disclosure intentions versus behaviors," *Journal of Consumer Affairs*, vol. 41, no. 1, pp. 100–126, 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1745-6606.2006.00070.x

[20] L. D. Roberts, D. Indermaur, and C. Spiranovic, "Fear of cyber-identity theft and related fraudulent activity," *Psychiatry, Psychology and Law*, vol. 20, no. 3, pp. 315–328, 2013.

[21] P. M. Thomson, "Criminal law & procedure," *Letter from the Editor*, p. 23, 2015.

[22] S. Romanosky, R. Telang, and A. Acquisti, "Do data breach disclosure laws reduce identity theft?" *Journal of Policy Analysis and Management*, vol. 30, no. 2, pp. 256–286, 2011.

[23] "Fundamental texts / Legal materials / About INTERPOL / Internet / Home - INTERPOL." [Online]. Available: https://www.interpol.int/About-INTERPOL/Legal-materials/Fundamental-texts

[24] I. Goldberg, D. Wagner, and E. Brewer, "Privacy-enhancing technologies for the internet," DTIC Document, Tech. Rep., 1997.

[25] N. Z. Post, "Youshop," Online. https://www.nzpost.co.nz/tools/youshop (Accessed 17/01/17).

[26] T. Knauer, "How To: Pay Anonymously Online," Online. https://greycoder.com/how-to-pay-anonymously-online/ (Accessed 10/01/17), February 2016.

[27] B. Goode, "Voice over Internet Protocol (voip)," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1495–1517, 2002.

[28] S. Chen, X. Wang, and S. Jajodia, "On the anonymity and traceability of peer-to-peer VoIP calls," *IEEE Network*, vol. 20, no. 5, pp. 32–37, 2006.

[29] "Avoiding technical support scams," Online. https://www.microsoft.com/en-us/safety/online-privacy/avoid-phone-scams.aspx (Accessed 17/01/17).

[30] "Becoming a trusted trader," Online. http://www.trademe.co.nz/seller-information-centre/new-sellers/becoming-a-trusted-trader/ (Accessed 17/01/17).

[31] B. Schneier, "Two-factor authentication: too little, too late." *Commun. ACM*, vol. 48, no. 4, p. 136, 2005.

[32] F. A. Aloul, S. Zahidi, and W. El-Hajj, "Two factor authentication using mobile phones." in *AICCSA*, 2009, pp. 641–644.

[33] R. E. Koenig, P. Locher, and R. Haenni, "Attacking the verification code mechanism in the norwegian internet voting system," in *International Conference on E-Voting and Identity*. Springer, 2013, pp. 76–92.

[34] "Authy," Online. https://www.authy.com (Accessed 18/01/17).

[35] A. T. B. Jin, D. N. C. Ling, and A. Goh, "Biohashing: two factor authentication featuring fingerprint data and tokenised random number," *Pattern recognition*, vol. 37, no. 11, pp. 2245–2255, 2004.

[36] J. Klensin, "Application Techniques for Checking and Transformation of Names," *RFC 3696*, 2004.

[37] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science," *ACM SIGMOD Record*, vol. vol. 34, no. 3, pp. pp. 31–36, Sept 2005. [Online]. Available: http://doi.acm.org/10.1145/1084805.1084812

[38] L. Carata, S. Akoush, N. Balakrishnan, T. Bytheway, R. Sohan, M. Seltzer, and A. Hopper, "A Primer on Provenance," *Communications of the ACM*, vol. 57, no. 5, pp. pp. 52–60, May 2014.

[39] R. K. Ko and M. A. Will, "Progger: An Efficient, Tamper-Evident Kernel-Space Logger for Cloud Data Provenance Tracking," in *2014 IEEE 7th International Conference on Cloud Computing*. IEEE, 2014, pp. 881–889.

[40] J. Kim, E. Deelman, Y. Gil, G. Mehta, and V. Ratnakar, "Provenance Trails in the Wings-Pegasus System," *Concurrency and Computation: Practice & Experience, Journal*, vol. vol. 20, pp. pp. 587–597, 2008.

[41] J. Garae, "User-centric Visualization of Data Provenance," *Master of Cyber Security (MCS), The University of Waikato*, 2015.

[42] J. Garae, R. K. Ko, and S. Chaisiri, "Uvisp: User-centric visualization of data provenance with gestalt principles."

[43] A. E. Schlosser, S. Shavitt, and A. Kanfer, "Survey of internet users' attitudes toward internet advertising," *Journal of interactive marketing*, vol. 13, no. 3, pp. 34–54, 1999.

[44] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proceedings Network and Distributed System Symposium*, 2010.

[45] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions," *ACM transactions on information and system security (TISSEC)*, vol. 1, no. 1, pp. 66–92, 1998.

[46] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected areas in Communications*, vol. 16, no. 4, pp. 482–494, 1998.

[47] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[48] "Tor," Online. https://www.torproject.org/ (Accessed 19/01/17).

[49] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," DTIC Document, Tech. Rep., 2004.

[50] J. Stone, "Is Tor Safe? Anonymous Browser Hacked, With Suspects Keeping Quiet and Privacy Advocates Shaken," Online. http://www.ibtimes.com/tor-safe-anonymous-browser-hacked-suspects-keeping-quiet-privacy-advocates-shaken-1645210 (Accessed 19/01/17), July 2014.

[51] C. Smith, "How the Dark Web's favorite anonymity tool got hacked," Online. http://bgr.com/2015/12/02/tor-dark-web-security-hack/ (Accessed 19/01/17), December 2015.

[52] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[53] N. Applications.com, "Desktop browser market share," Online. https://www.netmarketshare.com/browser-market-share.aspx?qprid=0&qpcustomd=0&qpsp=2016&qpnp=1&qptimeframe=Y (Accessed 19/01/17), January - December 2016.

[54] P. Eckersley, "How unique is your web browser?" in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2010, pp. 1–18.

[55] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 674–689.

[56] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless monster: Exploring the ecosystem of web-based device fingerprinting," in *Security and privacy (SP), 2013 IEEE symposium on*. IEEE, 2013, pp. 541–555.

[57] S. Garfinkel, *PGP: pretty good privacy*. " O'Reilly Media, Inc.", 1995.

[58] L. Franceschi-Bicchierai, "The Dream Of Usable Email Encryption Is Still A Work In Progress," Online. http://motherboard.vice.com/read/google-yahoo-end-to-end-email-encryption-work-in-progress (Accessed 16/01/17), March 2016.

[59] Google, "Transparency Report, Security and Privacy, Safer Email," Online. https://www.google.com/transparencyreport/saferemail/faq/ (Accessed 16/01/17).

[60] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," *Foundations of secure computation*, vol. 4, no. 11, pp. 169–180, 1978.

[61] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Stanford University, 2009.

[62] M. A. Will and R. K. L. Ko, "A guide to homomorphic encryption," in *The Cloud Security Ecosystem: Technical, Legal, Business and Management Issues*. Elsevier, 2015, vol. 1, pp. 101–127.

[63] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," *Journal of Privacy and Confidentiality*, vol. 1, no. 1, p. 5, 2009.

[64] M. A. Will, R. K. Ko, and I. H. Witten, "Privacy preserving computation by fragmenting individual bits and distributing gates," in *The 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-16)*, vol. 1. IEEE, 2016, pp. 900–908.

[65] M. A. Will, B. Nicholson, M. Tiehuis, and R. K. Ko, "Secure voting in the cloud using homomorphic encryption and mobile agents," in *2015 International Conference on Cloud Computing Research and Innovation (ICCCRI)*. IEEE, 2015, pp. 173–184.

[66] S. Gaw and E. W. Felten, "Password management strategies for online accounts," in *Proceedings of the Second Symposium on Usable Privacy and Security*, ser. SOUPS '06. New York, NY, USA: ACM, 2006, pp. 44–55. [Online]. Available: http://doi.acm.org/10.1145/1143120.1143127

[67] L. Tam, M. Glassman, and M. Vandenwauver, "The psychology of password management: a tradeoff between security and convenience," *Behaviour & Information Technology*, vol. 29, no. 3, pp. 233–244, 2010. [Online]. Available: http://dx.doi.org/10.1080/01449290903121386

[68] "1password," Online. https://1password.com (Accessed 18/01/17).

[69] K.-P. Yee and K. Sitaker, "Passpet: convenient password management and phishing protection," in *Proceedings of the second symposium on Usable privacy and security*. ACM, 2006, pp. 32–43.

[70] A. Inc., "Frequently asked questions about icloud keychain," Online. https://support.apple.com/en-sg/HT204085 (Accessed 18/01/17).

[71] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical report, SRI International, Tech. Rep., 1998.

[72] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[73] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217302

[74] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Managing dimensionality in data privacy anonymization," *Knowledge and Information Systems*, pp. 1–33, 2015.

[75] P. Christen, "Privacy-preserving data linkage," *Tutorial at The Australasian Data Mining Conference*, 2008.

[76] C. M. O'Keefe, M. Yung, L. Gu, and R. Baxter, "Privacy-preserving data linkage protocols," in *Proceedings of the 2004 ACM workshop on Privacy in the electronic society*. ACM, 2004, pp. 94–102.

[77] P. Christen, "Privacy-preserving data linkage and geocoding: Current approaches and research directions," in *Data Mining Workshops. ICDM Workshops.* IEEE, 2006, pp. 497–501.