Electronics and Computer Science
Faculty of Physical and Applied Sciences
University of Southampton

<Umar Anwar>

<2016>

<Data Anonymization>

Project supervisor: <Vladimiro Sassone>
Second supervisor: <Paul Brittan>

<A report submitted for GCHQ>

Executive Summary

The evolution of data in the real world as well as online has exceedingly increased over the last decade or so. With the likes of social media platforms such as Myspace, becoming very popular for status updates, blogs, entertainment purposes and sharing information with friends. At a later stage then came Facebook consisting of a more dynamically social platform for friends, families and photo sharing. Then last and most recently came Twitter that took a keen interest on a global scale, adapting to all types of environments for different users. Whether it was for business purposes or standard communication between people.

In particular Facebook and Twitter have drastically become more and more trendy as the years' progress and it continues to do exactly that. As both platforms have continued to increase in popularity, the risk of privacy and data of users has also risen to an enormous amount, thus leaving a dominantly vulnerable scale of information widely available for exploitation.

There's also a demand for a lot generic and demographical information that is required to help for research purposes. However, this poses a threat towards individuals as the data is given in a form of microdata which could directly or indirectly relate back to them, leaving the potential misuse of their information, as an unauthorised user may be able capture that set of data, resulting in the exposure towards being able to re-identify an individual and learn a lot of detailed sensitive and confidential information contained within the dataset.

There are two separate topics that will be widely discussed throughout this report in relation to anonymity, privacy and data.
These are:

- **Anonymization**
- **De-Anonymization**

Subsiding within these topics, there will be a discussion consisting of the different techniques that are used for **anonymizing** and also **de-anonymizing** data. Analysing the approach, dataset and results that each technique depicts, through a representation of the authors views expressed.

There are a number of anonymization techniques that have been discussed throughout a variety of research papers analysed within this report. Here are a number of techniques identified:

Number of Papers Discussing These Anonymization Techniques			
K-anonymity	**l-diversity**	**Tokenization**	**AES**
3	3	1	1

Contents

1 Introduction

The aim of this research task is to gather a collection of information on the data anonymization spectrum and the way in which all the past developments made compare to the newly proposed methods in anonymizing data in the most secure and proficient manner possible. Whilst also focusing on the de-anonymization processes that were previously used and the new approaches taken towards the re-identification of data. Whilst identifying the different techniques useably available residing amongst anonymizing and de-anonymizing data, presenting a comparison of the two coinciding with the strengths and weaknesses between with one another.

2 Literature Review

The aim of the literature review is to showcase the many existing and newly discovered anonymization and de-anonymization techniques expressed by researchers and to analyse each paper in a data driven and efficient manner. Primarily focusing on what the authors approach, dataset and results entail, whilst also considering other recommendations conveyed through till the end.

2.1 Anonymization

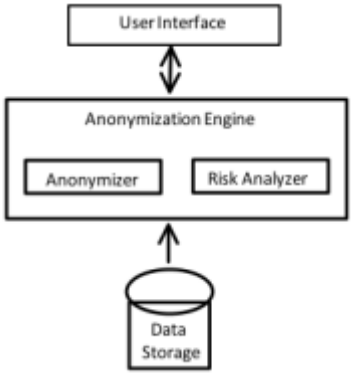
A process to sanitize information with the goal to protect privacy (either by encrypting or removing PII from sets of data), resulting in data remaining anonymous.

2.1.1 Interactive Anonymization of Sensitive Data

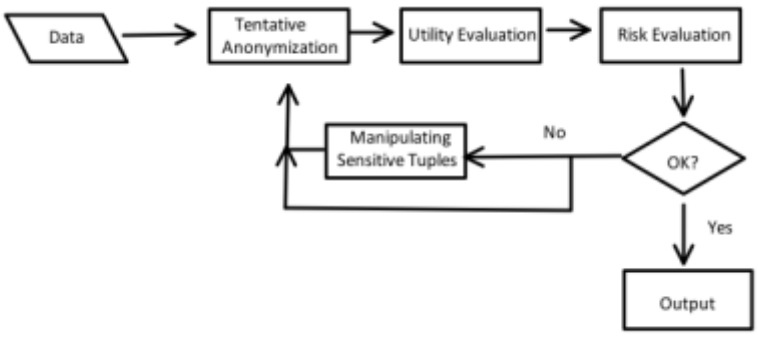
(Xiao, Wang, & Gehrke, 2009) proposes the use of a comprehensive usable toolkit for easy use with the ability to control the limit of disclosure when publishing data. Predominantly aimed at organizations (such as medical facilities or large governmental institutions) that tend to collect large sets of data that contains personal and sensitive information. When dealing with such data, it is important that the publication process is carried out properly. If carelessness occurs whilst undergoing this process, then this poses a grave threat towards the privacy of individuals not only who have contributed such data, but also the data itself that may contain other personal information.

Technique:
There has been a lot of research undertaken to evaluate and determine the most suitable anonymization techniques that can provide thorough anonymity guarantees. After all the research undertaken, a proposal was made to implement an interactive usable tool **CAT** (Cornell Anonymization Toolkit) that can intuitively guide users through the workflow of data publishing. Essentially this toolkit takes all the theoretical material of data anonymization and turns it into a practical usable tool. **CAT** is aimed to help users develop an intuitive understanding of the disclosure risk that resides in anonymized data, with a more educated decision-making mechanism upon the release of appropriate data. As well as offering users the ability to have control over the whole anonymization process (with the use of **l-diversity algorithm**) (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007), allowing access to changing any parameters and being able to examine the anonymized data itself in terms of quality (privacy and utility).

System Architecture:



Anonymization Process:



Dataset:
Contingency table (**Appendix A - I**) consisting of "Gender" and "Marital status" showing the correlation between the different pairs of attributes.

Results:
The system firstly calculates the disclosure risk of the records within the dataset based upon attributes within the background knowledge that the adversary may have (bottom right box in **Appendix A – II**). This is part of the process to be able to determine the risk of sensitive tuples that need to be removed and which can remain.

The user interface (**Appendix A – II**) illustrates all of the parameters used such as **l-diversity algorithm** to control **l** & **c**, two density graphs (original and generalized) to depict the corresponding contingency table and also a histogram that has a risk threshold of 20%(If the tuples in the dataset meet below the threshold then the privacy guarantee is deemed sufficient enough). All of these aspects can be altered to output a privacy and utility guarantee through the dataset's overall anonymization process.

2.1.2 Effects of Data Anonymization on the Data Mining Results

(Buratoviç, Miliçeviç, & Zubrinic, 2012) examine privacy preserving techniques for data mining, in an attempt to dignify whether or not it is feasible to anonymize data for research purposes. However, (Buratoviç, Miliçeviç, & Zubrinic, 2012) express that for data mining techniques to be effective with discovering patterns and relationships, the data published must be released in the original form of individual tuples i.e. microdata, as representative statistics or pre-aggregated data cannot be used for data mining purposes due to the lack of flexibility.

Technique:
(Agrawal & Ramakrishnan, 2000) depicts privacy preserving data mining as "a paradigm of exercising data mining while protecting the privacy of individuals". Sanitization is the process of removing all explicit identifiers (e.g. name, PIN, address) that can directly identify an individual from a set of released records. Although there may be some sense of security present, there are still threats posed such as linkage attacks, due to the possibility of de-identified data that may contain other data from quasi-identifiers, that can be uniquely combined and linked with publicly available information for the re-identification of individuals. To avoid such attacks and at the same time preserve the integrity of the released data, (Samarati & Sweeney, 1998) propose **k-anonymity** that uses generalization and suppression on a dataset.

An attribute relevancy estimator **Relief*F*** is also used to aid in fulfilling the most robust means of anonymization with preservation of utility and privacy in assisting with k-anonymity being achieved. The algorithm essentially estimates the relationships among attributes and then produces an average merit and rank score through the understanding of several different factors (**Appendix B – IV**).

Dataset:
Students' data, containing information of students' background and academic success (such as secondary school success, exam scores and first year university study success rate).

Results:

The aim of enforcing k-anonymity is to ensure that all students data that is released in the form of microdata is protected from any possible attack of re-identification. (**Appendix B – I**) illustrates the generalization and suppression procedure on the hierarchy for students' postal codes. Two groups "2***" majority of all students come from the same city Dubrovnik and Dalmatia (post codes starting with "2"). "OTHER" consisting of the fewer minority. Also the same process for students' GPA (**Appendix B – II**) and hierarchy for secondary schools **(Appendix B – III)**.

The experiment in this instance uses 2-anonymization(***k=2***) entailing that the data within the dataset is altered in a way that the combination of values of quasi-identifiers can be matched to at least 2 students indistinguishably (small dataset hence why basic anonymity level is used). Generalization of the attributes takes place each one step at a time, until anonymity level is satisfied(*k=2*), or there are less than 10% outliers left that can be supressed (to avoid overgeneralization of the significant attributes). During the anonymization process it is imperative to not let the drawbacks outweigh the benefits, it is important to maintain the utility of the data, whilst at the same time preserving students' privacy by considering the order of generalization of the attributes (most significant attributes are less generalized then those of lower importance).

Once the **Relief*F*** algorithm had been applied, (**Table 1**) showcased the hierarchical levels for the chosen attributes and the number of values present for each attribute through anonymization. 22 instances remained after generalization was performed, that did not satisfy the level of anonymity given. (Buratoviç, Miliçeviç, & Zubrinic, 2012) ensured they were suppressed.

Attribute	Distinct values				
	Original dataset	*After 1st step of generalization*	*After 2nd step of generalization*	*After 3rd step of generalization*	*After 4th step of generalization*
Postal code	52	32	19	11	3
Secondary school	35	7	3	-	-
Age at enrollment	9	3	2	-	-
Profession	13	10	5	-	-
Secondary school GPA	158	27	4	-	-
SGE results - Croatian language	114	79	12	4	2
SGE results - Foreign language	190	89	13	4	2
SGE results - Mathematics	54	48	12	4	2

Table 1 - Number of distinct values through the process of anonymization

2.1.3 Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues

(Sedayao & Bhardwaj, 2014) from Intel Corporation and (Gorade, 2014) from the University of Texas at Dallas focus on enterprise businesses and the way in which particularly online data on the cloud can be anonymized and out of an adversary's reach. Combining a triad of anonymization, privacy protection and big data techniques to analyze online usage data whilst protecting the identities of users. The main aim is to improve the usability of Intel's profoundly used internal web portal known as **Circuit** using web page access logs and big data tools. However, the primary concern lies with how to protect Intel employees' privacy, removing Personally Identifiable Information (PII) without affecting the use of big data tools to carry out analysis or the re-identification of log entries to explore unusual behaviour. Previous work holds claims of anonymization being useless, due to the Netflix dataset incident (data being compromised through correlation with IMDb). Apart from (Ohm, 2010) expressing a meaningful point about "the inherent tradeoff in anonymization – utility vs privacy".

Technique:
Propose the creation of an open architecture (open and readily accessible to variety of methods and open source tools) for anonymization used on behalf of carrying out both de-identification and re-identification of web log records.

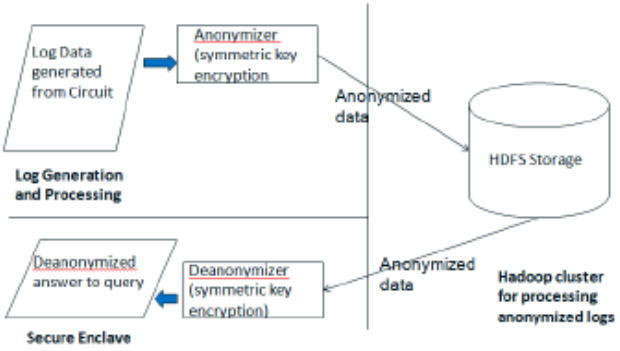


Figure 1 - Anonymization Architecture

(**Fig.1**) shows (Sedayao & Bhardwaj, 2014)'s architecture creation that carries out de-identification and re-identification within a secure enclave. First, sensitive fields in Circuit log files such as IP addresses and user ID's are encrypted using AES (Advance Encryption Standard) symmetric key encryption. Second, with the usage data anonymization process complete, the data is securely moved to Apache Hadoop File System (HFDS), where it stores the Circuit usage and ready for observation by the analysts of Intel (when re-identifying is needed for the log data, the logs can be moved back to the secure enclave and decryption of sensitive fields can occur with the same symmetric key).

Symmetric key encryption is chosen due to the nature of an open architecture, the adaptability of this standard allows simplicity of multiple tools to work on the data (using the same key and encryption standard lets a set of tools to read the same data). An alternative suggested to AES was to use tokenization, mapping a string to each item that needs to be de-identified, however, this would require a large token table (Lane, 2012).

(**Appendix C – I**) is an example of a Circuit log entry, which shows that the PII (IP address) and (username) for the user's official Intel account username "**jcsedaya**" is noticeable in more than one web trend formatted field. Implementing a canonical model for anonymization organized and transformed log records into fields that could fit in that exact format (canonical – organize data into relational database type tables). (Sedayao & Bhardwaj, 2014) placed field references for locating an identifier when it would be found within a field that wasn't an identifier field. To make it easily identifiable, once encrypted, all further occurrences with that identifier would be displayed as the first occurrences encrypted value (e.g. **1 – IP address** or **2 – username**). (**Appendix C – II**) puts this into practice by showing how the reference pointer technique and encrypted values work upon the identifier once encrypted (for username in particular, as there are multiple occurrences with "**2**").

Dataset:

The dataset consists of a Use Case (**Appendix C - III**), for studying purposes of the Circuit logs by Intel's analysts. (**Appendix C – IV**) with the eight most popular searches and traffic counts taken from the "Search terms" and "Aggregate page hits" use case in (**Appendix C - III**).

Results:

In the process of implementing this architecture, (Gorade, 2014) found that enterprise data has different properties to the standard data environment scenarios for anonymization.

Implementation phase

The first implementation phase was carried out in PERL, processing all log files consecutively. As AES was integrated in PERL, the period of anonymization on a day's log took hours to process. However, in the second phase later on, (Sedayao & Bhardwaj, 2014) decided to use Python, replacing PERL as it was a much faster written implementation and reduced the time for processing to minutes(using Hadoop, with the use of Pig Latin code being written to process the data in a more scalable manner).

Sampling one month's data, two use cases from (**Appendix C - III**) were focused on:

1. **Search terms** - top searches for that month
2. **Aggregate page hits** – Implementation of traffic counts by user (number of occurrences)

Implementing the open architecture (**Fig.1**) accompanied by HIVE (data infrastructure providing data querying and analysis), with the handling of the anonymized log files and undertaking traffic counts of the anonymized user names from the anonymized table, allowed decryption of the user names, generating a table of real usernames and traffic counts(re-identification). All in all, this results in a successful implemented anonymization software dealing with two use cases, with the use of big data tools.

Quality check & Improvements

Having masked the IP addresses and user names does not instantly mean that the data is not error-prone to attacks and that individuals can no longer be identified. There still stands the possibility with this anonymization process that an adversary can correlate side knowledge data with some noticeable fields to identify a user that outputted a specific log line. To measure the quality of anonymization, k-anonymity is used.

The higher the *k* value for a set of data, the better and more robust privacy becomes. HIVE was used again to run an analysis on each entry of the sampling month data, after aggregation of time stamps were put in place with one hour intervals. (**Appendix C - V**) illustrates the distribution of *k* values, identifying that the attempt on anonymizing the dataset had actually left it exposed and highly prone to correlation attacks.

Two improvements were put into place, known to be used in anonymizing medical records, the two metrics for privacy:

- *Average risk* - average of all disclosure probabilities and provides a realistic measure of disclosure risk
- *Maximum risk* – highest risk of all those disclosure probabilities (a view of maximum risk to an individual whose activity is recorded in the dataset.

Entropy was considered as a measure for utility. The main focus was on the utility metric of completeness (utility of the data), decreasing as records were removed that were prone to correlation attacks. (**Appendix C - VI**) – effects of removing data on average risk depicts that eliminating local & browser information each time aggregation occurred, the average risk tended to improve a lot. (**Appendix C - VII**) – improving average risk and maximum risk shown that by eliminating log entries that were at a certain vulnerability level, improved average risk, but was mostly dependant on the maximum risk level that was present. (**Appendix C - VIII**) – tradeoff of completeness vs maximum risk entails that the lower the maximum risk, the less completeness (utility) there will be available. Although due to the hourly time stamp aggregation at hand, this doesn't cause much of an issue, where the main cost lies is with a significant loss in precision.

2.1.4 Data Anonymization and Integrity Checking in Cloud Computing

(George & S, 2013) discuss the impact of data storage residing on the cloud and the overly increasing businesses and clients that store their meaningful data in these remote cloud servers without considering to keep a local copy for themselves, and instead put their complete trust in the cloud. There has been work carried out to ensure that these files stored in the cloud does not somehow get lost or corrupted, through the use of data integrity checks. Performing these computation checks on encrypted data can become difficult, instead anonymizing the data can be used as a means of a privacy preserving mechanism.

Technique:

Many proposals have been made on what may be the best way to undergo such a process of enhancing privacy whilst storing important data to the cloud in a secure approach. Such as Provable Data Possession (PDP) based on homomorphic verifiable tags or Proof of Retrievability (POR) using a single cryptographic key.

(Ateniese, et al., 2007) expressed a PDP approach entailing high efficiency based solely upon symmetric key encryption. Where the owner pre-computes a certain number of short verification tokens (each token covering a certain set of data blocks). The data is then passed onto the server, when the owner would like to gather proof of data possession, the server is then challenged with a set of random block indices. The server processes an integrity check on the specified blocks, returning it back to the owner. The returned integrity check must match the given values pre-computed by the owner in order for the proof to hold.

The proposed solution by (George & S, 2013) for preserving privacy at the same time as ensuring data possession is adequately sufficient consists of two divided parts:

1. Data Anonymization
   - Publication of data is usually stored in a table, with each record relating to an individual, consisting of three main attributes:
     I. Key attributes
     II. Non-sensitive attributes (Quasi-identifiers)
     III. Sensitive attributes
   - Within the proposed system, clients are able to send their data that they want to be anonymized along with the anonymization information directly to the secure enclave (secure area in enterprise network), where the data anonymization process is performed.
     Involving these techniques in a progressive order:
     1) **Removing or Obscuring** - Key attributes (name, PIN etc.) subject to information that is provided by the data owner
     2) **Generalization** - Replacing specific values with general values
     3) **Suppression** - Hiding values to ensure they are not released at all
     4) ***k*-anonymity** - Number of occurrences of each distinguishable combination of values within a table. If there are records in the list that are less than *k*, then add anonymous records to the table, enough to satisfy *k*-anonymity and then recalculate the records list. (*k*-anonymous only if each value in the list is $\geq k$)
     5) ***l*-diversity** – Number of distinct sensitive attributes relating with each combination of quasi-identifiers within a table. If there are values in the list that are less than *l*, then add anonymous records to the table, enough to satisfy *l*-diversity and then recalculate the records list. ***(****l*-diverse only if each value in the list of sensitive attributes is $\geq l$*)*

To enhance the privacy of data, the secure enclave uses a combination of techniques outlined above. Once the data has been anonymized it is then sent to the cloud server and stored. For the utilization of de-anonymization, translation tables are created and stored at the secure enclave (values are stored in the translation tables through hashing).

2. Integrity Checking
   - There are two options available for checking the integrity of data:
     I. Checking of original data in secure enclave
     II. Checking integrity of published data in cloud server
   - Both options use almost the same functions, apart from the (II.) being publicly known and differentiating slightly with a ***GenerateResponse()*** function that checks integrity of published data in the cloud server.

Dataset:
New Mexico Expenditure Data used as input data, containing:

- Employee names, ID's, hire dates, organization they work for, title and salary.

Results:
Employee names and identification numbers are the key identifiers that are removed in the process of anonymizing the dataset. Whereas employee hire dates, organization they work for and their titles are quasi-identifiers which undergo the procedure of generalization, suppression, *k*-anonymity and *l*-diversity being applied to them (employee salary is a sensitive attribute that is automatically de-linked from each individual after anonymization is complete).

2.1.5 The Anonymisation Decision-Making Framework (ADF)

(Elliot, Mackey, O'Hara, & Tudor, 2016) on behalf of UKAN publications, have proposed a data anonymization framework that focuses on helping people/owners to anonymize their data in a manageable and well-structured format. The framework known as The Anonymisation Decision-making Framework (ADF) incorporates two important frames of action: consisting of:

1. **Technical Aspect**
   - Thinking about both the quantification of re-identification risk and how it can be managed
2. **Contextual Aspect**
   - Thinking about and addressing those factors that affect the re-identification risk, in particular those of your given data situation (the data and data environment). Such as: data flow, legal and ethical responsibilities and governance practices, responsibility once you have shared or released data and backup plans in the case of a rare event when things go wrong.

The Framework Detail

- Depicts a new way of thinking, or some would argue on already existing mind approaches, as a new means towards the thinking about the re-identification problem.
- Focusing on how to enforce realistic measures of risk, in regards to the data situation. Although, it may seem evident in the importance of the environment in which data is shared and released, over a substantial period of years, the main focus underlying in the data confidentiality field has resided amongst the data itself. Thus, not much thought or effort had been put into looking at data from a larger scope (e.g. only the data alone was being considered, nulling out or even considering any other aspect around it).
  - Due to this, the re-identification risk was seen as emerging from the data. In turn, no there were no concerns with such issues as to why or how re-identification may occur, nor considering what skills, knowledge or other types of data a person would need in order to have a successful attempt of doing so (striving away from any real world analysis).

The ADF is made up of ten components that implements a systematic approach:

1. Describing the data situation
- The relationship between data and its environment
    - The data itself responsibility of the custodian, other data, people who interact with the data and their responsibilities and governance
2. Understanding your legal responsibilities
- Prior to anonymization of data, personal data that may be processed will reflect upon the Data Protection Act (1998)
    - Ensuring someone with suitable skills and knowledge oversees the anonymization process, documenting who does what, when and how, accurately recording the process (good governance).
3. Knowing your data
- Ensure that you understand your dataset
    - Where there are special or unique cases e.g. outliers
    - Which combination of variables are risky (consider carefully which variables may pose a risk to creating safe data) e.g. age & address and what is sensitive information e.g. name
4. Understanding the use case
- Context on how and what data you should release, who will want access that data and how those accessing it may want to use it for different purposes
    - Restrict access to the data e.g. accredited researchers only or you could share it with a wider audience under a different means, essentially restricting a lot of the detailed important information.
5. Meeting your ethical obligations
- Key is that you should do no harm
    - Avoid releasing personal data, e.g. deceased persons as ethical considerations being the cause of distress to the relatives of that person's family, if that data is released publicly.
6. Identifying the processes, you will need to assess disclosure risk
- How a data breach might occur, taking these steps to calculate: 1) Asses the data, 2) Scenario analysis (potential intruders, likelihood of attack), 3) Testing the Scenarios (Penetration testing)
    - How easy is it to manipulate the data and find people in actual practice?
    - Are there other sources of data that can be linked, that have not been considered?
7. Identifying the disclosure control processes that are relevant to your data situation
- Recognising that the technical work Is only a part of the anonymization process
8. Identifying who your stakeholders are and a plan of how you will communicate
- Being sure of what you want to explain about your anonymization method and how much information you want to give out
9. Planning what happens next once you have shared or released the data
- Data environment is likely to change (data is used in new ways), impacting the risk level
10. A plan of what you will do if things go wrong
- Risk of disclosure is not zero, therefore it should be negligible
    - Developing a breaches policy, identifying crucial factors e.g. malicious intrusion, who is responsible for doing what, when and how

2.2 De-anonymization

A process to in which anonymous data is cross-referenced with other sources of data to re-identify the anonymous data source.

2.2.1 Deanonymisation in Linked Data: A Research Roadmap

(Al-Azizy, Millard, Shadbolt, & O'Hara, 2014) develop a roadmap to highlight and identify the state-of-the-art privacy challenges and de-anonymization techniques in Linked Data. Coinciding with publication of data on the Web, as well as other data linking to it. Whilst discussing these techniques and the precautions that need to be taken into account for keeping an open point of view towards the leak of coverage in Linked Data in the Web, at the same time preserving privacy for when it is required. Particularly with focus on the collection of data from Online Social Networks (OSNs), untrusted applications that is beneficial for third party usage. Raising privacy concerns about personal information on social platforms and how easy it is to detect government data linkage with an individual's private information. Altogether, acknowledging the privacy challenges and risks that come with publishing data on the web in order to satisfy individuals' interest of privacy rights against the political and business interests that tend to dominate decision-making when publishing data on the web.

Dataset:
Online social networks (OSNs) consisting of personal and private data

Technique:
De-anonymization is popular in the context of OSNs, research indicating that 75% of studies found literature for analysis falling under this category.

Proposal of de-anonymization attack methods:

- **Data matching**
  - Exploitation of network structure and utilising nodes, edges and social relationships for matching values to identify users, account details and other attributes
  - **Co-reference resolution**
  - Between overlapping datasets, checking different URIs are the same by referring to the same entity.

- **Linkage attack**
  - Practical attack, where the adversary can link auxiliary information of certain users with anonymized records in a dataset.
  - **Linkability**
  - Connecting entities that are linked, that should be anonymized. Inference attacks use observed links or attributes of users'. E.g. link two accounts of the same user from different social network platforms and distinguish it that way.

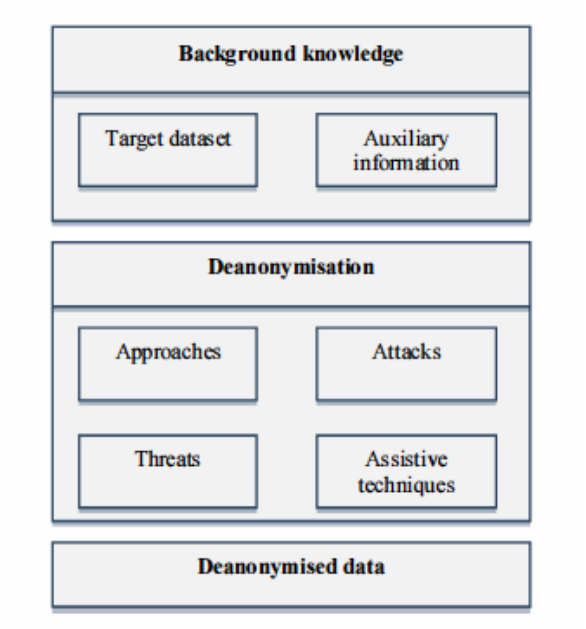


Figure 2 - De-anonymization Framework

(**Fig.2**) illustrates the de-anonymization framework, First, background knowledge is taken into consideration, consisting of target datasets and auxiliary information. Second, the de-anonymization process begins, including approaches, attacks, threats, assistive techniques and ends with the data that was once anonymized, becoming deanonymized.

The de-anonymization approach is classified based on the methodology that the adversary tends to use when exploiting the structure of the target dataset in conjunction with the auxiliary information.

As mentioned before Data matching is a traditional method used for the de-anonymization of data, there are also other attacks that have not been discussed yet. Such as:

- **Graph Matching**
  - Commonly used approach due to the web structure being a global scaled graph. Hence, acquiring the graph structure is a simple task and makes it possible to exploit in order to achieve de-anonymization attacks successfully.
- **Seed & Grow and Threading**
  - Two major approaches to begin a complex graph matching activity.

There are other matching features that consist of similarity and statistics, with the likes of classifiers such as link, group, similarity and sparsity and trail-based approaches that deal with de-anonymization within different contexts.

2.2.2 De-Anonymization of Dynamic Social Networks

(Ding, Zhang, & Wan, 2013) consider dynamic socials networks and particularly pay close attention to the re-publication of the dynamic data that is released from those platforms. As the evolution of social networks is continuously growing, there emerges a requirement for releasing data on a periodic basis, at the same time this raises a number of privacy concerns that can lead to de-anonymization of sensitive information regarding individuals.

Technique:
Re-identification can occur through the potential of users' identities becoming exposed via an adversary executing an attack with the use of background knowledge, then mapping these known individuals with anonymized nodes that are present in the released network. Subsequently, there are at times an external network called an auxiliary network that tends to overlap with the released network and from this re-identification of nodes can transpire, leading to gaining important information on individuals (e.g. attributes, relationships). Re-publishing could convey more of a risk towards adversaries as it could provide a more likely breach by them to a user's privacy. (By re-identifying each release separately and add up the results)

One main issue that lies with this type of technique is the lack of ground truth that is available upon the attempt of carrying out such a task. Reliability of the data is reduced by a significant amount, due to true mapping of the knowledge that the adversary holds and that from the released network on individuals (result of re-identification lowers considerably). If an adversary decides to synthesize results that contain false mappings, this will portray nodes as representing different individuals that may cause confusion as nodes for one individual may be different or the same. A limitation of this existing attack, is that it only exploits structural knowledge for re-identification resulting in new defences being introduced to protect anonymized nodes from different types of structural knowledge.

(Ding, Zhang, & Wan, 2013) propose threading to deal with this issue, which utilizes correlations between these releases. De-anonymization instead occurs in a simultaneous manner of all releases to ensure there is no ambiguity throughout the whole process, aiding in the ground truth becoming more accurate. Also introduces node attributes to loosen the re-identification process.

Seed threading is introduced to ensure the robustness of this de-anonymization process, with the algorithm reaching a threading expansion stage, where each thread is iteratively extended to its neighbourhood to a larger threading through the auxiliary and target network (followed by a matching score computation called **BestMatch**, this key for re-identification to succeed.

Dataset:
To evaluate the algorithm, data crawled from Netease Microblog, which is one of the most popular social networking sites in China was used (**Appendix D - I**).

Results:
(**Appendix D - II**) – Separate vs Threading-based attacks. De-anonymizing releases separately and then synthesizing results after does tend to have a decrease in reliability of the results notably.

Threading-based technique can provide results of a much higher precision. (**Appendix D - III**) depicts that the algorithm produces better threading when the releases are closer to each other in time. (line "1,2,3"). particularly when close to the background knowledge. (line 1,3,6 vs 1,4,5). The algorithm produces a result of higher precision when the seed network size is large.

(**Appendix D - IV**) shows the effect of node attributes to re-identification - by comparing a zero and non-zero value of a variable in matching score computation (bijection of nodes).

2.2.3 De-anonymizing Social Networks (Narayanan & Shmatikov)

Technique:
A very prominent and passively tough de-anonymization process of real world social networks on a large scaled, based purely on the network topology, without the use of a big number of Sybil nodes, that is robust to noise and all defences that currently exist and working to an extent even when the overlap between the target network and adversary's auxiliary information is miniscule. This paper's main aim is to call for a substantial re-evaluation of business practices that surround the sharing of social network data.

The algorithm is broken down into 3 main techniques:

- **Seed identification**
  - Key step in assisting the overall algorithm to succeed
  - Searches the target graph for unique node degree match and common neighbour count, if successful the algorithm then maps nodes to the matching nodes in the auxiliary graph (successful), else resulting in failure.
  - Running time can be greatly decreased as soon as one match is found, due to termination of the algorithm.

- **Propagation**
  - Considering factors such as eccentricity, edge directionality, node degrees, revisiting nodes, reverse match and complexity
  - Takes input as: two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ and a partial "seed" mapping between the two.
  - At each iteration, the algorithm begins with the accumulated list of mapped pairs between V_1 and V_2. It picks a random unmapped node in V_1 and computes a score for each unmapped node in V_2, equal to the number of neighbors of nodes from V_1 that have been mapped to its neighbors of nodes from V_2.

- **Mapping between Flickr and Twitter**

- Ground truth – mapping between the two graphs (exact matches based on username or name fields)
- Compute scored based on three fields (username, name and location), rejected if score is not above threshold.

Dataset:
Data crawled from Flickr and Twitter (**Appendix E - I**)

Results:
- Success rate of mappings for re-identification: 30.8%
- Incorrectly identified: 12.1%
- Not identified: 57%

(**Appendix E - II**) Shows a decrease in the percentage of nodes re-identified due to the imprecision of the auxiliary information, however, illustrates that it doesn't prevent large-scale re-identification from occurring.

With social networks growing larger by the day, with a huge population and their relationships, overlap is ever increasing and this entails that this algorithm is able to re-identify users on even larger networks in the years to come.

- 24% of names associated with accounts from Twitter occur in Flickr, whilst only 5% of names associated with Flickr accounts occur in Twitter.
- 64% of Facebook users are also present on MySpace.

A lot of detailed methods and techniques used in the approach towards achieving de-anonymization and resulting in the re-identification stature of sensitive and personally identifiable information of users (Twitter and Flickr comparison in this paper for their experiment on re-identification of users). The risk towards privacy is ever growing and is a topic that is not to be misconstrued as anonymity is clearly not sufficient enough to protect people's information from the likes of an attacker.

This paper showcases an algorithm that signifies the real importance of how anonymity should be taken serious, but above all, the level and severity as to how the term privacy should be considered, not only with dealing with people's sensitive data, but also in terms of business and for research purposes.

2.2.4 Robust De-anonymization of Large Sparse Datasets

(Narayanan & Shmatikov, Robust De-anonymization of Large Sparse Datasets) present a statistical de-anonymization attack upon high-dimensional micro data, the technique is robust to perturbation and allows for some mistakes to be made from the adversary's background knowledge. The model focuses on a much wider class of de-anonymization attacks rather basic cross-database correlation. Residing amongst two aspects that the algorithm is formed upon:

- Successful rate based upon the posterior probability of de-anonymization
- Amount of information recovered about the target

Technique:

Sparse datasets increase the probability for de-anonymization succeeding. (decreases amount of auxiliary information that is needed and improves the robustness to both perturbation in the data and any inconsistencies in the auxiliary information).

An improvement is made from the Algorithm **Scoreboard** which is not robust for a number of applications, as it fails if any of the attributes in the adversary's auxiliary information are incorrect, thus an improved version is proposed **Scoreboard-RH** that deals with what the previous algorithm can't.

Consisting of enhancements to the three main components of the algorithm:

- **Scoring function**
  - Assigns a numerical score to each record in the database based on how well it matches the adversary's auxiliary information (***Aux***).
  - Gives higher weight to rare attributes (rare attribute is more helpful than identifying a common attribute, e.g. name of a specific book rather than what novel it's from)
- **Matching criterion**
  - Algorithm applied by the adversary to the set of scores to determine if there is a match.
  - Improve robustness by having a top score that is significantly above the second best-score (to be able to identify which records stand out).
- **Record selection**
  - Selects one "best-guess" record or a probability distribution, if necessary.

The only two possible ways of the algorithm failing are by an incorrect record is matched as the highest score and the correct record may not have significantly higher score than the second best-match.

Dataset:

Netflix Prize dataset consisting of 500,000 records.

Results:

After applying the Algorithm **Scoreboard-RH**, a measure of similarity (**Sim)** which has a threshold function upon attributes (returning 1 only if the two attributes are within a certain threshold of one another).

For movie ratings that are scaled 1 - 5 on Netflix, they consider thresholds of 0 (indicating to an exact match) and 1. (Same goes for the rating dates of 3, 14 and infinite number of days)

(**Appendix F - I**) - shows that the adversary knows exact ratings and approximate dates. Even with just 2 movie records in auxiliary information, the adversary is able to de-anonymize at a minimal probabilistic rate of 0.4 and above.

(**Appendix F - II**) - Detects number of movies from sample that can be de-anonymized but also indicates the probability that the target record is not in the sample.

Even if the adversary has less auxiliary information, there's still a lot of extra information that is present within the dataset that can be used to eliminate false positives. (if the start date and total number of movies in a record are part of the auxiliary information, then the adversary knows enough information (e.g. when the target record first joined Netflix) that can then be used to eliminate candidate records.

99% of records can be uniquely identified within the dataset. With a 3-day error rate for two ratings and dates, 68% can be identified and the other 32% holds a list of possible candidates that has decreased considerably.

Even with a 50% error rate of knowledge about the candidate's record, the probability of de-anonymizing with auxiliary information doubles than before.

Proposal by Matthew Wright

- To release the records without column identifiers (i.e. movie names in the case of Netflix dataset). Although it is not clear how much worse current data mining algorithms would work under this restricted behaviour. It does not mean that de-anonymization is impossible in this manner.

2.2.5 Content-Based De-Anonymization of Tweets

(Okuno, Ichino, Kuboyama, & Yoshiura, 2011) discuss the different types of personal information that is easily accessible on the Web and the ways in which these can be retrieved. Highlighting the fundamental aspects that are needed to violate privacy and implement a method that can match individuals from one source of information to a post that may be on a social media platform, in this case it is information portrayed on a Resume (CV) and tweets by a user on Twitter.

Technique:
A de-anonymization approach known as the **Vector Space Model** (information retrieval method) was applied to compute the similarity between two documents, a tweet and a resume. The way this process would be successful, is through the author of the resume and author of the tweet sharing a high similarity result. The documents are represented as vectors, where Term Frequency (TF) is used for calculating the vector weight, measuring how frequently a term occurs in a document.

$$TF_j = \frac{n_j}{\sum_k n_k},$$

$\boldsymbol{TF_j}$ being number of times term j appears in a document/total number of terms in the document ($\mathbf{n_j}$ number of times that the term j appears in the document).

$$sim(d_1, d_2) = \cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|} = \frac{\sum_{j=1}^{m} d_{1j} d_{2j}}{\sqrt{\sum_{j=1}^{m} d_{1j}^2} \sqrt{\sum_{j=1}^{m} d_{2j}^2}}$$

Similarity(***sim***) is calculated using the cosine to compare the deviation in angles between the two vectors (similarity between $\mathbf{d_1}$ and $\mathbf{d_2}$ is calculated as: where $\mathbf{d_{1j}}$ and $\mathbf{d_{2j}}$ are weights of term **j** in $\mathbf{d_1}$ and $\mathbf{d_2}$.

$$IDF_j = \log \frac{N}{N_j},$$

IDF is used as the weight to consider the general importance of the term. $\boldsymbol{N}$ is the total number of documents in the document set and $\boldsymbol{N_j}$ is the number of documents that contain the term **j**.

$$x_{ij} = TF_{ij} \cdot IDF_j,$$

TFIDF - frequency and importance of a term in a document.

Weight of $\boldsymbol{X_{ij}}$ of term **j** in document **i** is: $\boldsymbol{X_{ij}} = \boldsymbol{TFij . IDF_j}$
$\boldsymbol{TF_{ij}}$ - is term frequency of a document and terms within it.

Dataset:

400 tweets collected from 5 different users and resumes (CVs) of two of them (Adam and Bob) (**Appendix G – VII**)

Results:

(**Appendix G - I**) shows the similarity between Adam's documents and resume was not the highest, whereas (**Appendix G - II**) shows that similarity for Bob's was the highest for most documents. Thus the Vector Space Model entailing that it does not always give best results with highest correlation between a person's tweet and resume, even if that person is the same author of both (due to the words used by someone writing a resume can differ from those that the same someone uses in their tweets e.g. short abbreviation or the full grammatical terminology for professionalism).

An improved scoring method for correlation was then designed to detect the unwritten words and to calculate their correlation to the original word (by the algorithm returning a value of 1.0 when an original word in the resume is found in a document). Automatically quantifying the correlation between the original word and the document, returning a value in the range of 0.0 – 1.0 (even if the original word is not matched and found). Thus displaying some sort of similarity. This algorithm represents a model in which a person finds a correlation between a set of words and documents with the use of a web search engine (keywords or key phrases are selected from the nouns that are morphologically extracted from the document, then a search engine is used for each set of selected keywords).

(**Appendix G - III**) process for the top *n* retrieved items to determine the score of a word. Through the use of this scoring system, they were able to estimate in a resume "University of Electro-Communications" and in a tweet "open campus and briefing session on entrance exam for graduate school will be held on May 22" to have some exposure of correlation as the score produced 0.14. (**Appendix G - IV**) - Improved scoring method used as the term weight in the vector space model in order to reflect the unwritten words to the measure of similarity. X_j Document and weight of term, **1/n** (number of tweets), sum of scoring of term **j** in tweet, **.** IDF_j total number of documents/ number of documents with term **j** in it. (**Appendix G - V**) shows much higher correlation of results for similarities of comparison between all documents and resume(cv) for Adam and same result for Bob (**Appendix G - VI**).

Expressing that there is no need for auxiliary information for selecting words and phrases from the resume as the vector contains all the words and phrases in the resume apart from those used most often in a resume (due to IDF and normalization of the words) rare words and those of importance are recognized and kept whereas other words that have been used many times have little importance (there is a need for weighing down the frequent terms while the need is really for scaling up the rare ones).

This papers results showcase an experiment that was able to match information from a resume, through possible short anonymous messages that showed that the author of that message could be de-anonymized through specific measures and techniques applied.

3 Anonymization Vs De-Anonymization

Anonymization	***Strengths***	***Weaknesses***
K-anonymity	• Allowing records to be indistinguishable when released as a set (by releasing multiple records with the same or similar attributes).	• If the adversary has substantial knowledge already on an individual, they are more prone to determining a record belonging to that specific individual.
l-diversity	• Preserves and gains privacy in datasets by reducing the granularity of the data representation.	• Lose effectiveness of the data and the purpose for which it were intended (e.g. for data mining purposes).
PII Encryption	• Encrypting personally identifiable chosen fields and data types with strong hash such as SHA-1. • Ensures integrity of the record is kept intact.	• Decryption can occur if an adversary has the correct technique (such as the correct decryption(public/private) key). • Background knowledge can aid the adversary in a successful attack.
De-anonymization	***Strengths***	***Weaknesses***
Vector space model	• Detect similarity and compare two or more documents.	• Doesn't always give the highest correlation between two documents (e.g. words may not be written the same in both documents).
Seed identification & Threading	• Strong correlation of detecting sensitive information if ground truth is correctly carried out and mapping is accurate.	• If ground truth is not accurate or anonymized nodes in a network are not identified correctly, then it causes false mapping.
Active attacks	• Very effective, work with high probability in any network.	• Risk of being detected.
Passive attacks	• Undetectable. • Victims are only those linked to the attackers.	• Attackers may not be able to identify themselves after seeing the released anonymized network.

4 Conclusion

Anonymity has become an important measure that is needed for all types of users on the Web, as well as large organisations that want to keep their data secure. There have been many different proposals towards anonymizing data and keeping it securely away from the possession of unauthorised personnel (attackers), however, there is an imbalance of the amount of privacy (data hidden and secure) one receives and amount of utility (completeness of the data) that remains once a tradeoff has been completed.

Adversary's that perform attacks against these enforced anonymization techniques tend to find different de-anonymization methods of their own to breach the privacy that they intended for from the beginning, such as sensitive attributes (name, age, address) or quasi-identifiers (city, education, grades) that can lead to gaining the information that they require. Although, there are many robust anonymization techniques implemented on several platforms and systems, this does not entail that it is completely impenetrable from an adversary that may hold certain background knowledge or the key to re-identification.

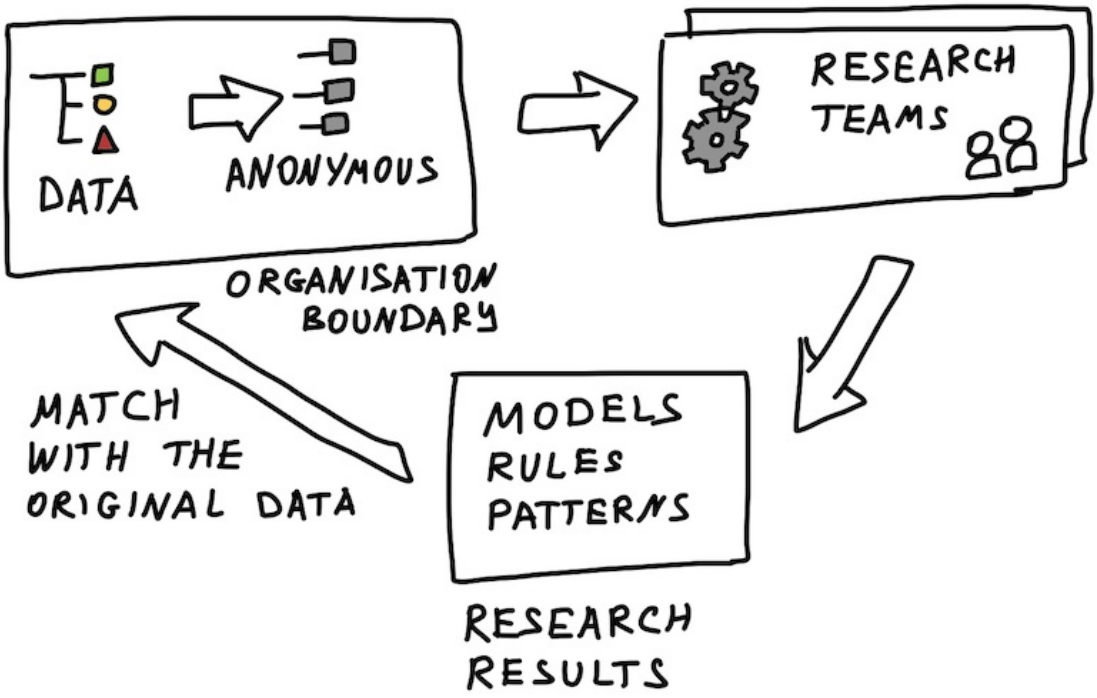
In terms of dealing with datasets and the process of implementing various techniques, findings have identified that Python is the most suitable programming language to deal with such techniques and data. Although C++ may be quicker in processing, it does not hold pre-defined libraries that are required for dealing with the process of large sets of data.

Overall, there is clear indication and sufficient evidence provided throughout this paper that highlights the ways and reasons as to why data for whatever the purpose may be, will never have a 100% success rate of being anonymized. As circumstances change and dynamic data in particular is never kept in the same state, there is always a need for regular maintenance and updates of ensuring that data is kept the way it is wanted and not only consisting of a one-time fix, as this will not benefit a user anonymizing data, nor an adversary who is attempting to de-anonymize and extract information from anonymized data.

5 Future Research

There have been many existing and newly proposed techniques for both anonymization and de-anonymization illustrated throughout this paper. However, there has not been a technique yet published that is significantly different than others already used by users or businesses. There is a need for continuous research as platforms, specifically social networks such as Facebook, Twitter as well as other applications are dynamically evolving and this poses a grave threat towards the privacy of users. As more and more users are joining these platforms, it makes anonymity an increasingly difficult topic to consume and secure in the most appropriate way possible.

It is certain that more research is needed to keep up with the latest online applications, the threats that are present as well as vulnerabilities that may exist. Subsequently, considering the different aspects that exist upon the term "data" as this holds more meaningful information towards decision-making that has not yet been fulfilled to the highest capability that it could of till present.



6 Acknowledgements

I'd like to thank the association of the University of Southampton and GCHQ for providing me this opportunity to undertake the research and thank my supervisors Vladimiro Sassone (University of Southampton) and Paul Brittan (Northrop Grumman) for all of their support.

7 References

Agrawal, R., & Ramakrishnan, S. (2000). Privacy-preserving data mining . *Conference on Management of Data* , 439 - 450.

Al-Azizy, D., Millard, D., Shadbolt, N., & O'Hara, K. (2014). Deanonyimsation in Linked Data: A Research Roadmap. *World Congress on Internet Security (WorldCIS-2014)*.

Ateniese, G., Burns, R., Curtmola, R. G., Herring, J., Kissner, L., Peterson, Z., & Song, D. (2007). Provable Data Possession at Untrusted Stores*. *Proceedings of the 14th ACM Conference on Computer and Communications Security*.

Buratoviç, I., Miliçeviç, M., & Zubrinic, K. (2012). Effects of Data Anonymization on the Data Mining Results.

Ding, X., Zhang, L., & Wan, Z. &. (2013). De-Anonymization of Dynamic Social Networks . *Information Technology Journal 12 (19)* .

Elliot, M., Mackey, E., O'Hara, K., & Tudor, C. (2016). *The Anonymisation Decision-Making Framework*.

George, S. R., & S, S. (2013, July 4 - 6). Data Anonymization and Integrity Checking in Cloud Computing. *Department of Computer Science and Engineering, College of Engineering*.

Gorade, N. (2014). Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues . *IEEE International Congress on Big Data* , 601 - 607.

Lane, A. (2012, January). *Tokenization Guidance*. Retrieved from https://securosis.com/assets/library/presentations/TokenizationGuidanceAnalysis_Jan_2012.pdf

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-Diversity: Privacy Beyond k-Anonymity. *Cornell University* .

Narayanan, A., & Shmatikov, V. (n.d.). De-anonymizing Social Networks .

Narayanan, A., & Shmatikov, V. (n.d.). Robust De-anonymization of Large Sparse Datasets.

Ohm, P. (2010). Broken promises of privacy: Repsonding to the surprising failure of anonymization. *UCLA Law Review, 57*, 1701.

Okuno, T., Ichino, M., Kuboyama, T., & Yoshiura, H. (2011). Content-Based De-Anonymization of Tweets. 53 - 56 .

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information . *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 188.

Sedayao, J., & Bhardwaj, R. (2014). Making Big Data, Privacy, and Anonymization work together in the Enterprise: Experiences and Issues. *IEEE International Congress on Big Data*, 601 - 607.

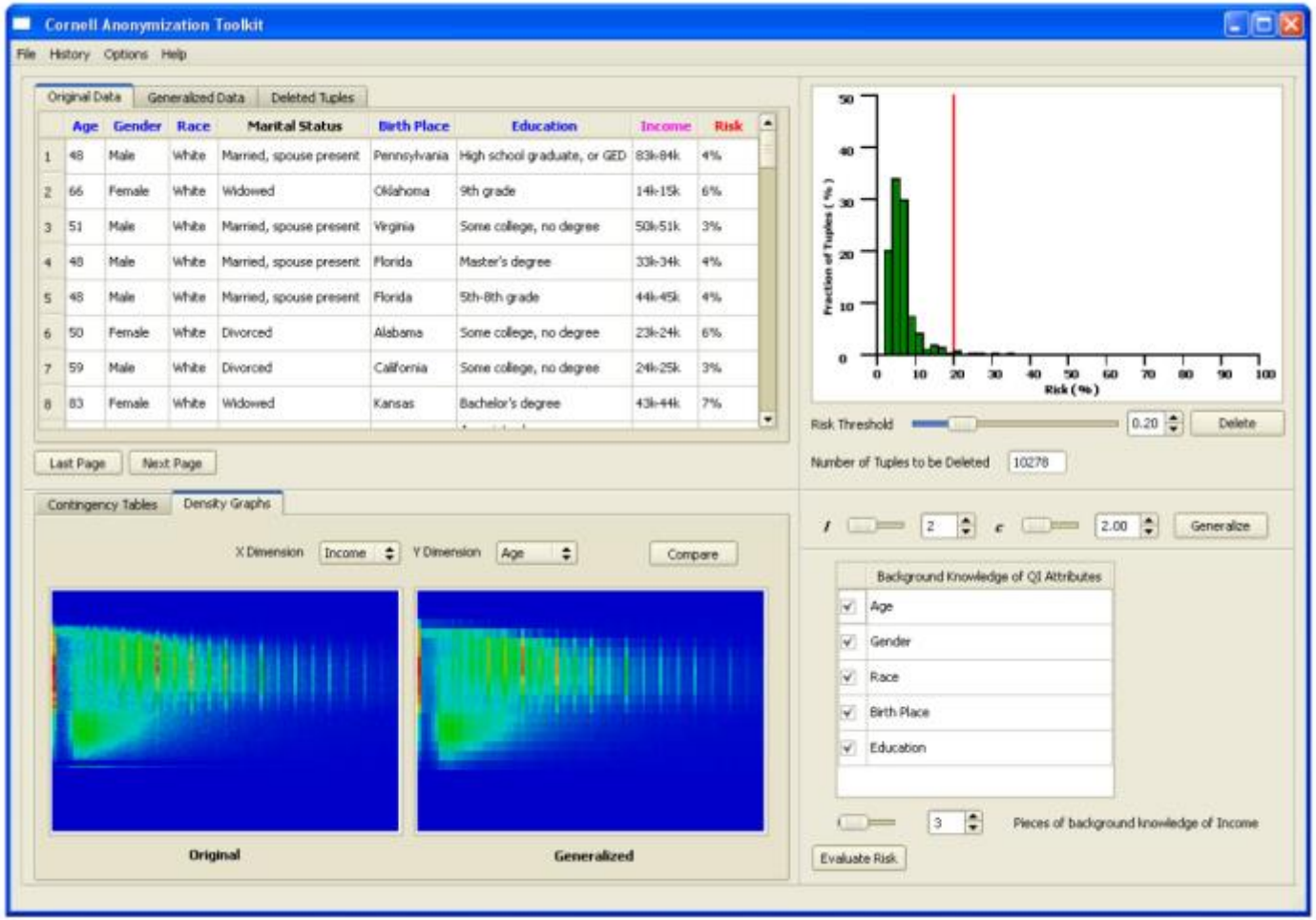
Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 557 - 570.

Xiao, X., Wang, G., & Gehrke, J. (2009). *Interactive Anonymization of Sensitive Data* , 1051 - 1053.

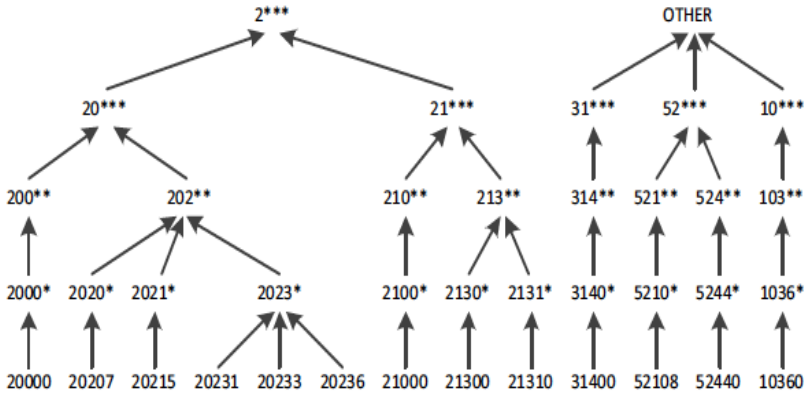
8 Appendices

	male	female	total
married	284421	48590	333011
divorced	37453	56581	94034
widowed	13546	61549	75095
single	48105	49755	97860
total	383525	216475	600000

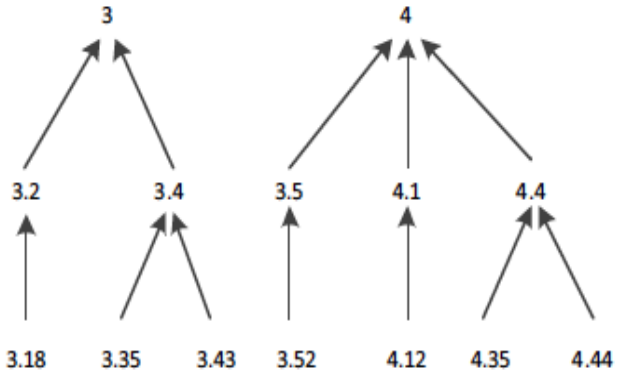
Appendix A – I (Contingency table - Page 6)



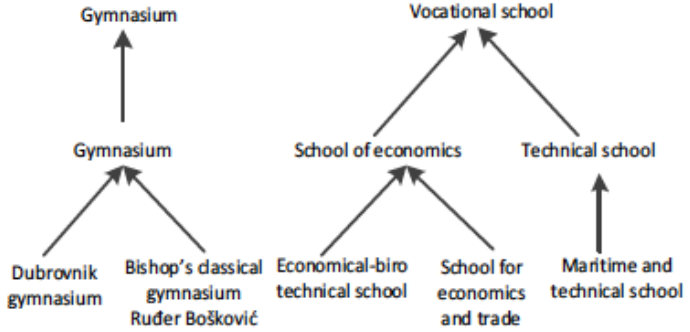
Appendix A – II (User interface – Page 6)



Appendix B – I (Postal code Generalization Hierarchy – Page 7)



Appendix B – II (Students Grades Generalization Hierarchy – Page 7)



Appendix B – III (Secondary Schools Generalization Hierarchy – Page 7)

0.1011765	2 Department
0.092549	1 Undergr_Study
0.0729412	18 Enroll_Date
0.05435	11 Sec_School_GPA
0.0478431	8 Profession
0.0321569	4 Sec_School
....	

Appendix B – IV (Average Merit Scores for Attributes – Page 6)

2012-09-30 08:20:17 10.255.146.153 jcsedaya employeeportal.intel.com/irj/portal/PurchasingServices GET / DCS.dcsqry=NULL&WT.tz=7&WT.bh=15&WT.ul=en-us&WT.ti=Purchasing%20Services&WT.auth=jcsedaya&WT.authname=jcsedaya&WT.js=Yes&DCSext.authname=jcsedaya&DCSext.tabid=QL&DCSext.campuscode=IHO&WT.co=Yes 200 - - Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.1;+WOW64;+Trident/4.0;+SLCC2;+.NET+CLR+2.0.50727;+.NET+CLR+3.5.30729;+.NET+CLR+3.0.30729;+.NET+CLR+1.1.4322;+MS-RTC+LM+8;+InfoPath.3;+AskTbSPC2/5.12.2.16749;+.NET4.0E;+.NET4.0C;+Zune+4.7) https://employeecontent.intel.com/EntryPage/Default.aspx?nodeid=54ceab56-c3b6-44c7-ae3f-08fc87af9045

Appendix C – I (Circuit Log Entry – Page 9)

2012-09-30 08:20:17 **%%70ddf03121acb3b20f581604fe2100a3e87304ee38028a11d76447075890acbc%%%%4cf838ef66d75c67618b6dda35d561aa3cad1b180cadb96af00d8d4f3faea658%%** employeeportal.intel.com/irj/portal/PurchasingServices GET / DCS.dcsqry=NULL&WT.tz=7&WT.bh=15&WT.ul=en-us&WT.ti=Purchasing%20Services&WT.auth=**%%2%%**&WT.authname=**%%2%%**&WT.js=Yes&DCSext.authname=**%%2%%**&DCSext.tabid=QL&DCSext.campuscode=IHO&WT.co=Yes 200 - - Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.1;+WOW64;+Trident/4.0;+SLCC2;+.NET+CLR+2.0.50727;+.NET+CLR+3.5.30729;+.NET+CLR+3.0.30729;+.NET+CLR+1.1.4322;+MS-RTC+LM+8;+InfoPath.3;+AskTbSPC2/5.12.2.16749;+.NET4.0E;+.NET4.0C;+Zune+4.7) https://employeecontent.intel.com/EntryPage/Default.aspx?nodeid=54ceab56-c3b6-44c7-ae3f-08fc87af9045

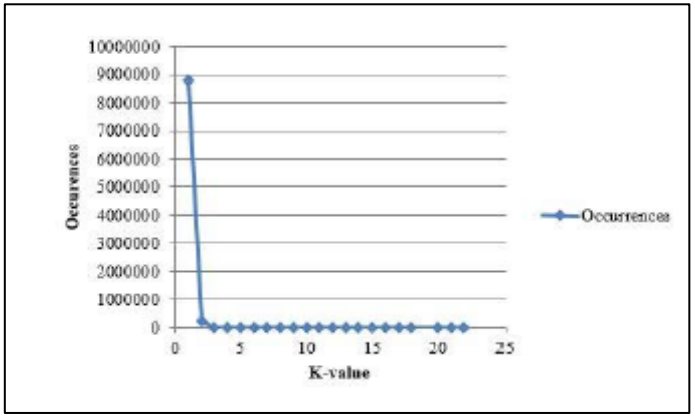
Appendix C – II (Anonymized Version of Circuit Log Entry – Page 9)

Use Case	Description
Aggregate page hits	Per user, how often do they hit the circuit website as a whole? Need to identify users and aggregate by Calendar Month
Session Time	What is the aggregate time per session by users?
Aggregate Search vs. Browse	Aggregate hits, by user per month, of search (identified by iSearch in the string) vs. browsed page hits
Search terms	What search terms were used by users in a given month?
Browse vs. Search in a session	How do users leverage circuit? Within a session, how do users navigate the environment? Do they browse for items of interest before searching? Do they directly search and browse through topics?
Search Efficiency	How different searches do users enter in a session and what is the variation between searches?
Referrer Pages	Which pages are referred by what page? Can we see how users get to different information and where they are coming from to get there?
User Demographics	Who are the frequent users? Where are they from? What business group? - Test merging CDIS data with Circuit output
Peak Usage Times	What are the peak use times by users by region? Aggregate at the hour using system time
Anonymous Circuit and Anonymous Circuit CDIS	Can we bring 2 anonymized data sources together and get joined information while the data is anonymized? – Can we still use simple key encryption or is tokenized a requirement?

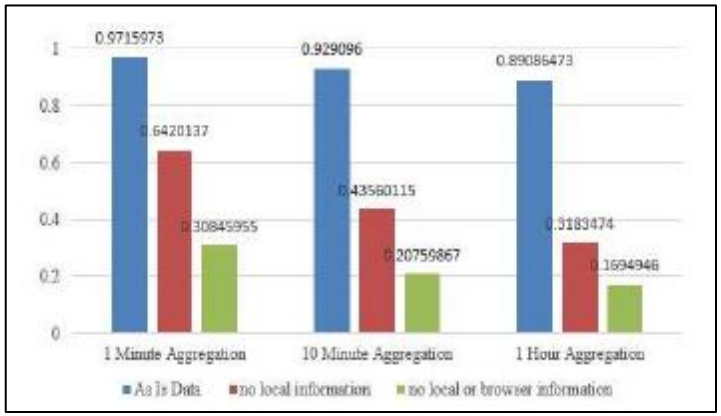
Appendix C – III (Use Case for Study of Circuit Log – Page 9)

Search Term	Occurrences
Recognition	2308
Plts	1294
Flu shots	1120
Mft	892
Recognition awards	861
PLTS	824
MFT	818
Health for Life	783

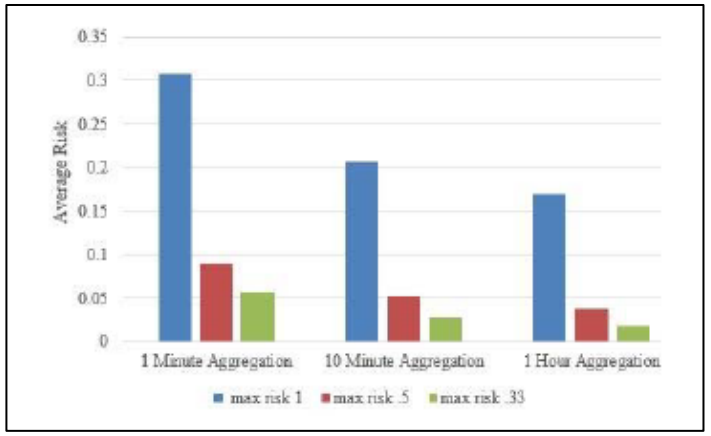
Appendix C – IV (Top Searches on Circuit – Page 9)



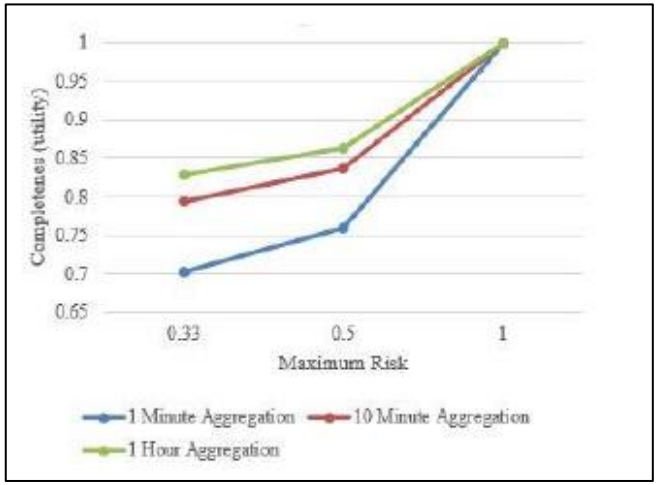
Appendix C – V (Occurrences of k-anonymity levels – Page 10)



Appendix C – VI (Effects of Removing Data on Average Risk – Page 10)



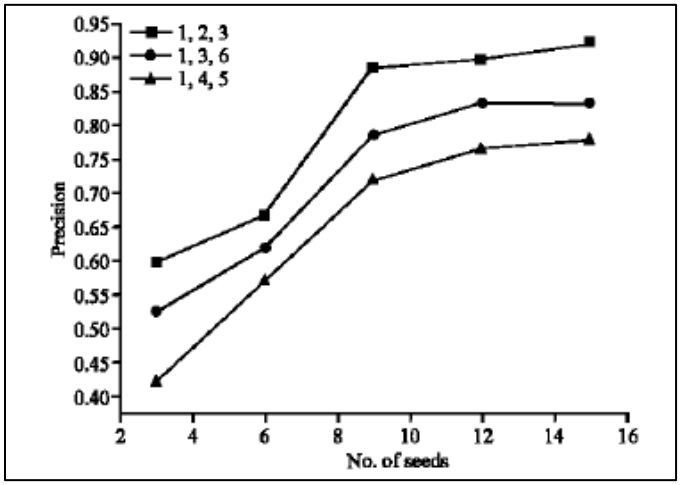
Appendix C – VII (Improving Average Risk & Maximum Risk – Page 10)



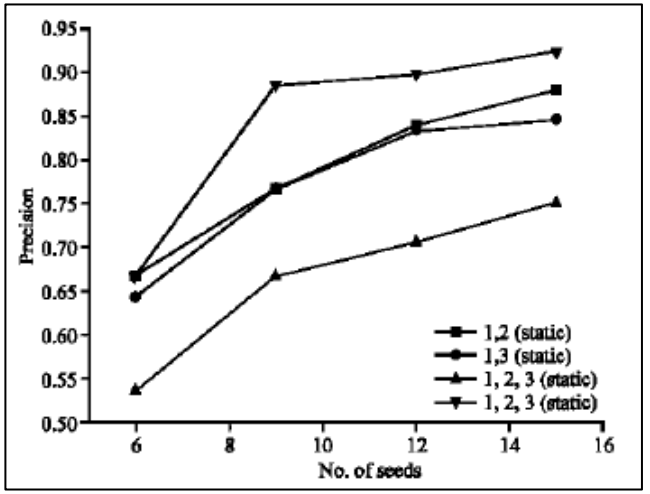
Appendix C – VIII (Tradeoffs of Completeness Vs Maximum Risk – Page 10)

No.	Date	Nodes	Edges	Av. deg
S1	2010-03-27	47,367	784,508	33.1
S2	2010-03-30	54,190	858,244	31.7
S3	2010-04-02	64,002	977,889	30.6
S4	2010-04-07	117,864	1,478,188	25.1
S5	2010-04-11	161,267	1,894,840	23.5
S6	2010-04-16	204,790	2,308,505	22.5

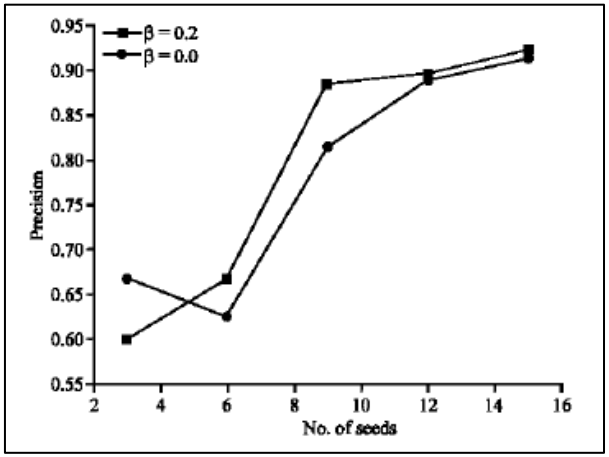
Appendix D – I (Data crawled from Netease Microblog – Page 16)



Appendix D – III (Re-identification of Different Release Combinations – Page 17)



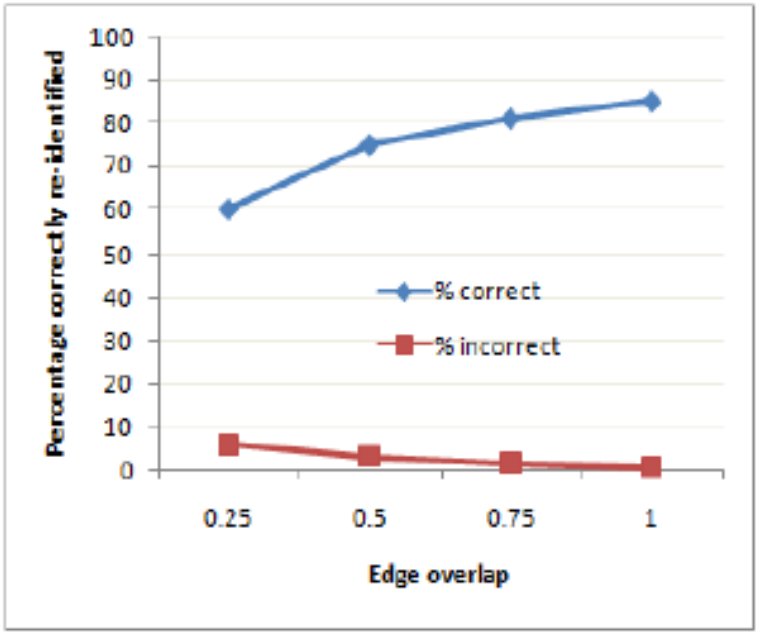
Appendix D – II (Separate vs Threading-based attacks – Page 17)



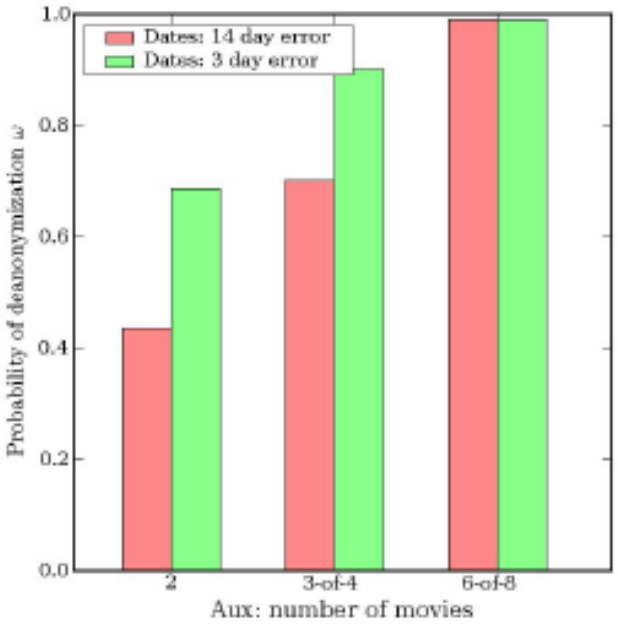
Appendix D – IV (Effect of Node Attributes – Page 17)

Network	Nodes	Edges	Av. Deg
Twitter	224K	8.5M	37.7
Flickr	3.3M	53M	32.2
LiveJournal	5.3M	77M	29.3

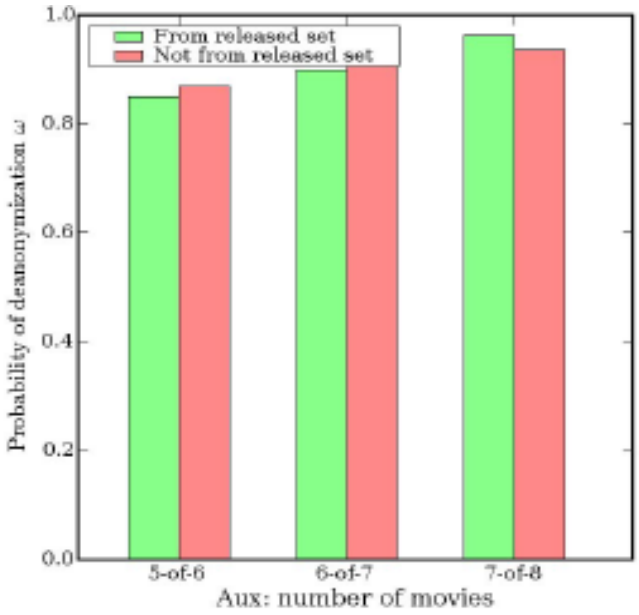
Appendix E – I (Data Crawled from Flickr & Twitter – Page 18)



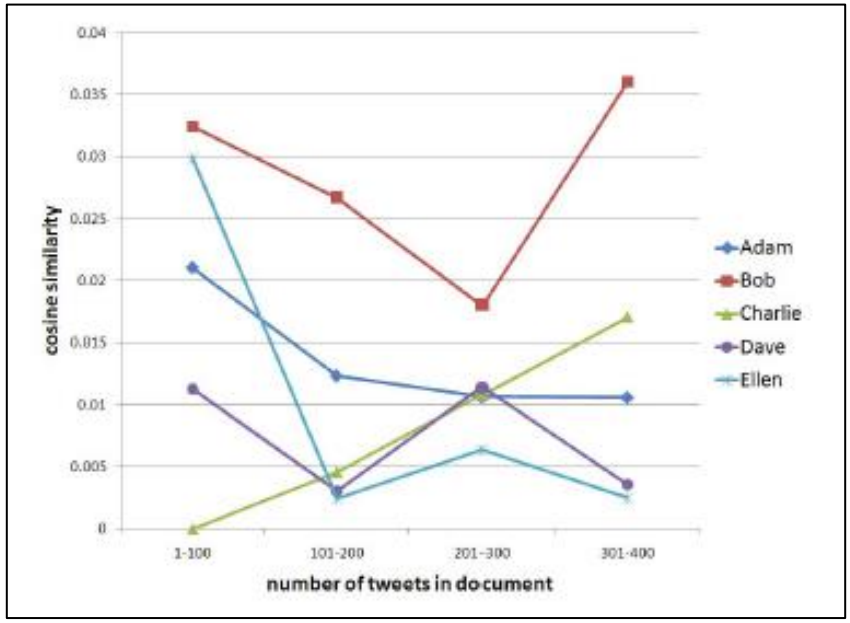
Appendix E – II (Decrease of Nodes Re-identified – Page 18)



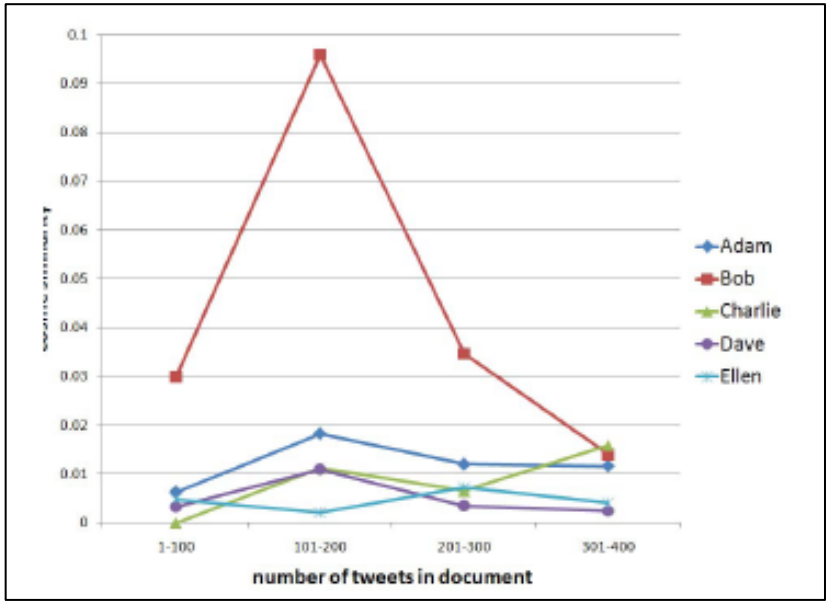
Appendix F – I (Exact Ratings & Approximate Dates – Page 20)



Appendix F – II (Detecting When Target Record is Not in the Sample – Page 20)



Appendix G – I (Similarity between Adam's resume & all documents – Page 22)



Appendix G – II (Similarity between Bob's resume & all documents – Page 22)

selected keywords, and the following process is executed for the top n retrieved items.

1) Determine if the title of the result contains the original word.

2) If the original word is found, calculate the score on the basis of ranking of the retrievals.

$$\text{score}(i) = \begin{cases} n - i + 1, & \text{if the original word is found} \\ & \text{in the title of } i^{th} \text{ retrieval} \\ 0, & \text{else} \end{cases} \quad (6)$$

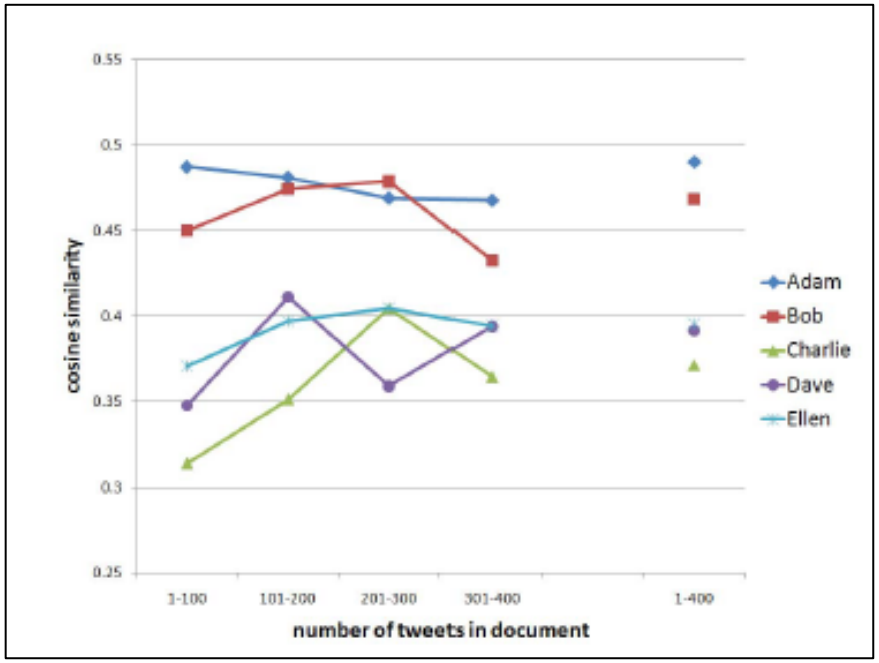
3) Calculate the summation of scores and normalize the summation by dividing it by the maximum score:

$$normalized\_score = \frac{\sum_{i=1}^{n} score(i)}{\sum_{i=1}^{n} (n - i + 1)}. \quad (7)$$

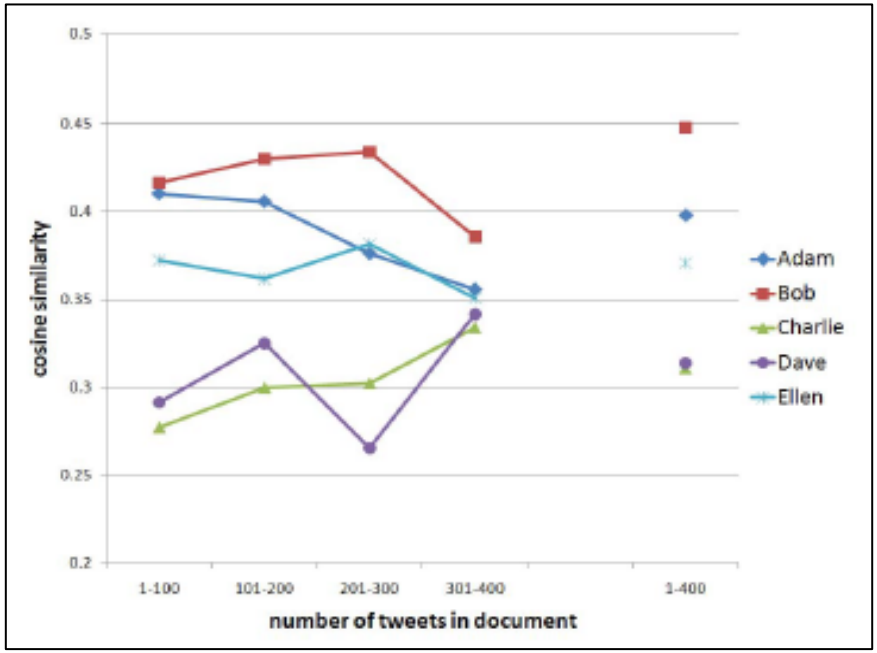
Appendix G – III (Algorithm for Determining Score of a Word – Page 22)

$$x_j = \frac{1}{n} \sum_{i=1}^{n} s_{ij} \cdot IDF_j,$$

Appendix G – IV (Improved Scoring Method – Page 22)



Appendix G – V (Adam's resume and all documents with improved method – Page 22)



Appendix G – VI (Bob's resume and all documents - Page 22)

User's Name	University	Place of Work	Area of Expertise
Adam	UEC	NTT	Network security
Bob	Waseda University	NTT → UEC	Networks and multimedia
Charlie	UEC	(student)	Media security
Dave	(unknown)	Computer game maker	Game creator
Ellen	(unknown)	(unknown)	Singer

Appendix G – VII (User Attributes – Page 22)