Data Quality Measures for Identity Resolution

Matthew John Edwards, BSc. (Hons)

School of Computing and Communications

Lancaster University

This thesis is submitted for the degree of

Doctor of Philosophy

April 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This thesis contains fewer than 80,000 words.

Matthew John Edwards, BSc. (Hons)

April 2018

Acknowledgements

This thesis includes work carried out and published in collaboration with others, though in each case I was the principal author.

Throughout the development and execution of the body of research which forms this thesis, I have been supported and advised by Dr. Paul Rayson and Prof. Awais Rashid, to both of whom I owe a great debt of gratitude for their patience, insight and motivation. I can specifically identify them as co-authors acting in this role in [68–70], and thank them for their many helpful comments on this thesis.

The social engineering vulnerability detection project [67] was a project instigated by Dr. Alistair Baron, with myself as the main author, and the research, design and evaluation of that project were carried out as a joint effort between myself and others. Robert Larson and Benjamin Green carried out and reported on interviews with social engineering experts, which data I make use of with attribution in this thesis. I would like to thank Dr. Baron for the opportunity his project provided me to consolidate my understanding of my thesis. My work as detailed by that publication was funded by an *EPSRC Impact Acceleration Account* grant awarded to Dr. Baron.

Dr. Stephen Wattam has been a reliable sounding-board throughout my research, but must be particularly acknowledged with regard to [70], where his experience in and theoretical understanding of sampling web documents was applied to my great advantage in ensuring the evaluation was rigorous.

Finally, I would like to acknowledge in general all the members of the `#lucs` IRC channel, who have been simultaneously a support group, a circle of critical reviewers, a terrible distraction and a great source of motivation.

Abstract

The explosion in popularity of online social networks has led to increased interest in identity resolution from security practitioners. Being able to connect together the multiple online accounts of a user can be of use in verifying identity attributes and in tracking the activity of malicious users. At the same time, privacy researchers are exploring the same phenomenon with interest in identifying privacy risks caused by re-identification attacks.

Existing literature has explored how particular components of an online identity may be used to connect profiles, but few if any studies have attempted to assess the comparative value of information attributes. In addition, few of the methods being reported are easily comparable, due to difficulties with obtaining and sharing ground-truth data. Attempts to gain a comprehensive understanding of the identifiability of profile attributes are hindered by these issues.

With a focus on overcoming these hurdles to effective research, this thesis first develops a methodology for sampling ground-truth data from online social networks. Building on this with reference to both existing literature and samples of real profile data, this thesis describes and grounds a comprehensive matching schema of profile attributes. The work then defines data quality measures which are important for identity resolution, and measures the *availability*, *consistency* and *uniqueness* of the schema's contents. The developed measurements are then applied in a feature selection scheme to reduce the impact of missing data issues common in identity resolution.

Finally, this thesis addresses the purposes to which identity resolution may be applied, defining the further application-oriented data quality measurements of *novelty*, *veracity* and *relevance*, and demonstrating their calculation and application for a particular use case: evaluating the social engineering vulnerability of an organisation.

Table of contents

List of figures

List of tables

Glossary of acronyms

ACID A framework defined by Goga et al [83] for understanding the reliability of identity-resolution systems.

ACU The *availability*, *consistency* and *uniqueness* model presented in Chapter 4, which builds upon and refines ACID.

API An *application programming interface*, defining subroutines for use in software. Used here to refer to such interfaces offered by web services.

FOAF *Friend-of-a-Friend*: an online ontology system and project for describing people.

NB *Naive Bayes*, a probabilistic classifier with strong feature independence assumptions.

OSINT *Open-source intelligence*, intelligence gathering from public sources such as newspapers, television and the Internet.

OSN An online social network, or the service provider underlying the same.

ROC The *receiver operating characteristic* plot or curve, describing the performance of a binary classifier.

SMS A systematic mapping study, a particular replicable method for surveying the available research literature for a topic.

SVM A *support vector machine*, a classification method which maps data to a multidimensional space in which classes can be separated by a hyperplane.

URL A *uniform resource locator* for locating a resource on a network, most commonly a web address.

Chapter 1

Introduction

The explosion in popularity of online social networking services over the past decade has led to increased interest in *identity resolution* from security practitioners. With many different platforms recording different information about a person's life, efforts at *data fusion* – connecting together these strands of information – tend to centre on the identification of which online profiles refer to the same person. In turn, these methods are powering security analytics for tracking malicious or criminal behaviour, and providing powerful insight into the privacy protection necessary for users of online social networks. For example, data releases from scientists studying social networks have demonstrated how anonymising their contents is more challenging than simply removing classical categories of "personally identifying information" [258].

Existing literature has explored how particular components of an online identity may be used to connect profiles in the absence of ground-truth: researchers have demonstrated that content ratings [155], friends lists [128], writing style [154] and other incidental data [83] can be used to link online accounts that correspond to the same person. However, few if any studies have attempted to assess the comparative value of information attributes for this purpose. Questions about which of the many kinds of profile elements are more useful cannot yet be answered outside of individual cases on particular datasets and tasks. The field is lacking a grounded understanding of the important dimensions by which data can be judged suitable for identity-resolution purposes, and exploitable profile elements may be being missed as a result.

In addition, few of the methods being designed and published are easily comparable, due to difficulties with obtaining and sharing ground-truth data which necessarily reflects personally identifiable information about its subjects. Where these methods have been reproduced, there are indications that methods underperform compared to original results [84], in part due to the sparsity of real ground-truth data, and in part due to differences between 'messy' raw data and the more idealised datasets used in prior studies. Attempts to gain a comprehensive understanding of the identifiability of profile attributes are hindered by these issues.

Such deficiencies in comparative understanding and replication are becoming increasingly critical as identity resolution matures as a research area. The use of these technologies by law enforcement relies upon their methodology being defensible in a court of law, which requires in the long term comparative standards of evidence and reproducible results. More broadly, identity-resolution technologies relate directly to privacy risk for social network users, and being able to prioritise and quantify privacy-risk behaviour in a reliable manner could be critical to preventing harm.

The central position of this thesis can be understood as being that, ongoing developments notwithstanding, the field of identity resolution will be hindered until the problems of data quality – in both theoretical understanding and practical quality control – are properly addressed.

As an illustration of some of the potential applications of this work, consider the following motivating identity resolution scenarios:

1. A police officer has a suspect for a physical crime such as theft or vandalism. In questioning, the suspect has presented activity on one of their friends' online accounts as an alibi. The officer is suspicious that this activity is misleading, and wants to check other online accounts from this individual to see whether they corroborate the story or provide evidence contradicting it. They don't want to give said individual the chance to hide any such evidence. Simply searching for the name on the given profile brings up a large number of profiles on other social networking services. What should the officer compare between the volunteered

profile and these hits, to find the right individual? More broadly, what should a tool being built to support this activity do to handle this problem, and which social networking sites should it target?

This is a *one-to-many* identity resolution problem. There are a number of possible online services on which the individual in question might hold an account, and they may hold multiple accounts, or conversely may not necessarily hold any.

2. An administrator of an online community is looking to identify potential disruptive elements in their large user-base. They have a database of profiles for known misbehaving users from another online community, and want to match any of their own community's profiles that appear to belong to the same individuals, to flag them for monitoring. What are the important features they should compare between their own data and the misbehaving profiles from their competitor?

   This is a *many-to-many* identity resolution problem. While the number of datasets to search is constrained, there are multiple profiles that may have multiple matches in the community database.

3. A credit rating agency is tasked with identifying documents and debts associated with an individual, based on a few personal details, across a range of databases storing records of debt and financial obligations. This check could well be instigated by the subject themselves, aiming to verify a particular claim about themselves (e.g., that they have no debts) to a third party. Which details should the agency value most for performing this matching?

4. A maintainer of a university department's internal bibliographic database is updating their system and merging records with the institution-wide database. They need to identify which authors in both datasets are the same individuals. How can they understand which fields are likely to be useful for this matching? Which fields should they make mandatory in their new system to best preserve the linkability of the dataset in the future?

5. A corporation's HR department is performing a review of staff security habits. They are concerned about staff use of social media, and want to compare the internal staff directory to online social media and see whether any of their staff are visibly engaging in risky security behaviours. There are a large number of staff, so they want a program to attempt identity resolution between the staff records and online profiles, and then report on security risks. How can this tool best choose which features to use and online networks to check? How can the tool judge which profile attributes carry the most 'risk', to feed back into policy guidance?

It should be noted that these motivating scenarios are not what is known as *adversarial*. That is, specifically, they do not assume that profile owners are attempting to disguise a connection between profiles (e.g., using different names, pictures, etc.). Nor, however, are the subjects attempting to help identity resolution – the purposes for which it is being applied are not in all cases aligned to their interests, and in some cases may well be contrary to their interests. The area addressed is a broad spectrum lying between the cases where identity resolution is voluntary, and subjects might willingly provide connections, and where it is undesirable and they are able and/or precognisant enough to act to prevent it. In the first two examples, the profile owner would likely benefit from such an action, but are unlikely to have anticipated the need, or found it pressing enough to outweigh other benefits from non-adversarial use of social platforms. The methods discussed in this thesis could well be adapted to adversarial use cases (e.g., amongst professional criminals' profiles on underground fora) but particular attention would have to be paid to potential differences from general population identity resolution, and different data sources would be required as a basis of estimates.

1.1 Limitations of the State of the Art

A broad survey of data mining for law enforcement purposes, presented in Section 2.1, suggests that technology and methodology for handling online identities – such as linguistic de-anonymisation through authorship attribution methods – is one of the most

fruitful areas for practically assisting security efforts. This review also suggests that methods combining different data types (such as text, images, video, and structured data from online social networks) may be relatively scarce in the context of direct application to law enforcement. As online social networks (OSNs) are of large and growing interest to law enforcement, and present a broad selection of heterogeneous data types as features, this area is of particular interest.

Further investigation reveals a number of works which do integrate multiple data types in this form of OSN identity resolution. However, this research area appears hampered from fully forming as a research community. Novel methodologies are proposed and evaluated for identity resolution, but these results are rarely comparable to prior work. This limitation stems in part from a lack of reliable ground-truth data sources, meaning that most authors use different sources, which are rarely available to others at a later date. A resulting lack of replication efforts leads to a reduced reliability of published results on 'real' identity-resolution datasets with realistically challenging candidate sets and a high incidence of missing profile data [84, 222].

Most importantly, however, few if any publications working on identity resolution in modern web data show an awareness of the foundational theories of probabilistic record linkage, for the most part positioning their work in a general context of machine learning classification problems. Much of the work exists in isolation, with methods not only failing to compare with alternatives attempting the same goal, but lacking even a framework for expressing results about the identification value of approaches in the same language as other work.

It is at these areas of identity resolution – the practical and theoretical grounds that are hampering the cohesion and comparison of valuable research – that this thesis aims its contribution.

1.2 Thesis Objectives

A solution to the ground-truth problem in identity resolution

Identity resolution research is being hampered by a lack of readily-available ground-truth data which can be shared between researchers. All identity-resolution datasets by definition contain personally identifiable information, and the scale required for developing reliable methods does not permit the solicitation of individual consent for its release. This thesis will approach the problem from the perspective of sampling theory, aiming to show that certain sampling methods can resolve this conflict by allowing researchers to draw comparable samples from social networks.

A model enabling comparative assessment of the identification value of all common profile attributes

The lack of a common framework for understanding the identification value of profile attributes is a hindrance to identity resolution research, with results left in isolation, improperly comparable. This thesis will build up an abstract model of online profile attributes, and develop a model for understanding the value of each attribute, along with initial estimates of the values of these attributes, grounded in both the available literature and original measurements. An additional target of exploration is understanding the value of attributes for the purposes of *applying* identity resolution to some particular task, (such as, for example, identifying a company's employees in online social media), an activity which raises related but distinct concerns.

Improving the reliability of identity resolution methods on real datasets

Drawing upon these measurements, this thesis explores how concrete gains in performance and reliability might be obtained. One data quality problem of significance for identity resolution is missing data in data sources such as social networks. An exami-

nation of missing data mitigation strategies was carried out, complemented with novel methodology which is drawn from domain understanding of data quality measures.

1.3 Approach & Contributions

It is the contention of this thesis that knowing how to measure certain qualities of personal profile information, and thus compare and contrast different identity-resolution systems, is of paramount importance. Armed with knowledge about the important qualities of profile attributes and their similarity measures, researchers can create measurements for their populations of interest, find profile attributes which are useful features, build upon insights drawn in each others' papers, and produce more accurate and reliable identity-resolution systems. In turn, these more reliable approaches would improve the accuracy and reliability of security and privacy research relying on such systems.

Survey of data mining research for law enforcement

To situate this thesis in the wider context of data mining for security practice, a literature survey is carried out. Following recent adoption in software engineering, which mirrors longstanding conduct in medical research, this survey is carried out as a systematic study, with a reproducible method. Further background material pertinent to this thesis is surveyed in a more traditional manner.

Matching schema

To begin understanding the value of information in online profiles, it is first important to understand what types of information even exist in the relevant population. Accordingly, this thesis presents a well-grounded matching schema for online profile information. Coming to a standardised model of what information might be found in online profiles is not necessarily a simple task, given the broad and rapidly-evolving nature of social media. An inclusive approach is taken, drawing a schema from multiple highly-ranked websites which might contain user profiles, referred to hereafter as *profile networks*.

Data quality measures for identity resolution

The schema is then used as a basis for applying and grounding six key data quality measures. The first three of these – the *availability*, *consistency* and *uniqueness* of information items – are based on developments of fundamental theory in the related statistical field of record linkage, and relate directly to the value of information for the purpose of resolving identities. The full definitions and background to each measure are developed in the relevant chapters, but in brief:

- *Availability* is the degree to which the presence of an information item on profile networks can be expected.
- *Consistency* reflects whether a given user presents the same value for this information item across different profile networks.
- *Uniqueness* reflects the diversity of values that an information item might take in a population.

Data quality measures for identity resolution applications

The remaining three, the *novelty*, *veracity* and *relevance* of information, measure the value of information for the end purpose of any particular identity-resolution system. They are developed in a separate chapter, with a case study to demonstrate application. Again, in brief:

- *Novelty* brings in the question of context, valuing information based on contrast to a prior set of information.
- *Veracity* measures how likely it is that this item is accurate as it pertains to the person who made the profile.
- *Relevance* refers to the usefulness of an information item to the end-goal of an attempt at identity resolution.

Sampling ground truth data

In addition to addressing the value which can be ascribed to profile attributes, this thesis makes other contributions to improving the comparability and reliability of data for identity-resolution systems. The first focuses on the notable data-collection and sharing obstacle within the field, where a solution is proposed based on comparable sampling strategies from ground-truth resources. By providing a common reference point for ongoing research, results can be more readily compared.

Handling missing data

The second practical contribution focuses on an aspect of data processing which is often overlooked – the handling of missing data. The profile data used in identity resolution is often high-dimensional, and real-world data from social networks routinely presents missing values. Naive approaches to handling this data are likely to bias results and reduce reproducibility in application. This thesis provides an exploration of mitigation methods, including novel approaches grounded in the previously developed measurements of the domain.

In combination, these contributions aim to improve both the core understanding of identity resolution, and the standard of replication and comparability within the field.

1.4 Structure of this Thesis

From this point on, the thesis is structured as follows.

Chapter 2 presents a summary of the systematic survey which situates the field within identity resolution, followed by an overview of the field of identity resolution. This starts with relevant background in the field of probabilistic record linkage and works up to the modern day applications in an online social networking context, highlighting contributions which move towards an understanding of information value.

Chapter 3 tackles the problem of data collection and sharing, covering the situation facing identity-resolution researchers and proposing and evaluating a solution.

Chapter 4 forms the core of the thesis, and sets out the basis for the *availability*, *consistency* and *uniqueness* quality measurements, followed by a common schema for online profile information, and concluding with detailed literature-based estimates and original measurements of these properties.

Chapter 5 then addresses the problem of missing data in identity resolution, and makes use of the measured attribute properties to generate a novel solution, alongside observations about trialled mitigation approaches.

Chapter 6 outlines the *novelty*, *veracity* and *relevance* measures and applies them to a particular application in social engineering vulnerability detection.

Chapter 7 concludes by discussing the impact of this approach to privacy preservation, and outlining directions for future work.

1.5 Publications Emerging from this Thesis

All work presented in this thesis, unless otherwise indicated, is that of the author. Some of the work presented in this thesis has been previously published in various venues, with occasional support from other authors, as described below.

1. The matching schema for online profile information, described in Chapter 4, Section 4.3, has been previously published under the title *A service-independent model for linking online user profile information* [68], by the author and his supervisors.

   [68] Edwards, M., Rashid, A., and Rayson, P. (2014). A service-independent model for linking online user profile information. In *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 280–283. IEEE Computer Society.

2. The systematic survey of literature pertaining to data mining for security informatics purposes, described in Chapter 2, Section 2.1, has been previously published under the title *A systematic survey of online data mining technology intended for law enforcement* [69], by the author and his supervisors.

   [69] Edwards, M., Rashid, A., and Rayson, P. (2015). A systematic survey of online data mining technology intended for law enforcement. *ACM Computing Surveys (CSUR)*, 48(1):15.

3. The sampling mechanism to enable replication in identity resolution studies, described in Chapter 3, has been previously published under the title *Sampling labelled profile data for identity resolution* [70]. Alongside the author and his supervisors, Dr. Stephen Wattam provided insight and statistical expertise in evaluating the methodology.

   [70] Edwards, M., Wattam, S., Rayson, P., and Rashid, A. (2016). Sampling labelled profile data for identity resolution. In *Proceedings of the IEEE International Conference on Big Data*. IEEE.

4. The social engineering vulnerability detection case study, given in Chapter 6, is based on a publication by the author and others, titled *Panning for Gold: Automatically analysing online social engineering attack surfaces*. The case study relates to the application described in the paper, and is partially founded on data drawn from that study, but the paper does not describe the measurement of *relevance*, *novelty* or *veracity*. The authors Robert Larson and Benjamin Green carried out and reported on the interviews which ground the *relevance* items, and Dr. Alistair Baron was the instigator of the project.

   [67] Edwards, M., Larson, R., Green, B., Rashid, A., and Baron, A. (2017). Panning for gold: Automatically analysing online social engineering attack surfaces. *Computers & Security*, 69:18–34.

5. In addition, reporting on the *availability* of profile attributes, in combination with an outline of the ACU model, an exploration of the impact of missing data on identity resolution performance and the development of the *a-priori* and *a-posteriori* models for feature selection, all form a publication which is currently under review, and may be published while this thesis is under consideration.

   Edwards, M., Rashid, A., and Rayson, P. (unpublished). Hunting through empty fields: robust identity resolution attacks on online social networks. *Under review*

Chapter 2

Background

This chapter reviews the background to this work from three perspectives. In Section 2.1 immediately below, a report is made of a systematic survey of data mining literature targeted at law enforcement professionals, reviewing the problems addressed and methodology employed, and scoping out areas for valuable contribution within this broad domain. Having identified such an area in multi-attribute identity resolution, the chapter continues in Section 2.2 by outlining the foundational background of identity resolution as a probabilistic technique, along with an illustrative example of some common issues in the field. Section 2.3 then concludes with an overview of research from the last decade or so which approaches identity resolution in the context of the web and social media, these being the immediate context which this thesis addresses. A short summary then recaps the chapter and the situation of the thesis against the covered material.

2.1 A Systematic Survey of Security Informatics

This section aims to situate this thesis, which focuses on identity resolution, in the wider application context of data mining for the support of law enforcement. The increasing fusion of digital and physical life presents two key challenges to law enforcement agencies: the population's online presence means law enforcement must learn to adapt to crimes taking place only online, and at the same time digital interaction also provides a valuable evidential resource for officers investigating both physical and online crimes.

Understanding and connecting identities online emerges as a crucial aspect of this field, but it is helpful to understand also the methodologies, opportunities and problems across the domain.

With Internet accessibility widening and ever more crime taking on a digital aspect, online investigation is becoming a critical tool for law enforcement organisations, and scientific examination of such processes becomes ever more a key issue.

With manual inspection of online information being labour-intensive and the unprecedented scale of information online, law enforcement agencies seek to optimise their surveillance or investigation of online data sources through the use of various data mining technologies. This behaviour has diverse implications, including raising social and ethical questions about privacy and the role of state surveillance as well as posing unique technical challenges for the data mining technologies being employed.

The field of computer science, particularly data mining research, has a key role to play in shaping the future of these investigations. This section aims to support that role, by identifying within the literature some open research problems and highlighting a research agenda for the community at large. While the broader aim is to survey the literature for gaps, guiding questions were developed to help target the extraction of information:

1. What are the problems (crimes, investigative requirements) which are being addressed in the literature?

2. Which online data sources are being used?

3. What are the methods (data mining techniques) which are being employed to provide solutions?

4. Are studies making use of multiple data sources?

5. Are studies validating their contribution's utility to law enforcement practitioners?

These questions are answered through a comprehensive search for and evaluation of peer-reviewed computer science studies concerning the mining of digital data sources for

law enforcement purposes. The study is aimed at examining the visible state-of-the-art with regards to both techniques and the criminal activities being addressed.

Taking inspiration from a recent trend towards evidence-based practice in software engineering, the survey takes the form of a systematic mapping study (SMS), with the intent of producing a survey which not only covers the available literature and has replicable results, but also can be methodologically examined for deficiencies.

It is an important part of a review's design to make clear not only the scope of the survey, but the intended purpose. The primary concern is that the results of the survey identify gaps in the published research regarding data mining of online sources for crime detection or investigation purposes. The results of the study can then be used to inform ongoing work in this area.

Two terms should be considered as key here. Firstly, the specification of *online* data sources, meaning data which can be gathered from examination of Internet-based sources. This distinction separates this study from other areas of research such as work which makes use of restricted criminal records or other police databases, as well as distancing the study from many areas of digital forensics which focus on the investigation of hard disks or active machine memory. Secondly, the specification of data mining with application in *crime* detection. While many data mining methods have plausible application in this domain, only publications which make an explicit reference to such employment are considered. For purposes of scope, data mining which is performed for purely Information Security reasons is excluded, thus leaving aside a mature literature on Intrusion Detection Systems and related work which has already been surveyed [15].

2.1.1 Method

A systematic literature review (SLR) attempts to provide answers to a specific research question through a transparent and objective approach to the collection and synthesis of existing scientific literature on the topic. This method can be contrasted with non-systematic literature reviews, whose contents may be unrepresentative of a field of research due to, for example, narrative-driven distortion, where reviewers include only

Table 2.1 Search queries were constructed by the combination of quoted forms of the following term-sets

First Term-set
Crime
Police
Law Enforcement

Second Term-set
Artificial Intelligence
Data Fusion
Data Mining
Information Fusion
Natural Language Processing
Machine Learning
Social Network Analysis
Text Mining

papers whose findings support their line of argument; or narrowness of study, where reviewers are unaware of a large number of relevant publications because they were never personally exposed to them.

The design of this survey drew heavily on the work of [163], which recommends an explicit eight-step system to SLRs. Certain key deviations from this procedure adapted the process to a systematic mapping study (SMS). A description of the main features of SMSs is provided by [36], but the key distinction between the two types can be summarised as an SMS being an SLR which aims to more broadly survey the available literature rather than answer specific research questions.

The search process, carried out between December 2012 and January 2013, was designed as an automated search, targeting four key computer science publication databases: IEEExplore, The ACM Digital Library, SpringerLink and ScienceDirect. In each database, 24 queries were carried out[1], as defined in Table 2.1 and the resulting papers' metadata collected. In total, 13,246 unique results were collected.

The title and abstract of each result were then examined by a reviewer and classified as either relevant or irrelevant to the study according to the following criteria. If the answer to any of these questions was no, then the study was not included.

1. Does the study appear to address or make use of online data — that is, types of data which may be discovered online (either on the Web or otherwise)? Specifically excluded are data such as disk images from a crime scene and restricted databases.

[1]These were identified through a series of pilot searches to determine the relevance and effectiveness of particular search strings.

2. Does the study have a stated or heavily implied application in law enforcement, crime detection, monitoring or investigation? For the purposes of this study, studies dealing primarily with attacks against computer infrastructure (intrusion detection systems) are specifically excluded.

3. Does the study appear to have a methodology which involves either fully automated or machine-assisted processing of data?

Following the screening process, all references were gathered from each of the 116 accepted papers, along with all papers identified by Google Scholar as having cited the accepted items. These were also put through the screening process above.

Following the search and screening stages, the full text of each accepted paper was obtained, along with a full citation. Items for which a full text could not be located, those which turned out to be in a non-English language or which on review of the full text did not meet the screening criteria, were discarded. This resulted in a final included list of 206 accepted papers. Each paper was examined to answer specific questions regarding its quality. Each paper was given a value of 0, 0.5 or 1 for each of the following points, with 0 being a negative response, 0.5 being a partial positive response and 1 being a positive response. The overall quality rating for each item is the sum of its individual scores on these responses.

1. Does the paper outline its method in a replicable manner?

2. Does the paper make its evaluation replicable?

3. Where evaluation is qualitative, does the evaluation make use of domain experts?

4. Where evaluation is quantitative, is an appropriate statistical assessment of results carried out?

Alongside the quality analysis, questions related to this review's main aims were answered for each paper.

1. What problems are being addressed?

2. Which data sources are being used?

3. What methods are being employed?

4. Does it make use of multiple data sources?

5. Does it validate the contribution's utility to practitioners?

All examination was carried out by the same examiner, making use of a predefined data extraction form made up of these questions. Papers were categorised according to whether papers were addressing a similar problem and whether they were using the same methods or data sources.

2.1.2 Papers surveyed

Identification with computer vision

Identification tasks in computer vision mostly rest on visual identity, a troublesome concept in an environment where a new face – or no face at all – is so simple to obtain. The majority of the uncovered literature looks at a particular subset of visual identity – recognising the visual representation of a person in a particular online environment, such as the game Second Life. The challenges here have significant overlap with the development of facial recognition systems in general, including automating adequate pre-processing to find comparable facial images and minimizing the runtime of any face-matching system. Additional problems are raised by the possibility of 'different worlds' where the same person may use a different visual identity, something not generally possible in the real world. Some authors posit that avatars made by the same person may be consistent across different services, or in some way connected to their actual visual appearance, but this appears unproven.

[240] discuss the application of computer vision to the facial recognition of online avatars, justifying the research topic with reference to criminals and especially terrorist groups using virtual environments — particularly the online game Second Life — as

training simulators, and studying two key applications. The first of these involved inter-reality avatar-to-photograph matching, where avatar faces generated from photographs were matched against other photographs of the same subject. Off-the-shelf face recognition technology sufficed here, given that the avatar was generated automatically from an actual image of the target. While a useful result, this merely suggests that their automatic avatar generation system preserves key information for facial recognition, and not that users will do so when crafting their own avatars. Addressing this, the second application used a collection of actual Second Life avatars and attempted to match different images of these avatars. The authors discovered acceptable classification accuracy, although they reported a performance bottleneck in face and eye detection, with significant improvement in accuracy coming from manual eye location. A remaining question left unanswered is whether avatars can be recognised across different digital platforms.

[17] introduce a new method for avatar facial recognition, employing wavelet transforms alongside a hierarchical multi-scale local binary pattern (HMLBP). The authors build on other developments in this area, including earlier propositions of the use of wavelet transforms and local binary patterns. The study focuses on re-detecting avatars from Second Life and Entropia — another online world — from pictures in different poses. While the results show improvements over earlier work, the method still relies on correctly-cropped input, achieved in the dataset through manual effort. This hurdle must be addressed for a completely automated system for detection and recognition of avatar faces.

[149, 150] both continue with the use of wavelet-based local binary patterns, for re-detecting avatar faces, but with new variations, one making use of Eigenfaces and the other making use of directional statistical features. Both experiments re-used the Second Life and Entropia datasets presented in the previous publications. In the first of these two publications, the authors mention the design of a fully automated system for addressing the cropping problem as a target for ongoing work. In the second, the authors include a comparison of the classification time for a number of leading techniques, an important consideration for any near-realtime detection and recognition system.

In summary, this niche area of facial recognition has shown significant development with regards to the core task of re-identification of avatars from within an online environment like Second Life or Entropia. Still awaiting research is meaningful deployment of these classification systems, with work evaluating automated cropping and exploring usable interfaces to the online environment appearing on the horizon. In crossing realities, some initial work evaluating the traceability of the results of automatic avatar generation has been undertaken (e.g. [240]), but it remains to be seen if links between a user's visual appearance and virtual avatar can be determined, or indeed if avatars are consistent across online environments.

Other computer vision work on identification includes that of [250] and [231]. [250] focus specifically on image spam — spam emails which make use of text presented as images in order to avoid text-based filtering techniques. They cluster spam images by visual features, and report a high success rate with respect to a manually-identified ground truth. Their approach analyses images for evidence of various types of template being reused by spammers, as divined by layout, colour and texture. What they do not report in this paper is whether typical spam filtering, or indeed linguistic analysis, can be applied to text extracted via optical character recognition. [231] describe a more general spam-origin toolkit which makes use of website image comparison (from following links in spam emails) as one tool in its arsenal, alongside WHOIS and IP lookup information and more typical email attributes such as subject lines. While their analysis of a researcher-gathered dataset appears to reveal interesting clusters of spam, and the utility of the website image comparison in particular is demonstrated by all but one cluster centring on one website image, an evaluation against manually-identified ground truth would be stronger justification of the method's validity.

Computer vision and crimes against children

Computer vision's role in preventing crimes against children is mostly connected to the recognition of child abuse imagery in an online population of images. This can be either searching for known child abuse imagery in order to filter it or identify distributors, or else

identifying new examples. A common problem in the second domain is distinguishing between ordinary adult pornographic material and images of children in pornographic context, which is highly visually similar.

[92] present the FORWEB system, which focuses on forensic applications of existing signature analysis and web-crawling systems, the key motivation of the authors being to automate the search and discovery process involving networked servers. They clearly distinguish their approach from established storage media analysis tools like *EnCase* in much the same way as this review separates these areas of study. Their file-fingerprinting scheme aims to identify images based on properties which are much less likely to be affected by the simple alterations which throw off hash-based file comparisons, and combined with the spidering bot this becomes a useful tool for detecting known malicious images. Like all such tools, this relies on the existence of an up-to-date database of known suspect files (a resource which may in itself bring significant performance overhead) and does not address the aim of identifying unknown media of a suspect nature.

[99] takes a quite different approach, detailing a method whereby image files are identified on the network, reconstructed and then classified as either child abuse media or not by both a machine learning system and an image matching system similar to FORWEB's fingerprinting scheme — the intent being that such a system would be installed on network boundaries to filter child abuse material. Aside from concerns about network performance, a major weakness of their trial of the system is that for legal reasons their system only attempted to distinguish between nude and non-nude images, which is clearly a far easier task than distinguishing child features from adult ones.

This also applies to [220], which opens with a motivation of preventing child abuse, but in a sudden switch focuses on identifying pornographic video scenes as a proxy. The paper neither provides an implementation nor an evaluation of the system, merely outlining methods to be explored.

This more difficult child-recognition task is tackled by [110]. They focus on detecting child exploitation material on social networks, but contribute an algorithm which could equally apply to P2P networks. Their skin detection technique is specifically tuned for

the detection of child skin tones, and they also suggest techniques which help detect pornographic context. While these proposed methods are indeed critical research areas for computer vision in child protection, the authors do not provide the results of an evaluation or even a completed system.

Also attempting this task, [198] aim at detecting child abuse material on the network level, as an alternative to manually searching suspicious venues or application-layer networks. Their classification system, consisting of a stochastic learning weak estimator combined with a linear classifier, was trained and trialled on a sanitised dataset provided by Candian law enforcement, a rare example of child abuse imagery being available for training. Notably, the classifier was tested on partial as well as whole images, taking into account likely fragmentation of images over a network link. While valuable for this alone, the classifiers being trialled still produced less-than-ideal rates of false positives for a tool to be deployed at the network level. Estimations of the base rates for child abuse material versus adult pornography suggest that alerts generated may be mostly incorrect – though this does not invalidate the utility of the classifier as a tool for network monitoring, given appropriate human supervision.

Computer vision and threats or harassment

There is a specific use-case for computer vision in detecting visual (as opposed to verbal or written) forms of harassment in video communication. As these systems are intended to be deployed on large video-streaming populations, performance is critical to creating a deployable solution. The particular misbehaviour discussed in these papers has some particular visual challenges regarding lighting and the potential detection of faces.

[236] outline an unusual problem specific to the anonymous video-pairing site Chatroulette – users exposing themselves to strangers. They stress that a considerable proportion of Chatroulette's userbase would be classed as minors, and that site policy on age restriction and obscenity is difficult to enforce due to the anonymous nature of the service. The authors note that de-anonymising the service could solve this issue, but would damage one of the site's key features in the process, and so turn to video-

analytic approaches for detecting offensive users. Their key observations include that misbehaving users usually hide their faces, and that misbehaving users' images differ from pornographic images in that they often stay partially clothed and only expose their genitals. Their system therefore focuses on detecting user faces as a key feature in making a decision about the probability of misbehaviour, along with a novel skin detection system which takes into account the abnormal context of webcam images. While they evaluate their classification accuracy, they do not report on performance speed, an issue which would appear critical for their problem domain, as extra delay in connection would impair the appeal of the Chatroulette service.

[48] refine this first approach into a fine-grained cascaded classification solution which filters out easily disambiguated images earlier in the process for the sake of efficiency. They also integrate new work on gathering contextual information from webcam images and a new fusion system for combining probabilities of misbehaviour. The improved system is evaluated against their older system and other contenders, showing significant improvement, particularly in regard to the previously unaddressed matter of classification latency.

Computer vision and terrorism/extremism

A limited deployment of computer vision techniques in counter-terrorism is seen in the context of the analysis of propaganda videos released by jihadists. The problems they address are of coding the content of the videos in a pseudo-automated fashion, where correct identification can be an aide to intelligence work.

[192] present an exploratory study of jihadi videos which attempts to highlight the research and intelligence need for automatic exploration of jihadi video content, and produce a tool to support manual coding of videos for this purpose. The results are demonstrative of the effectiveness of their analysis on a set of terrorist videos and not that of the performance of their coding toolkit.

[193] provide an extended version of the same research, again with more focus on the content analysis than on the support tool. In both cases, while the authors' work

is presented as a stage towards automated video content analysis, the requirements for progression from manual intervention are not fully detailed.

Computer vision and financial crime

Computer vision has seen deployment in anti-piracy efforts. The systems in this section are attempting to detect copies of restricted material from being distributed online by comparing content to a stored visual fingerprint of pirated material – techniques also deployed in detecting known child abuse media. The problems addressed in these publications are primarily infrastructural, attempting to resolve detection efforts with minimal impact on legitimate traffic.

[97, 248] describe a system for large-scale online monitoring at the Content Distribution Network level, wherein videos are fingerprinted based on certain visual cues and compared to a blacklist of pirated material. Their system is particularly notable due to the fact that it was actually deployed on a large CDN, although the evaluation presented seems to be from laboratory results rather than real-world performance. Nonetheless, it would appear that their system is resilient to minor tampering such as is common with pirated material. Notably for a system to be deployed at a large scale, the performance overhead is quite significant, with fingerprinting and search time together incurring a 40 second delay, raising questions about usability.

[98] address problems linked to the computational and networking overhead of this large-scale video processing by deploying server clusters closer to the user in the content distribution network and distributing tasks between nodes based on proximity and computational load. This results in reduced processing time as compared to existing approaches, but still requires well over a minute to perform detection on movie-length items. It would appear that, despite ongoing work to address this issue, there is scope for improvement in the efficiency and scalability of video copy detection.

Other computer vision applications

[95] target the detection of pornography, but do so with reference to illegal or offensive activity — whether the authors suspect that pornography is illegal, or target illegal pornography particularly but work with proxy data, is unclear. Their method addresses not only image recognition, but also the text processing of suspected pornographic web content, combining this information in their classifier. Their contour-based detection method appears to perform better than region-based skin detection, specifically with regard to false positive rates including bikini or face-focused images.

[226] describe existing general-purpose information filtering systems which they suggest could be used to defend users against various types of information, insult or crime. A range of methods and systems for information filtering are outlined, but neither methods nor systems are subject to a great deal of scrutiny. How information filtering technologies such as those presented can be linked to the prevention of crime is also not clearly outlined.

SNA and terrorism/extremism

As a set of tools for analysing communities and graphs, social network analysis has seen particular deployment in counter-terrorism context, where the analysis of groups can be useful in identifying key nodes and group behaviour. Particularly, it is applied to graphs which are mined from online forums and blogs, where relationships between individuals can be determined structurally from links.

[40, 237], focus on mining and analysing online communities in blogs, specifically communities of blogs frequented by hate groups. These two studies both make use of the same 28 anti-black blogrings from the Xanga blogging platform. While the studies include semi-automatic detection of hate groups as a key aim, the selection process presented relies on manual filtering of search results. A more automated means of selecting hate groups could aid in making their approach generalisable.

[185], is the first of a number of studies making use of the Dark Web Forum Portal collection. The authors focus on detecting overlapping communities by using latent

dirichlet allocation to detect topics, with a positive evaluation on an English-language forum from the Dark Web Portal. The treatment of networks as allowing members to be part of more than one community is perhaps a useful model, but whether topics of conversations reflect actual networks rather than simply ideological leanings is not clarified.

[131] focuses on the process behind online radicalisation. This work includes a well-written motivating example, and a review of current theory related to online radicalisation, but most importantly for this review it also includes a social network analysis using forum data from two Dark Web fora, one from the middle-east and one from Europe. Interestingly, the author reports technical issues with a module of the Dark Web Portal. The analysis suggests that radicalisation is happening between the most involved members of the community, as identified by several measures of centrality.

[241] gathered discussion data from MySpace, using the DBSCAN algorithm to cluster topics as points for a social network visualisation tool. While the level of detail in the description of the algorithm is adequate, the authors' choice of example in their demonstration of the tool is the only link specifically to terrorism. Further detail on what may constitute interesting patterns within the network resulting from their clustered topics would make the tool's utility to terrorism investigators clearer.

[171] describe the Dark Web Attribute System which applies content and link attributes to items from the Dark Web collection, calculating measures of technical sophistication for various linked terrorist websites. The evaluation lacks rigour, however, and doesn't effectively demonstrate what might well be useful annotation work.

[41] describe the application of SNA techniques as part of a system for identifying and monitoring terrorists at the ISP level, also advocating their system's use for targeted disruption of terrorist networks through identifying key nodes. The paper describes only a theoretical system and provides no evaluation. Most pressingly for a paper advocating large-scale surveillance, they include no discussion of the likely rate of false positives. Their baseline is also likely to be misleading, as they base their threshold of typical terrorist behaviour on only terrorist content, ignoring the possibility that terrorist

individuals may access other sites. A more behaviourally sound model of terrorist web usage would be of use in improving such a system.

[157] describe a method utilising social network analysis for detecting changes in a group's behavioural patterns, as observed via email communications. They particularly highlight homeland security and intelligence applications of this method. They do not provide an evaluation in this paper, but discuss their ongoing development of a simulated email dataset for that purpose. As they discuss, their current model does not handle dynamic social networks such as those they expect in real data, an area which needs redressing. A key limitation of any such simulation would be its validity as a predictor of performance on a real email network – it would seem more advisable to work with real email datasets in developing the analysis methods outlined, even where this means working with proxy data rather than actual terrorist network data.

[213] analyse YouTube's social graph to discover extremist videos and communities. Their system works from a seed list of videos to discover YouTube videos which are hate speech and users advocating acts of aggression. The authors discuss the network properties of the connections they found – including the different types of YouTube network – alongside brief topic analysis of user comments. The main contribution here is the development of search support tools for an intelligence analyst, adding structure and ranking content, but there is limited comment on the scope of the approach.

In summary, social network analysis has been applied to a number of terrorism-related datasets with some success, but current studies tend to present either toolsets which, due to the nature of terrorist content, often cannot be evaluated easily, or else exploratory analyses of a particular network which demonstrate some value but do not generalise. A theme common to a small number of papers has been using topic analysis of text to better subdivide communities of interest, but it would appear that this approach has yet to be validated in a meaningful manner.

SNA and police intelligence

As with terrorist organisations, social network analysis has been applied to online information about criminal organisations, often mined from news reports or other unstructured text documents. This provision provides for opportunities – additional information on time or space of interactions may be available – but also additional challenges in that relationships are not necessarily correctly represented in such secondary sources.

[176] provide a case study where link analysis – with links in the form of webpage co-occurrence – is used to trace a notorious violent criminal, producing link charts for known members of his gang and related individuals. The method presented relies on Google search results to identify relevant web pages, which may lead to narrowed results due to personalisation if countermeasures are not taken. A comparison with other methods for identifying web sources could prove useful.

[94] provide a review of web mining for input into criminal network analysis, and propose a framework which integrates the identification of crime hot spots and criminal communities into the workflow of a web crawling agent. Detail on how the more relevant tagging modules will be implemented is omitted.

[219] focus on term networks, presenting a novel algorithm for key term extraction, and presenting a case study similar to that of [176] where news related to a particular gangster was gathered and mined to describe relationships between gangsters. The term model presented appears more powerful than simple entity collocation, but the study presented does not make a convincing case for the utility of this method, demonstrating only simple relationships as could be found through more traditional means.

From this sample, the area of web mining criminal networks, like terrorism network analysis, appears to suffer from a lack of rigorous evaluation. Identification of a means of better evaluating the performance of information-gathering agents such as these could help focus research efforts. A standard marked dataset suitable for evaluation could be considered an initial step.

[129] describe attempts to discover the social networks of criminals by mining spatio-temporal events such as web usage. A detailed explanation of the problem and

algorithmic approach are given, and the theory is validated against a data set collected from a university campus' wireless network. While their system appears technologically sound and is well-presented, the intended deployment scenario is not clear.

[117] discuss integrating SNA concepts into common digital forensics practice for investigation of email. The validating case study involves transforming the Enron email dataset into a form suitable for social network analysis and highlighting key actors from within that dataset. As the thesis itself acknowledges, social network information is not 'hard' evidence which can be considered directly in court, being instead useful in guiding further investigation. The analysis of the Enron dataset presented does show some utility, but it is worth noting that the analyst's interpretation of results seems likely to be informed by previous knowledge of the dataset's context. A blinded study would mitigate such issues.

[64] makes use of Twitter data and geolocation for building a social network based on ongoing terrorist events, and then provides a modifiable visualisation to aid interpretation. Several areas for ongoing development are highlighted, including incorporation of temporal and sentiment dimensions into the visualisation tool.

[22] theoretically demonstrates a means of detecting hidden friendships — relationships in a network which are not formal connections. While a potentially valuable intelligence tool, the paper does not provide an evaluation of this method's efficacy.

[209] use SNA as part of a range of tools for investigating email data for various crime-related purposes. The social network analysis component is only one part of the tool, which is described only very briefly and not evaluated.

[9] describe the process of mining and analysing criminal networks from collections of unstructured text documents, in an approach which relies on the recognition of named entities and the detection of prominent communities of connected names. Their approach was validated in a case study from a real cybercrime investigation, with an instant messaging database provided by law enforcement and their investigation being compared to an expert's manual analysis of the chat logs. It is notable that the analysis was guided by the researchers' own identification of suspicious information – while fully

automated analysis is not necessarily desirable, for purposes of evaluation it is necessary to distinguish the performance of the support tool from the performance of the authors. A blinded study with a number of analysis engine users compared to a number of manual analysis users would provide more objective assessment of their network-mining engine's utility.

SNA and cybercrime

[136] focus on the construction of social networks from email and blog data linked specifically to cybercriminal activity. The paper refers most often to cybercrime as its motivation, but also to terrorists who 'upload obscene pictures'. The degree to which authorship identification techniques were applied is unclear.

[160] apply SNA techniques — as part of a toolkit with other subsystems — to help identify cybercriminals from email data. They appear to have implemented their system and even gathered a dataset (Enron) to trial it on, but provide no evaluation in this paper.

SNA and financial crime

[90] address financial crime through application of social network analysis in mining corporate emails to prevent fraudulent transactions. Taking the approach that data outside the accounting information system should help protect against fraud involving senior management figures, they strive to mine both the textual content and social networks of email data. They provide a competent review of relevant work and use the Enron email dataset as an illustrative example.

[170] also use SNA in detecting fraud, but as applied to transaction data from online auctions. Their method, working as a third-party service, applies Markov Random Fields to model the networks and belief propogation to detect fraud within the network. Their positive evaluation includes both a synthetic dataset and a transaction dataset scraped from the popular auction network eBay. A third-party approach such as this would appear to allow their system to adapt to a number of auction platforms, but with a risk of being rendered ineffectual by changes to site templates or APIs.

SNA and identification

[91] turn SNA methods to forensic (i.e. identification) analysis of temporal email data. The SNA component of this mostly text-mining tool is employed to provide behavioural, temporal and geographic modelling information. A partial evaluation of a different module of the toolset is provided using the Enron email database, but the SNA component is presented merely as a useful analytics and visualisation workbench.

SNA and crimes against children

[77] examine the structure of online child exploitation networks, building networks of websites based on their links and a set of predefined bad keywords, with the ultimate goal of identifying the major nodes whose removal would most disrupt online exploitation. They demonstrate their deployment on four networks crawled from websites identified through search results, identifying the key nodes through the top 10 values for in-degree and for severity of content as identified through keywords. They also find that centrality does not correlate with severity of content, but severe websites were highly linked to each other, suggesting scope for targeting subnetworks of the most extreme material where law enforcement resources are scarce.

Information extraction in terrorism and extremism

A variety of information extraction techniques can be applied in analysis of terrorists and extremism, including topic mining and summarization. Websites and forums frequented by these groups are particularly rich source of information. The main body of research tends to focus on either white supremacist groups in the US or Islamic fundamentalists.

[141] describes a system called ProfileMiner for combating cyberterrorism. This system amounts to an interface or series of interfaces to a database of online information, the compilation of which is left unspecified but appears to be tied to a commercial product called MAVIS. No evaluation is provided, nor is it clear whether the interface was actually constructed rather than simply designed.

[254] briefly outline the motivation for and design of the Dark Web Portal, a resource used in several papers addressing information extraction from terrorist and extremist sites. [255] describe a semi-automated system for collecting and analyzing information on 'Dark Web' sites, and apply this to a selection of United States extremist web sites. Their results and methodology are subjected to an expert evaluation with a positive outcome. While their automated collection stage (outlined in more detail itself in [253]) appears effective, their approach to filtering the results of said searches involves manual filtering of hundreds of URLs followed by a second stage of search to manually bulk out the results. If value over that of a typical search engine is to be added in a semi-automated collection and filtering system, it must be to reduce such loads on the analyst. The work by [42] appears to be linked, in which a case study is carried out to collect and analyse examples of Arabic fora. The same levels of expert evaluation and manual workload are evident, suggesting that the only key difference in the two works is the community being analysed.

[43, 183] take a similar approach in what may be a continuation of the same line of research. A coding system referred to as the Dark Web Attribute System is developed to look specifically for signs of technical sophistication and content richness in the design of the websites of extremist groups. The first paper uses this framework to compare terrorist sites to those of US government agencies while the second compares the internet presence of extremist organisations drawn from three geographical regions. The latter's detailed analysis of these technical indicators highlights how relatively innocuous details can be of interest when examined at scale. A combination of this attribute system with an automatic collection system could prove useful in identifying new groups that show above-average sophistication, perhaps thus better helping identify key emerging threats.

[81] report on the observed typical behaviours of those holding supremacist or separatist beliefs, as determined through an examination of 157 purposefully-selected extremist sites. Their findings include interesting results such as disavowal of racism and a low rate of direct incitement to violence. The authors also comment on the utility of the internet to widely-scattered extremist groups. Though their motivation is given in terms

of extremism generally, their sample appears focused particularly on a certain group of white supremacist sites, with Islamic extremists appearing only in an 'Other' category.

[243] describe how web crawling technology is integrated with NLP techniques to extract common topics from websites hosted by extremists or terrorists. Though their evaluation does include an attempt to assess the compactness of topics, a notable problem with their LDA results is the generation of several reasonably similar topics. A means of better combining (or representing the distinction between) such groups could be considered an area for study in the field of topic extraction in general.

[242] relate how a clustering opinion-extraction method targeted specifically at opinions expressed in online discussion is trialled on a corpus drawn from Myspace, including discussions about terrorism. Their clustering method attempts to overcome some limitations with the DBSCAN clustering mechanism, focusing on a distance-base clustering method. More detail on the TFIDF mechanism by which web opinions can be represented as a vector of core concepts could help generalise their method to non-clustering applications.

[114] describe a process of mining hyperlinks from terrorist web pages as part of link analysis. However, only a simulation of output from a toolkit is provided.

[102] focus on gathering open-source information about terrorism via summarisation of web-based news articles. Though some details appear obscured by poor translation, the authors seem to find support for an ontological approach to detection of terrorist events, comparing this approach to a gazetteer and some form of grammatical parser in an evaluation on Thai news articles.

Of additional interest to this problem topic and method is an information extraction tool which is designed for general intelligence use [201]. It makes use of terrorist subjects as an illustrative case study. The network study of blogging sites [237], which focuses on extremist hate groups, is also relevant.

Information extraction and police intelligence

Information extraction can be applied to online sources for intelligence on organised criminal activity. That the core of this is a collection of web-mining toolkits suggests common research interest around synthesis of web-based news articles for intelligence purposes, heavily reliant on named entity recognition techniques, with some recurring problems including the identification of relevant articles for processing and the reliance on a domain lexicon for identifying key information.

[80] focus on extracting 'story' patterns from web-based news articles as part of open-source intelligence efforts, with pattern-matching begun by the detection of trigger words indicating certain events. Their system is demonstrated on a collection of Chinese news articles on the 2008 Mumbai attacks, but what their published results show is unclear. Their system's reliance on trigger words seems to suggest that individual implementation will either require existing domain knowledge, reducing utility for emerging events, or else be general terms which may not fully capture specific narratives. The question of how appropriate news articles are gathered for processing is also critical for deploying such a system.

[127] focus on extracting crime information – in the form of key phrases – from narrative reports, which is applied primarily to police and witness reporting, but is noted for potential application to web news. Similarly to the previous study, their approach struggles with a scalable system for managing a crime lexicon, which they resolve with manually-created lists supplemented by dictionary resources.

[14] also describe efforts to gather structured knowledge from web-based news articles for EU security purposes, clustering articles based on textual similarity and geographical location, then applying other event extraction tools. There is a lack of detail on the operation of these tools.

[232] discuss an extension of the Encase forensics toolset to allow analysis of web pages regarding some form of illegal gambling activity. They attempt to mine not only entities, but also the relationships between entities, in an unsupervised manner. However,

their approach to information extraction appears to be overly tailored to their chosen problem domain for it to generalise to other scenarios. No evaluation is provided.

[201] focus on the problem of gathering *novel* information about a topic, relative to a specified set of existing knowledge. They make use of web search engine results regarding the known topic and use these web pages to form new queries based on prominent nouns, clustering the results based on descriptive nouns. Their ATHENS approach is trialled on terrorism topics, but could equally apply to other law-enforcement or defense uses. Their method for selecting descriptive nouns compares the frequency in the web pages under review to the frequency in the British National Corpus, a standard English corpus. While this approach is domain independent, strict comparisons are likely to lead to spurious noun-phrases being identified, so it would be better to search only for nouns whose frequency is significantly greater than in the reference corpus in order to prevent common variations diluting search terms. In the same vein, the BNC relies on texts now well over a decade old, and is not likely to include a number of now-common proper nouns. A different reference corpus, perhaps drawn specifically from web sources, might make a more suitable baseline.

[228] use Twitter as a source of general crime prediction, drawing on automatic semantic analysis, event extraction and geographical information systems to map crime hotspots. In an evaluation on actual hit-and-run crime data, their system outperforms a baseline uniform model. While there may be scope for improvement in the predictive technique, more interesting developments are likely to be found in modification of the model for deployment on a streaming Twitter feed. [227] do so, using Twitter data to model criminal incidents geographically. They apply a spatio-temporal generalised additive model to a combination of geographical and demographic features of an area and textual features extracted from the Twitter feed of a news agency, evaluating their performance against actual crime incidence rates. Their analysis shows that the textual features provided by the Twitter data improve prediction accuracy as compared to a previous model only using geographic and demographic information.

[82] aim at gathering information in extreme events, describing an approach to open-source intelligence which was applied in an artificial competition environment (searching for red balloons), and how experiences in the challenge may relate generally to intelligence-gathering, particularly with regard to false-reporting. Their overview is high-level and rather specific to their challenge, but includes reference to a number of techniques and technologies not otherwise captured by this review.

[64] focuses on the detection and analysis of 'dark networks', with specific focus on visualisation tools for handling networks parsed from Twitter and placed by geolocation. No formal evaluation is provided, but the paper discusses trial usage on real networks of interest.

[115] look at finding relationships between unstructured law enforcement texts (emails) and using said relationships to help augment information of interest, analysing the semantic relatedness of documents and linking identified entities. A demonstrative application is presented, acting on a sanitized corpus of real law enforcement emails. Given appropriate consideration of scalability, this information linking tool would appear to be an impressive resource for augmentation of police intelligence.

Information extraction and crimes against children

Information extraction techniques are sparsely deployed in child protection, mostly aimed at child sexual trafficking. Their aims include identifying children known to be missing by monitoring trafficking networks and chatrooms for mentions of their names or other identifying details.

[224] present an approach to combating the sexual trafficking of children through examination of open sources such as classified advertisement sites and bulletin boards. They examine such resources for evidence of trafficking networks and introduce techniques to search for victims under aliases and misspelt names. Though the authors do not present an evaluation, they discuss ongoing trial deployment, highlighting challenges specifically related to anonymisation of the toolkit's interactions with sites to prevent counter-intelligence, and with scaling their approach to wider monitoring.

[187] approaches the same problem from a different angle, applying intelligent agents to identify missing children on the internet by connecting information in open databases of missing children with web crawling and IRC chat monitoring. The approach was partially implemented as the SADIE system at the time of publication. The proposed ecosystem of agents dealing with specific data sources appears flexible, but the exact means of calculating results' similarity to a short query – a very key detail for any of the agents – is left unspecified.

A common theme to both these sexual trafficking technologies is the integration of information from multiple sources, but both publications appear to focus on different sources for their information. This may, in part, be due to the large time gap between the two papers. The SADIE system outlines a high-level approach to multiple data source integration, but leaves many implementation matters unresolved.

[142] attempt to extract crime information (Who, Where, When, How, What, Why) from chat logs, drawing on published examples of sexual abuse from adult dating and scam interactions as their data source. Tokenisation and part-of-speech tagging of the data is discussed. Classification accuracy results for their crime information categories are also presented, though how these results were derived is unclear. While mining instrumental crime information as would fit the given categories could well prove useful to investigators, the paper does not present a coherent solution for the purpose.

Information extraction and cybercrime

Cybercriminal activity has also been mined from online data sources, the primary source being fora where they sell or exchange information.

[203] uses an XML framework to mark up relationships extracted from hacker fora, essentially mapping the social network of said fora for usage by police. The paper focuses heavily on the choice of technology and representation for the task – XPath queries on tidied HTML source of the fora – with no real evaluation of the value to law enforcement. Furthermore, the approach relies on manual exploration of the XPath query

space for page sources, a process which could, at the least, have been guided through generated templates.

[257] present a study of the cybercriminal economy on the Chinese web, attempting to model the extent of this black market and the amount of malicious code involved in its constituent websites. In addition to these contributions, the authors offer detailed description of a cybercriminal infrastructure. Their estimation of the value of cybercriminal assets, and particularly their attempt at breaking down these totals by classification allows law enforcement and cyber-security vendors to focus their efforts where greatest impact can be effected.

Additionally, some papers discussed earlier fall into this category.[136] construct networks of cybercriminal activity from email and blog data while [141] describes a system designed for fighting cyberterrorism through handling of collected intelligence sources.

Information extraction and finance

[27] address a financial criminal matter (money laundering) by helping track financial services through web mining. Their tool crawls the web, identifying online financial trading sites through a generalised linear model applied to textual features. While the presented accuracy appears impressive, the results were obtained via an artificially balanced dataset wherein roughly a quarter of all websites were actually OFT (Online Financial Trading) sites – a situation which is unlikely to be the case when the system is deployed on the web generally.

Information extraction and online identification

[7, 19, 256] all cover various aspects of an email de-anonymisation workbench built on a set of mature UNIX tools. This UnMask toolkit focuses specifically on detecting and countering spoofing attempts within email messages, examining email bodies and headers particularly for examples of spoofed links, forms and headers, and storing evidence in a manner suitable for law enforcement use. Their aim of achieving this through combining

a variety of pre-existing tools is laudable for its software reuse, but there is a lack of rigorous analysis of the performance of the anti-spoofing components. The papers, however, do include a case study demonstrating potential uses of the toolkit during an investigation.

Other information extraction applications

[32] describe a study aimed at improving the analysis of forensic network traces from investigations, presenting a high-level packet analysis tool which has been developed and compared to existing tools, but not formally evaluated.

[11] cover the detection of 'suspicious' or 'deceptive' emails. They do not provide a clear definition of what that may mean, but an ominous reference to national security. They apply a series of classifiers to an insufficiently explained dataset, and report high classification accuracies, particularly for the IBk decision tree.

Machine learning and online identification

Machine learning techniques have been applied in online identification tasks, often working with email data, attempting in particular to identify scammers and phishers from their campaign output.

[8] present the ScamSlam project. It focuses on identifying the common origins of scams, particularly advance fee fraud, through the use of unsupervised hierarchical clustering on scam emails detected with a Poisson filter. Their method appears to detect a small number of scammers (20) sending most of the advance fee fraud messages in a corpus of 534 such scams, but they are unable to verify this result. It is not clear how broadly their system may be applied, as the advance fee fraud they focused on has a fairly large text body, which may be atypical of scam email.

[204] cover the same use-case, but make use of email headers to build clusters of scam originators using WHOIS data. Using this approach, they identified 12 email addresses which were key in registering spam-origin domains. Such an approach holds

benefits in that it may be applied to many scam or spam emails without requiring specific additional feature in the body of the scam, but also risks vulnerability to email spoofing.

[246] present an approach looking at profiling rather than simply detecting phishing attacks. Their study makes use of hyperlinks from the body of an email as well as structural features and WHOIS information in a pair of classifiers. They profile phishing emails by having the classifiers apply multiple labels to each email regarding the presence of scripts, images, etc. and the apparent legitimacy of linked sites from WHOIS information. The strength of the paper lies in its clear formulation of the problem of profiling phishers rather than merely detecting phishing.

[58] present a combined approach for profiling large volumes of phishing email. The results of several independent unsupervised clustering algorithms working on a random subset of a large dataset are combined with a variety of consensus algorithms and, in turn, used to train a number of fast classification algorithms for use on the whole dataset. This approach of using unsupervised clustering to prime supervised clustering would appear to work well for classification of emails into clusters detectable in the training set, but may suffer in a deployment where new clusters of phishing email begin to appear.

The other identification studies using machine learning more generally address the identification of criminals from email data. The approaches discussed below are somewhat unusual as compared to the more standard classifiers discussed in the later NLP section, but contain some overlap.

[108] make use of an existing speaker recognition framework from the field of speech processing in an attempt at authorship analysis, using several classifiers. The framework is evaluated against the Enron email dataset with results indicating a competitive approach, although the requirement of 200 training emails per author is not insignificant.

[195] explores the application of associative classification to authorship attribution of text, in an approach which requires the extraction and amalgamation of rules. However, the system performs poorly on a multi-author dataset, with a best classification accuracy of 50% on only 10 possible authors.

[205] covers authorship attribution in a trial dataset of an online newsletter. The approach uses classifier ensembles, demonstrating how a range of diverse classifiers can be constructed through exhaustive disjoint subspacing, and showing that the approach outperforms a simple SVM model using word frequencies. The author goes on to enhance the model with a cross-validated committees technique.

Machine learning and terrorism and extremism

Machine learning is deployed both in detecting terrorism-related activities and in identifying terrorists from their online footprint.

[49] explore microblogging within the terrorism informatics domain. They perform an observational analysis of the Twitter network's response to two real-life terrorist events, and use this as inspiration for the design of an information-gathering framework. They later apply the framework to a synthetic dataset of events which share some properties with terrorism events. They also apply a variety of common machine learning analyses to their dataset in an exploratory manner.

[191] link streams of Twitter data to other resources through Open Data mechanisms. They apply named entity recognition to the content of Tweets. They mention the terrorism domain, their aim being to allow for structural links to entities to be be imposed on unstructured Twitter data to better allow law enforcement to parse and respond to events detected via Twitter. However, the implementation of this is relegated to future work.

[161] evaluate a number of machine learning methods (the ID3 decision tree algorithm, logistic regression, Naive Bayes and SVM) for the purpose of detecting suspicious emails. As well as developing a terrorism-related email dataset for the purposes of this comparison (including real messages gathered from newsgroups), they develop a feature selection system that provides consistent improvement to the results of all of the tested classifiers. They report that for their application, logistic regression and ID3 outperformed the Naive Bayes and SVM classifiers.

[197] use a qualitative formalism as the basis for a fuzzy analysis, applying this to link analysis and the determination of aliases. They evaluate their system against

unspecialised unsupervised learning systems on a constructed terrorism dataset gathered from web articles, an author publication dataset (DBLP) and an email dataset. Their system appears to outperform a number of similar link-based algorithms.

[71, 72] employ web usage data to identify terror-related activities, training a classifier on the web usage of ordinary users and a collection of known terrorist web sites. The aim is to deploy a system which monitors the web access of users (at an ISP or organisational access provider level) and raises alerts whenever a user accesses abnormal content. The civil liberty implications of such a mass-monitoring system could rightly be challenged, but more practical issues may prevent adoption. Their detection system reached an AUC of 91% on their experimental dataset, rising to 99.7% with additional components. Given the large number of normal users and relatively tiny number of real, detectable terrorist usages of actual networks, even such a classifier would produce an unreasonable volume of false alerts for every true event it captured. This issue is not unique to the work of these authors, but applies to all systems of this kind. Nonetheless, these two papers consist of a coherent description of the development of a high-performance classifier for web usage data.

[218] suggest a self-organising map approach to classifying web users from usage data. They provide no evaluation or appraisal of their proposed system, and indeed minimal description of its proposed operation.

[73] have developed what they term an intelligent search procedure for webmining cyber-terrorism information, feeding a vector representation of 600 articles, half related to cyberterrorism, into a self-organising map, the results of which they then briefly dissect. Their presentation of the SOM as a heat-coloured grid seems ill-suited for law enforcement analysts.

[244] focus on identifying extremist content in social media sites, drawing their design inspiration from biological immune systems. They build a mathematical representation of lymphocytes which incorporates lexical, sentiment and syntactic features of text as a precursor to a semi-supervised classification system. In an evaluation of this system on

violent messages scraped from a white supremacist web forum, their system outperformed two benchmark labelling systems.

Machine learning and harassment

Machine learning can be deployed to detect threatening textual communications, the aim in most cases being to produce a classifier which separates threatening messages from normal communication.

Appavu and Rajaram [10] compare a decision tree classifier with SVM and Naive Bayes classifiers, using two email corpora and two different feature selection mechanisms (information gain and term frequency variance). They find decision trees to outperform SVM and Naive Bayes in detecting examples of threatening email. A follow-on paper [12] repeats this analysis, but includes the Ad Infinitum algorithm, which outperforms the other methods. [21] later revisit this work, looking also at the detection of threatening emails. The authors compare the data of Appavu and Rajaram to their own Naive Bayes approach, which makes use of different features (single and multiple keywords as well as weighted keywords with context matching). Measuring the accuracy of results with the F1-score rather than simple percentage accuracy, they find that their weighted multiple keyword system with context matching performs in a manner competitive with the better methods from Appavu and Rajaram's analysis. They do not make a direct comparison due to the different datasets underlying results, but a review of F1 scores indicates that some of the methods presented by Appavu and Rajaram may be better classifiers.

[239] identify emotions common in cyber-bullying, and develop a training procedure to help recognise these emotions from text without reference to a labelled training set. They evaluate their zero-label-trained SVM system on a labelled Wikipedia corpus, finding that it has lower cross-validation error than three baseline methods. They also apply it to Twitter traces involving bullying, finding that only a relatively small proportion of said traces showed emotion, and that where emotion was detected it did not necessarily reflect severity or sincerity. [247] instead focus on a supervised learning approach to detecting cyber-bullying, using term frequency as a primary measure, and supplementing

it with sentiment and contextual features. Their model performs fairly poorly on their web datasets, with the best F1-measure accuracy being less than 50%.

[225] take inspiration from biological immune systems in much the same manner as [244], also integrating term frequency into their mathematical adaption of it. Though they claim 'good results', there is no evidence of any evaluation.

[196] turn to a more conventional Naive Bayes classifier, testing it on a small corpus and a bag-of-words feature set which appears to be extended with some user-level attributes. The presentation is somewhat ambiguous, describing classification rules for detecting a 'threat' class of message, but presenting classification results for 'movie' 'food' and 'travel' topic classes, none of which are alluded to a-priori.

Machine learning and crimes against children

A small number of rule-based systems have been generated to help with detecting predators in textual exchanges.

[93] present a knowledge-based system for detecting sexual predators, with a Naive Bayes subsystem with reasonable classification accuracy. Interestingly, their hand-coded rules for predator characterization were originally written in and for Spanish, but were automatically translated to apply to English, and appear to still be effective in identifying the main predation phases.

[144] compare previously-developed rule-based classifiers to decision tree and a k-nearest neighbours classifiers. They find that the machine learning systems improve classification of predation when working with specific transcripts, but fail to reject the null hypothesis in a more general case comparison against their rule-based system. The average accuracy of their rule-based classifier is 68%.

[174] move away from rule-based systems, applying and combining two separately-trained SVM classifiers in a weighted manner. They achieve an F1-score of 0.9 for the task of classifying authors as predators, but much lower accuracy for detecting specific grooming posts.

Machine learning and financial crime

[148] use SVMs and Random Forests to detect advanced fee fraud scams in an email dataset. They report high classification accuracy on a synthetic dataset where roughly a third of all mail was advanced fee fraud messages, and find that their SVM classifier outperformed the Random Forests classifier. An evaluation more comparable to real deployment base rates would be preferred.

[223] focus on click fraud prevention using web usage data. They detail a multi-level data fusion mechanism which takes input from a click map module, an outlier detection module and a knowledge-based rule module, and stores levels of suspicion regarding specific IP addresses, referrers and countries. They provide a detailed analysis of the results of their system as applied to publicly-available click-through data.

[27] track online financial services through web mining, gathering textual features to reach conclusions about the probability of a site being an online financial transaction site. Their evaluation against human subjects shows demonstrable benefits in terms of speed, and generally high precision.

Machine learning and police intelligence

[51] studies appropriate machine learning systems for categorising temporal events collected from web data. Using a case study involving web articles related to an incident of domestic terrorism, the performance of Naive Bayes, SVM and neural network methods at applying temporal group labels across a range of feature set sizes is demonstrated. The results show that while all three systems performed in a satisfactory manner, SVM and Naive Bayes increased in accuracy as the number of features increased, while the neural network peaked at 70 features.

[211] describe the application of a general-purpose email-mining toolkit to behavioural analysis, with a case study in detecting viral emails in an archive of the emails provided by 15 users. The system performs well when introduced to sudden and abnormal flows, but struggles to detect slow campaigns for the delivery of email. The degree to which virality can sensibly be detected from such a small user corpus is

debatable. The paper also provides a lengthy demonstration of the overall capabilities of the email-mining toolset.

Other machine learning applications

[62] approach the problem of pornographic web page identification with two classifier components. One component classifies web pages into various predefined categories, which can then be used to filter these web pages from access. The other component analyses the behaviour of users with respect to the category of sites accessed. They test their system – with a variety of classifiers – against commercially-available web filters, and find that their best classifier outperforms them.

[217] focus on the effect of pornography on young people as their motivation. They note failures of strict rule-based and keyword-based systems for filtering undesirable information, and propose a system which gathers a broader range of features from a page to assist in classification. They do not evaluate the performance of this proposed system.

[132] attempt to detect abnormal patterns of email traffic using a hierarchical fuzzy system. They develop three different system architectures, and trial these systems on a selection of threads from the Enron email dataset, finding that all three agree with each other in the ranking of abnormality of communication links. Whether such a test holds external validity is hard to determine.

[25] apply machine learning to recognise the traits of key actors in hacker communities. Their regression analysis of the social structure of hacker fora from the United States and China, determines that involvement in a number of threads, total message volume and number of attachments uploaded are the major factors which explain the reputation score of members of the community. While the paper focuses on cybercrime as a domain, its results could be said to apply more to certain online forum communities, of a criminal or otherwise nature.

Authorship attribution and online identification

Authorship attribution is naturally tied closely to the problem domain of online identification, and a wide range of techniques have been applied on a number of datasets. Typical problems for such studies include deciding upon the most appropriate feature set to use in classification, and finding appropriate methods for different data sources. Other for or highly related to this field include authorship verification, authorship similarity detection and stylometric comparison, with different clusters focusing on either the conflation of author identities or the assignment of specific texts to an author, but the technical challenges of both tasks are expressible within the same framework.

[138] apply authorship attribution specifically to phishing emails, aiming to cluster messages based on othographic features using an adapted form of the K-means algorithm. They reason that the semantics of phishing emails are often too similar to be useful for disambiguation. They provide an evaluation on a collection of 2048 known phishing emails, with several differing initial parameters for their clustering algorithm and gradually refined feature sets. While their method appears to produce reliable clusters, a validated dataset would be useful for verification purposes.

[59] make use of both structural and linguistic features and an SVM classifier. They validate their approach on a collection of emails to particular newsgroups, finding high accuracy in most cases. They additionally investigate the use of word collocation and the dimensionality of function words in a bid to improve classification accuracy. However, this does not improve performance.

[252] compare decision trees, neural networks and support vector machines on a corpus drawn from English email messages and both English and Chinese BBS postings. The best results are for SVM classification of the English newsgroup postings, with neural network performance lagging slightly behind. They note a drop in performance in their Chinese dataset, which they ascribe to fewer style features for that language. A follow-up paper [251] makes use of an extended set of features and the same set of classifiers, again finding that the SVM classifier outperforms the C4.5 decision trees and the neural network.

[53] also covers the reduction of authorship attribution to a pattern of certain writing features, applying this approach with an SVM classifier to public-domain books, theses and the author's own email collection, with good results in each case. The results show that function words appear to make the best features.

[216] suggest the use of an SVM classifier for authorship attribution on emails, explaining the operation of the classifier and listing some structural features of email which might be useful, but providing no evaluation. Given the prior existence of work such as [59], this would appear to be of at best explanatory value.

[205] explores authorship attribution via an ensemble of SVM classifiers and a feature set subspacing approach. Exhaustive disjoint subspacing is compared with the k-random classifiers method of ensemble construction, finding that the former outperforms the latter and also outperforms an SVM classifier when small subset sizes are chosen.

[206] covers the class imbalance problem in authorship attribution, where the volume of available training text for some candidate authors is extremely low. A new method for handling imbalanced datasets through variable-length sampling of training data is presented. The method is compared against a re-sampling variant to the existing under- and over-sampling methods, making use of both English and Arabic datasets. The results show that the method resulting in the best net improvement to the accuracy of an SVM classifier trained on the resulting training set is the random re-sampling of text from the available training data.

[3] provide a useful review of the state-of-the-art and go on to demonstrate a classification method which makes use of individual author-level feature subsets from a large feature space. They compare this method to an SVM classifier with a feature set drawn from previous literature and to an ensemble of SVM classifiers with an extended feature set, using a range of online text forms (the Enron email dataset, eBay comments, posts from an online forum and chat logs). Their system outperform both competitive methods on the email, comment and chat datasets, but not on the forum messages, where the ensemble of SVM classifiers performed best. Alongside the identification experiment,

they also distinguish the task of detecting similarity, and perform a similar evaluation for that purpose, finding their method outperforms the competitive baseline methods.

[56] focus particularly on blogs, covering the ethical debate over why bloggers may legitimately seek anonymity, and why law enforcement may wish to circumvent this barrier of anonymity. The paper covers technical approaches to stylometry only briefly and at a high level. [63] explores the same topic with more technical detail, creating a baseline model of authors based on frequency of characters and words, and using individual deviation from this baseline as the features for classification. Both Naive Bayes and SVM classifiers are evaluated, finding low average accuracy across all authors, but that certain authors were extremely well-predicted.

[135, 137] focus on applying authorship attribution to Chinese online texts. The first paper focuses on authorship attribution in email, covering issues such as the lack of explicit word boundaries in Chinese text and the selection of sequential patterns from texts, passing said patterns to an SVM classifier. In their evaluation they provided 30 training examples for three authors, and had 20 further emails classified as belonging to one of these three authors, with a classification rate of 90%. In the second paper, the authors also apply their classifier to blog and BBS messages, drawing a comparison between three classifiers, one of which uses linguistic features, another which uses structural features, and one which uses both. They find the classifier using the combined feature set outperformed the others, though they all performed at above 65% accuracy. They also examine the effect of varying the number of authors to classify texts, finding that larger numbers of authors caused accuracy to drop.

[104–107] all address authorship attribution through frequent pattern mining. The first of the publications focuses on the notion of frequent patterns as a means of ensuring the forensic worth of authorship attribution techniques, objecting to the lack of intuitive explanation in an SVM classifier. They use a combination of lexical, syntactic, structural and content-specific features in their method, detecting frequent writing patterns in an author's text and filtering out frequent patterns which are common to a large number of authors. They validate the viability of their method in an evaluation on the Enron

email dataset. In the second publication, the authors use standard clustering algorithms to group texts together as a prerequisite to mining frequent patterns for author identification. They examine the accuracy of the resulting output as a means of evaluating which clustering mechanism is best-suited to the task, again using the Enron email dataset as a source. The third publication presents frequent-pattern writeprints as a 'unified' solution to authorship analysis. The authors describe use cases involving small and large training samples, and also extend their system to discovering characteristics of an author. The evaluations on the Enron email dataset are repeated, and alongside these results a trial of the characterisation application is carried out. The results show that for gender prediction the approach performs slightly better than random assignment, and for location prediction, with three classes, it again performs with accuracy slightly above that which one would expect for random assignment. Finally, [104] combines this research into one volume, providing greater detail on the difference in approach between two versions of the classifier, with extensions covering somewhat separate problems of extracting cliques and topics from chat logs.

[164–166] cover authorship analysis on instant messaging communications. The first publication focuses on examining character frequency as a stylometric feature, examining the frequencies of characters in a small four-author dataset and testing for whether frequency of characters is distinct. The results show that uppercase characters, numbers and special characters are distinguishing and may be used as a form of intrusion detection system. In the second publication, the authors analyse what appears to be the same dataset, but with an extended range of features, including sentence structure and pre-defined sets of special characters. They apply three classifiers – the J48 decision tree, the IBk nearest neighbour classifier and a Naive Bayesian classifier – to these features, and find a high accuracy in each case, though given the sample size this would not be unexpected. They analyse the distinguishing features and find that abbreviations are the best discriminators, followed by the use of special characters. In the final publication, the authors expand their evaluation to include two larger datasets, examining the useful features for accurate classification in each system, and using different classifiers. They

find high accuracy with an SVM classifier trained on a range of 356 features, including lexical and syntactic features as well as the previously-used structural and frequency attributes. On a dataset of 105 authors, they achieve 84.44% accuracy.

[130] cover a particularly constrained form of authorship attribution which is particular to online discourse – attribution of Tweets to their authors. They detail the structural properties of Tweets and present a preliminary analysis of the viability of attribution using Tweets. They find classification accuracy of approximately 60% for 20 training examples, and highlight that adding training Tweets increased accuracy up to 120 examples, after which increases appear not to be significant.

[46] apply a frequent-pattern mining approach on the Enron email dataset. Writeprints – consisting of numeric representations of the relative frequency of stylistic features extracted from an author's text – are constructed and then compared in order to determine whether authors are similar enough to be the same. They compare SVM, PCA, K-NN, DT and K-means approaches, finding SVM to have superior classification accuracy.

[118] uses a bag-of-words model of email bodies and applies a Naive Bayes ensemble method to attribute emails drawn from the Enron email corpus. The method chieves a respectable classification accuracy, outperforming previous work on the same dataset, but it does perform best when given messages over 100 words, which slightly limits application to online texts.

[172] cover the detection of 'authorship deception', which includes both a normal attribution use case and an imitation attack whereby authors attempt to imitate the writing style of a victim. Their method involves building a writeprint of stylometric and content features, and applying logistic regression as a classifier. Evaluation on a blog dataset shows good performance in the classic attribution case, and a small evaluation of the imitation case shows highly positive results.

[134] attempt to address issues with the difficulty of writeprint comparison through a novel semi-random subspace method, which also aims to overcome redundancy in feature sets. A detailed description and theoretical analysis of the methodis provided, followed by an empirical evaluation on a subset of a large English corpus, displaying

accuracy results with regard to both the number of authors and the number of texts available per author. In all cases they compare their method to other well-performing classifier ensembles, with a positive result.

Given that the aims and methodologies of many authorship attribution papers targeting identification are comparable, results and methods from various studies may be contrasted with each other. Table 2.2 gives an overview of some of the best results from each paper, noting the dataset and number of classes being attempted. Important additional information including length of texts used, texts per author, and features used in classification are all left to the original texts. Items are sorted chronologically.

Authorship attribution and cybercrime

[89] cover the application of authorship analysis techniques to software source code, demonstrating how the means of expression can vary even when programmers are solving the same problem. Their motivation is the attribution of malicious code, just as in natural language analyses the application is attribution of malicious or incriminating messages. They identify a number of features which could be useful for authorship attribution, and present two short case studies of events where malicious code has been examined for attribution to its owner.

A number of other papers focusing on online identification of authors cited cybercrime in a general sense, but made no specific link to the domain as defined in this study, and hence have not been included in this analysis.

Authorship attribution and terrorism and extremism

[2] discuss authorship attribution with particular application to the forum postings of extremist organisations, with a focus on selecting an appropriate feature set for classifying Arabic text. They describe a study using SVM and C4.5 classifiers, applied to both English text from a Klu Klux Klan forum and Arabic text from posts associated with the Palestinian Al-Aqsa Martyrs group. They find slightly better performance at classifying

Table 2.2 Summary comparison of authorship attribution approaches

Source	Classifier	Dataset (#Messages)	Authors	Accuracy (%)
[59]	SVM	Newsgroup (1,259)	4	-
[53]	SVM	Email (253)	4	85.2
[252]	SVM	Newsgroup (153)	9	96.08
[252]	SVM	Email (70)	3	91.43
[252]	SVM	BBS (70)	3	82.58
[251]	SVM	Newsgroup (c.960)	20	97.69
[251]	SVM	BBS (532)	20	88.33
[205]	EDS Ensemble	Web news (200)	10	99
[107]	AuthorMiner	Enron Email (120)	6	90
[107]	AuthorMiner	Enron Email (100)	10	90
[3]	Writeprint	Enron Email	25	92
[3]	Writeprint	Enron Email	50	90.4
[3]	Writeprint	Enron Email	100	83.1
[3]	Ensemble	eBay comments	25	96
[3]	Writeprint	eBay comments	25	96
[3]	Writeprint	eBay comments	50	95.2
[3]	Writeprint	eBay comments	100	91.3
[3]	SVM	Java forum	25	94
[3]	SVM	Java forum	50	86.6
[3]	Ensemble	Java forum	100	53.5
[3]	Writeprint	CyberWatch Chat	25	50.4
[3]	Writeprint	CyberWatch Chat	50	42.6
[3]	Writeprint	CyberWatch Chat	100	31.7
[135]	SVM	Email (150)	3	90
[137]	SVM	Blog (1,379)	7	89.49
[137]	SVM	BBS (410)	5	73.97
[137]	SVM	Email (95)	5	80.61
[63]	SVM	Blog (c44,000)	100	54.53
[166]	NB	IM (?)	4	99.29
[165]	SVM	IM (950)	19	88.42
[165]	SVM	CyberWatch Chat (1250)	25	84.44
[130]	SCAP	Tweets (100,000)	50	72.9
[105]	EM	Enron Email (200)	5	80
[105]	K-m	Enron Email (200)	5	88
[105]	Bisecting K-m	Enron Email (200)	5	83
[106]	AuthorMiner2	Enron Email (160)	4	$x \approx 90$
[106]	AuthorMiner2	Enron Email (800)	20	$x \approx 70$
[46]	SVM	Enron Email (750)	25	88.31
[118]	NB	Enron Email (6,109)	10	86.92
[118]	NB	Enron Email (5,799)	9	87.05
[134]	PSemi-RS	Text corpus (2500)	50	77.04

English text authors than Arabic authors, and note that SVM significantly outperformed C4.5, going on to dissect the important features in classification for both languages.

Author profiling and crimes against children

Author profiling is widely deployed in the detection of sexual predators from chat transcripts. The typical classification is between text written by a child and text written by an adult. Some problems arise from attempting to parse net-speak with traditional linguistic tools, although there are indications that use of such language can itself be a useful age-determining feature.

[175] applies SVM and k-nearest neighbour classifiers to binary classification of predator and victim in chat logs from a vigilante website. They make use of word n-grams, with minimal pre-processing, as their input to a feature extraction function. They find their best classification rate (94.3%) comes from the k-NN classifier with a k of 30 and 10,000 features, both inputs being the largest of various levels tried.

[215] detail a method for distinguishing between teen and adult conversations, with application to detecting sexual predators. Using a chat corpus, they attempt to distinguish between teens and chat users of different age brackets, using word and character n-grams in a Naive Bayes classifier and then an SVM classifier, finding that the SVM classifier outperformed Naive Bayes. Unsurprisingly, the most difficult-to-distinguish age group were the authors in their 20's.

[122] compare a rule-based approach to log classification to a human analysis. They describe how a new iteration in a rule-based analysis of chat logs differs from a previous version of the tool in more appropriately identifying combinations of keywords in chat lines. The inter-coder reliability between their new tool and human analysis is reported as much improved. A follow-up work by the same authors [123] surveys the literature regarding both sexual predation and cyberbullying, and as such covers some work on profiling sexual predators.

[146] focus on detecting a grooming author by classifying messages into one of three attack categories and then combining classification probabilities. They perform

a comparative evaluation with a number of different classification algorithms, finding SVM to perform poorly next to k-NN, Naive Bayes, Maximum Entropy and Expectation Maximisation. They consider their Naive Bayes approach the most suitable.

[173] aims at predicting both age and gender of chat authors, with application to checking the truthfulness of reported profiles on social media sites. Working with a corpus drawn from a Belgian social network, they discuss several issues particular to online chat corpora, including shortness of texts and the variability of Dutch net-speak. They avoid issues of stemming and more involved linguistic analysis by utilising word and character n-grams. They find that word unigrams are the features best used for distinguishing between age and sex categories, achieving good accuracy in both cases.

[100], is notable in covering both author characterisation and topic detection, addressing general text-based surveillance. They describe a short characterisation experiment wherein Twitter users who are informative for a particular topic are identified, using a broad feature set and the expectation maximisation clustering algorithm.

[29, 30] focus on two different sub-problems in the identification of sexual predators. The first draws on the concept of fixated discourse – that predators will return to the subject of sex throughout grooming conversations. The authors apply a sentiment similarity measure to lexical chains identified from text. They hypothesise that long lexical chains related to sex are indicative of authors being sexual predators. They find some evidence for this in a comparison of the length of sex-fixated lexical chains in both a sexual predator corpus and a cyber-sex corpus. The second publication turns to the use of sentiment and emotion features in conversations involving sexual predators. The authors identify from related work that several sentiment features are linked to sexual predation, and construct a feature set based on sentiment markers. This feature set is compared to a number of simple character and word-based feature sets in a Naive Bayes classifier running over a corpus of chat logs from a vigilante website combined with ordinary cyber-sex logs.

[101] cover the 2012 International Sexual Predator Identification Competition, detailing a common evaluation framework against which 16 methods for identification

Table 2.3 Summary comparison of author profiling approaches applied to crimes against children

Source	Classifier	Dataset (#Documents)	Task	Accuracy (%)
[175]	SVM	PervertedJustice (1,402)	Victim/Predator	90.8
[175]	k-NN	PervertedJustice (1,402)	Victim/Predator	94.3
[215]	SVM	Lin2006 (2,161)	Author Age	78.6
[215]	NB	Lin2006 (2,161)	Author Age	69.8
[173]	SVM	Netlog (1,537,283)	Author Age	66.3
[221]	SVM+NN	PervertedJustice	Victim/Predator	93.5

of sexual predators could be evaluated in a comparable manner. The competitors were provided with a sample of 30% of a synthetic dataset constructed from a vigilante site and publicly-available IRC logs, and evaluated based on the F-score of their method's performance on the remainder of the set. The paper provides an overview of participants' approaches as well as their results. A more detailed account of the method used by the winning competitor [221] is also included in this review, as are the methods used by two other competitors [152, 174] who were placed 4th and 6th respectively. A master's thesis [151] by the author of the 4th-ranked paper provides additional detail on their method's unsuccessful behavioural analysis addition to an SVM classifier using unigram and bigram features.

Some of the approaches taken in author profiling in this domain can be broadly compared to each other, so the summary comparison provided in Table 2.3. The two types of task attempted are distinguishing predators from their textual contributions and determining the age of authors as part of such a system. Note that many important details on the precise features and processes used in classification are best explained in the original publications, and note also that the figures given for author age classification are figures focused on general child-versus adult classification, and the results for more specific age groups vary from this figure within publications – the general effect being that older adults are easier to distinguish from teens and children. Additional comparable approaches can be seen in the results reported by [101], the figure for [221] being their (the topmost-ranked) performance on that evaluation.

Author profiling in terrorism and extremism

[79] use a vocabulary of group membership markers to rank documents by the degree of militancy of the author, with the aim of building more efficient search tools for such material. Working with a corpus of white extremist websites, they find that these hand-selected features, when weighted by TF-IDF, correlate more closely with human rankings of militancy than full feature-sets or feature-sets selected from those words with the highest mutual information. They also outperform a variant using weights based on a cosine similarity measure. They also find that an SVM using TF-IDF and the full vocabulary performs best at classifying texts as militant.

Author profiling in threats and harassment

[47] aim to profile users who are likely to send out abusive messages. They do this through a combination of lexical and syntactic features and independent sentence offensiveness measures which draw upon a form of sentiment analysis. They distinguish the degree to which profanity determines offensiveness and produce some tailored rules for identifying name-calling. Their evaluation against manual markup of 249 Youtube users' comments shows high abusiveness classification accuracy.

Other author profiling applications

[179] do not make specific reference to a particular type of crime – beyond a generic formulation of 'cybercrime' – but undertake a range of author profiling, ambitiously attempting to derive not only age and gender information, but also the occupation of the author. They gather a corpus of well-used Vietnamese blogs, and run a large number of classifiers in a comparative evaluation for each classification task. They find, for the most part, that an IBk decision tree algorithm is best-performing, with the exception of the occupation classification, which is best served by a random forests classifier.

Sentiment analysis and terrorism and extremism

[1, 4] explore sentiment analysis on US and Middle Eastern web forum postings. In the first paper, the authors focus on the detection of emotions or affects in web-based discourse. The authors manually construct a lexicon mapping terms to a score of intensity in a particular category of sentiment. They proceed to a case study comparing the US and Middle Eastern extremist groups based on the intensity of the hate and violence intensity of their postings, finding a linear relationship in both cases and a strong one in the case of the Middle Eastern groups. In the second paper, the authors move towards automated sentiment classification of English and Arabic content. They use a range of stylistic and syntactic features and make use of a genetic algorithm to aid in feature selection. SVM classification using this system performed well at classification on a benchmark movie review dataset and on manually tagged English and Arabic forum postings.

[245] study detection of radical opinions in web forum postings. They particularly focus on detecting the features which are most relevant to such a specific form of classification task, working with a full range of lexical, structural, syntactic and content features. They validate their method on two U.S. based hate group fora, and find that a choice of lexicon for context is highly important. Their experimentation with a number of classifiers found SVM to outperform Naive Bayes and Adaboost.

[186] provide a very high-level description of mining of online information about agro-terrorism, providing no detailed implementation steps nor evaluation.

Sentiment analysis and threats and harassment

Sentiment analysis has a particular role to play in detecting threats and harassment in text, due to its ability to detect the tone of conversation. It has been applied with some success to posts in both online fora and social media.

[202] study the dynamics of political discussions on Polish internet fora, drawing on them as a source of strongly bipolar exchanges. They perform a topological assessment of the discussion network, and undertake a detailed analysis of the nature of user interactions and thread popularity based on political affiliation of participants. They note a connection

to analyses of hate groups, and contradict existing understanding of contrasting views leading to averaging of opinion.

[229] focus on detecting hate speech on the web, discussing issues with clearly defining hate speech — such as distinguishing reclamation or discussion of racial slurs from their offensive deployment. They perform a manual coding of hate speech related to Jews, and compare an SVM classifier using a number of feature sets to this ground truth, finding acceptable classification accuracy on a unigram feature set.

[182] cover cyber-bullying, detailing the design of a tool to help assist parents and school personnel in spotting malicious online posts. Drawing on a dataset of manually-gathered cyber-bullying instances, they perform a comparative affect analysis to distinguish the degree of emotion associated with cyber-bullying texts, drawing on an existing affect analysis framework with emoticon support. They found that there were not notably more emotive items in the positively labelled set, but that there were significantly more vulgarities. Interestingly, they also found evidence of sarcasm in their bullying dataset, with the category of 'fondness' ranking unexpectedly high. Based on this analysis, they build a machine learning system to be integrated into a web crawler for classifying malicious posts.

[238] present social media as a valuable resource for facilitating academic study of bullying, and highlight a number of key challenges for the NLP community to overcome, using Twitter as a source for example data and a number of exploratory analyses. Their detailed exploration of the topic is a broad starting-point for researchers to expand on.

Text classification in crimes against children

A number of publications focus on estimating the volume of child abuse media in filesharing networks, identifying files which may contain child abuse based on their filenames. A critical component behind many of these approaches is the collection of appropriate keywords to identify in filenames, these terms being drawn from a specialised vocabulary used by sharers of this media.

[208] focuses on evaluating the volume of child abuse material on the Gnutella network based on a keyword-based evaluation of filenames and search queries, also investigating a number of common claims regarding characteristics of such material on peer-to-peer networks. They find that just under 1% of queries and 1.45% of files were related to child abuse material.

[181] specifically attempt to identify the path to the use of child abuse material, presenting results drawn from a three-month study of the isoHunt filesharing network's top 300 search terms, where 3 of 162 terms were linked to child abuse material.

[169] aim at building an automatic classifier for child abuse material based on filenames. As an early step, their SVM-based and logistic regression-based classifiers are trained and evaluated on pornographic filenames as a proxy, with promising initial results. Other work by the same authors [168] presents more detail on the implementation of the filename normalisation and classification procedure, but no new results evaluating the viability of their classifier in distinguishing child abuse material filenames from adult pornography.

[76] perform a comparison of paedophile activity in KAD and eDonkey, two distinct filesharing networks. Using an existing classification tool to label queries, they find that eDonkey contained more child abuse-related queries (0.25%) than KAD (0.09%).

This collection of studies, though working on different networks, tend to arrive at similar results regarding the extent of sharing of child abuse material, with a small but significant percentage of a number of filesharing platforms appearing to contain child abuse material.

[177] focus on detecting predation – rather than predators – in chat logs taken from online games such as World of Warcraft. Their method uses a keyword-lookup system whereby suspicious messages are those which reveal personal information. A small trial evaluation found that their system highlighted two synthetic suspicious messages inserted into ordinary chat logs, though it would also appear that a large false-positive rate is inherent to their approach.

Text classification in terrorism and extremism

[199] focuses on detecting related messages, using a form of term frequency analysis to correlate and cluster messages using certain unusual words. Their focus is on detecting groups such as terrorists, that are aware of being monitored by keyword systems and are thus unnaturally altering their word usage. They demonstrate their approach on a synthetic dataset. Stronger demonstration that the word usage behaviour expected exists in communication traces would help validate the approach. [75] similarly aim to detect word substitutions in messages, a measure which might be adopted by those seeking to avoid keyword-based surveillance. They draw upon a range of weak sentence oddity indicators which, combined in a decision tree classifier, achieve good classification accuracy for sentences drawn from the Brown and Enron corpora where a noun has been replaced with another noun with a similar frequency.

[200] applies a number of word-usage models to posts on an English-language forum, drawing on measures of radicalisation and deception to rank forum posts and providing some analysis of the distribution of posts. They find that highly-radical posts are ranked low for deception, signalling sincerity.

Text classification in police intelligence

[212] focus on the prevention of drug abuse through monitoring social media. Their framework is designed to identify the popularity of posts within specified topics. Specifically, they focus on the prediction of comment arrival as a proxy for popularity, finding good results in an evaluation on Twitter and the Hong Kong Discussion forum.

Text classification in threats and harassment

[10, 12], and [21, 196, 225] cover building classifiers to detect threatening emails, all having been covered together under machine learning applications to harassment above.

Text classification in cybercrime

[39] appear to address cybercrime – though not clearly in the sense the term is used in this review – applying Naive Bayes, C4.5 and SVM classifiers to the somewhat ambiguous question of deciding whether or not texts are useful to cybercrime investigations. A trial on manually-coded case descriptions from a United States Department of Justice website suggests that all three classifiers have acceptable performance, with Naive Bayes the best-performing.

Text classification in finance

[230] focus on copyright infringement, sampling the BitTorrent network to gather information on the number of shared files, assigning files individual categories, and then checking a random sample of filenames manually to determine how many files appeared to contain copyright-infringing material. They find that the vast majority of shared files contain infringing content.

2.1.3 Summarised results

Within the 206 papers reviewed, there were 8 broad problem topics that papers sought to address:

1. Financial crime, which relates to fraud or crimes like copyright infringement whose principle damage is economic. Financial criminal activity does not always leave visible traces in online data sources, as much financial information is kept private. However, a number of specific areas are visible. Primarily there is the example of copyright infringement, one of the more widespread criminal activities visible online, can be examined via a number of public interfaces, not least the P2P filesharing mechanisms often used to commit it. Another online lens into financial crime comes via online auction sites, whose transactions are to some extent available to the public for scrutiny. Finally, when an allegation of fraud

is being investigated, the email records of suspects can indicate collusion and/or implicate co-conspirators.

2. Cybercrime, which is intended to cover crimes focused on information systems. While a majority of cybercrime will take place online, and leave traces in data sources such as firewall and server logs, much of this category of crime is excluded from the study, as it involves a vast body of work in intrusion detection and similar fields. The works which were considered in the scope of this study which dealt with cybercrime mostly focused on the social and economic background to cybercriminal activity, often mined from online fora where criminals share or sell information.

3. Criminal threats or harassment. The type of threat dealt with in this category ranges from the identification of serious bomb and murder threats in messages, to filtering instances of 'trolling', where the aim is merely to provoke shock. Identifying such messages has proven more difficult than the identification of spam mail, due to the varied possible representations of threats, and some form of sentiment analysis may prove critical to any solution to this problem.

4. Police intelligence — the creation of tools to support government or law-enforcement in the general detection of crime – is one of broader problem categories when it comes to criminal acts addressed. Generally speaking, the interest is in either the investigation of criminal organisations or some spatially-restricted prediction of crime, but other minor crimes are also addressed. A large body of this work aims to augment police investigations by filtering knowledge from web-based news articles, the intent being to provide situational awareness and keep investigators abreast with public information.

5. Crimes against children, including grooming and child trafficking. Online grooming of children has become very high-profile, and a number of publications focus on means of detecting it – either by identifying the age of a conversational partner or by directly modelling predatory behaviour in instant messaging conversation.

Other online data is also examined in relation to these crimes, most significantly filesharing networks which are often used to distribute images or videos of child abuse.

6. Criminal or otherwise links to extremism and terrorism. The specific nature of the problem addressed is entirely focused on either white supremacists from the United States or else Islamic fundamentalists. In both cases, the primary online lenses into the groups are the online fora they use to discuss matters pertaining to their ideologies, and much of the research effort is in examining their social networks and analysing the persuasive techniques they have employed.

7. Identification of online individuals in criminal contexts. The Internet being a theoretically anonymous medium, a critical issue for many criminal matters is identifying a person. Given the highly-textual nature of much online activity, means for identifying a person from their writing dominate this problem, but data sources can be diverse, including email, online posts, instant messaging logs and even images.

Additionally, some papers made reference to criminality in a broad sense, but appeared not to address specific crimes or categories of crime. These were labelled as 'Unclear'. Some papers were labelled as addressing multiple problem topics. The most prominent problem topic was online identification — broadly, the problem of identifying individuals based on only online data, a problem which was particularly related to the analysis of malicious emails and language-based classifiers. This topic was closely matched numerically by those papers addressing extremism or terrorism. Generally, the crimes most often focused on were terrorism or extremism related, or else linked to crimes against children.

Also identified were five broad classes of method common to several papers, including different types of online data being gathered and analysed. Of the method categories, the largest was natural language processing (NLP), with machine learning (ML), information extraction (IE), social network analysis (SNA) and computer vision (CV) falling far

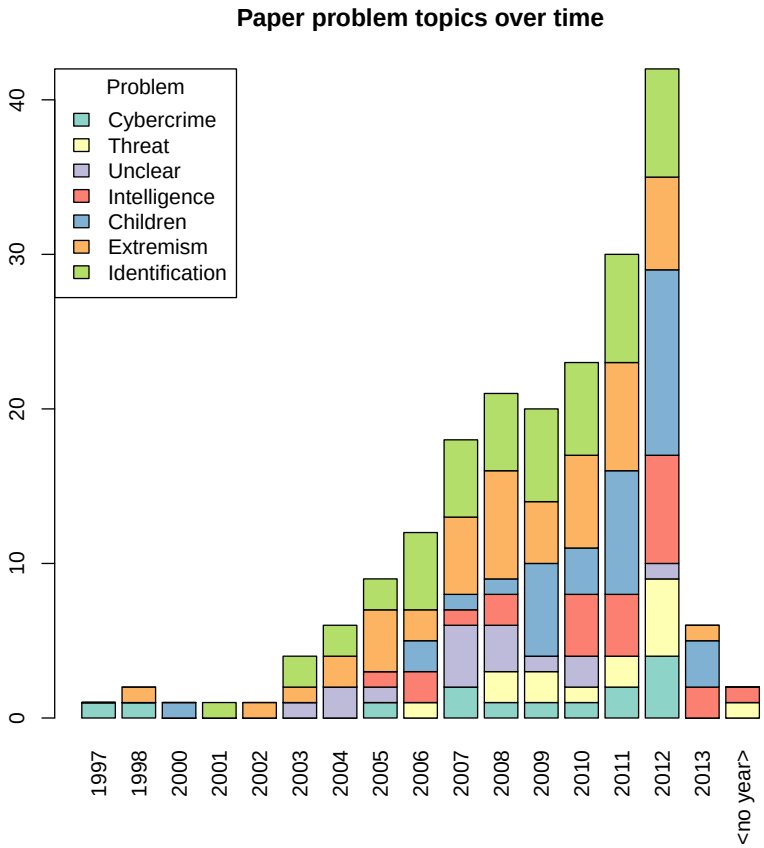
behind. Some papers did not fit into these five categories neatly, so a miscellaneous ('ETC') category houses them. The NLP subsection, due to being much greater in size, is broken down into Authorship Attribution (AA), Author Profiling (AP), Sentiment Analysis (SA), Text Classification (TC) and Other Methods (O). Data types observed included web page and forum contents, including data from social networks, email data, instant messaging data and network traces.

The first questions which may be answered are the guiding research questions previously outlined.

What are the problems (crimes, investigative requirements) which are being addressed in the literature?

As can be seen in Figure 2.1, a number of high-impact crimes such as terrorism and the sexual predation of children are prominent topics, alongside more broadly applicable aims such as identification of offenders using online data.

Fig. 2.1 The most common problem topics over publication years



In addition, Figure 2.1 shows the most common problem topics over time. Note that individual papers could fall into multiple topics. It can be seen that topics such as the identification of internet users and the investigation of terrorism or extremism remain relatively stable (as a percentage of research output) over time, while the attention to crimes against children appears to have increased since around 2009.

The problem of online identification was most often associated with NLP approaches to uncovering the author of a given written text, an aim relevant to legal debates about incriminating texts such as emails or blog posts. The authorship attribution literature connected to this aim appears to be rich and mature, with a number of comparable studies. This topic is extended somewhat in combined NLP and ML work – sometimes including computer vision techniques – which aims to cluster spam or phishing email campaigns to identify common origins, and similar aims motivate more traditional IP-lookup approaches. A very much distinct body of research by A.A. Mohamed and R.V. Yampolskiy also addresses identification, their focus being on approaches to identifying people via online avatars in virtual games [149, 150].

Papers addressing extremist or terrorist problems almost uniformly apply themselves to investigating and monitoring online communities as part of information-gathering efforts. Several studies look at the links between different sites and communities, while some others look at means of identifying the most radical members of groups where discussion is visible. The two key demographics targeted are groups linked to Jihadist terrorism and far-right extremist groups in the United States, suggesting a U.S.-centered publication bias. The predominant trend in the Cybercrime publications was also investigating and monitoring online communities, which suggests that there may be a binding theme of investigating online criminal communities.

There are two main categories of crimes against children visible in the reviewed publications. The first, and most common, is the detection of sexual predators engaged in online conversation with children, the aim being to detect attempts at grooming children for contact, a problem which by its nature draws heavily on NLP approaches. The second is the detection of child abuse material, which includes both CV attempts to discern such

content from images and videos and filename-based attempts to quantify the volume of such content on a number of P2P filesharing networks.

Financial crime publications either address copyright infringement on P2P networks, or else the detection of fraud, usually from auction sites. Intelligence tools are most often concerned with either mining criminal social networks from open sources, or providing alerts about potential criminal activity, often with respect to certain geographic or temporal limits.

Those papers whose focus was least obviously a criminal matter often made reference to pornography, which may be indicative of different legal frameworks and cultural backgrounds. Such results in this review might be considered to address parental control systems rather than strictly handle criminal content.

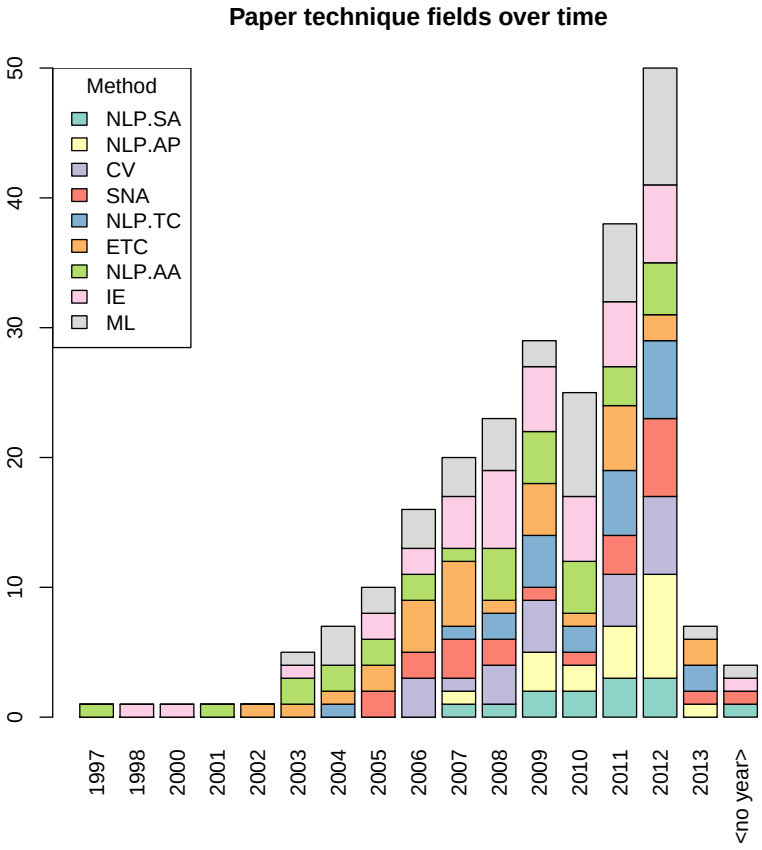
It is worth noting that 'terrorism' and 'cybercrime' were both often used as general motivations, not necessarily specific to the paper's focus, with 74 papers containing a reference to 'terrorism' and 50 papers referencing 'cybercrime' compared to 47 and 12 papers actually labelled as addressing these topics.

What are the methods which are being employed to provide solutions?

Natural Language Processing (NLP) is highly dominant in this review's results, with around half of all collected papers making some use of NLP techniques in some way. The presentation in Figure 2.2 breaks down this category along closer lines. The heavily textual nature of most electronic communications makes this a somewhat unsurprising result. Machine Learning techniques are also well-represented, with common classifiers like SVMs and Naive Bayes being applied to a variety of problems.

There are 21 papers in the review (10.2% of the corpus) which make some use of computer vision or image processing techniques. The low proportion of such papers may be linked to the choice of search terms in the discovery phase of the review — there was no CV-linked term included, but there were NLP and SNA terms. Of these 21, 16 papers made use of only CV techniques. As is to be expected, most of the data sources used in

Fig. 2.2 The most common techniques over publication years



this area were forms of image and video, with only a couple of exceptions where web and email data was processed visually.

22 papers in the review (10.7%) made use of some form of social network analysis (SNA). This number appears relatively low given a search term specifically selected for these methods, perhaps indicating a research area which requires further exploration. Of these 22 papers, 13 were labelled as only using SNA methods, the others overlapping with techniques from the domains of information extraction and natural language processing. The data used in these papers were primarily web data, including blogs and fora, but email data also formed a sizeable proportion of the study.

39 papers in the review (18.9%) focused on helping combat crime by mining information from public resources. Of these, 24 were labelled as solely oriented towards information extraction, while the remainder also used methods involving natural language processing and social network analysis. The vast majority of information extraction studies made use of web-based data, including online fora and social networking services like Twitter.

There are 43 papers (20.9%) which make use of machine learning techniques, only 20 of which make exclusive use of such techniques. Of the other 23, 18 use some form of NLP technique, indicating a significant overlap between those papers labelled as using machine learning and those labelled as using natural language processing. Such a relationship is retrospectively unsurprising given the close relationship between these fields. The 43 papers were fairly evenly divided with respect to the data types studied, with email and web data each featuring in nearly half of all studies. Four studies handled chat data and two studies — one overlapping with CV techniques — made use of image data.

92 studies from the review (44.7% of the total) used some form of natural language processing, making this by far the largest category of methods. Of these 92, 65 used only NLP techniques, making this also the category with least overlap with other methods (closely followed by the much smaller group of computer vision). The large number of

NLP-related studies collected may be linked to the inclusion of two terms in the search procedure which link to NLP.

There were 27 papers which did not fit within any of the broader technique categories. Chat data and network trace data were more prominent amongst these papers than in the main categories. Often these papers described frameworks or abstract processes for combating a threat.

Which online data sources are being used?

A breakdown of the different broad categories of data is provided in Figure 2.3. Most commonly examined was web data, with nearly half of all publications making some use of textual or semistructured web data. Within this category, simple web pages are most favoured, with social media – particularly Twitter – second and online fora third most popular. Behind Web data comes the other significant data source, email, which has the advantage of being both long-established and well-used. Chat data from instant messaging applications forms the third key data type under analysis, with comparatively few papers making use of images or videos.

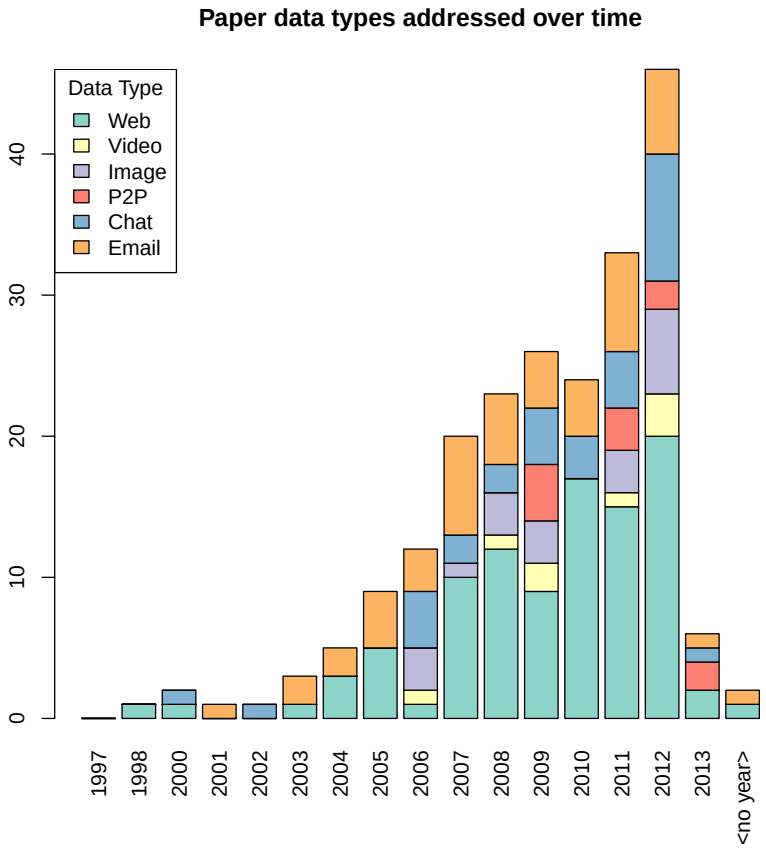
With regard to specific data sources, a few common elements were observed across papers.

- **The Enron email dataset** is a dataset of roughly half a million email messages from roughly 150 users. It was originally made public as a result of Federal Energy Regulation Commission's investigation into the Enron corporation. A full explanation of the dataset is provided by [119]. As a labelled dataset of authors and a large volume of messages produced by them, this corpus was often a standard reference for studies attempting authorship attribution, but also used in some studies based on social network analysis. A total of 22 papers make reference to the Enron dataset.

- **PervertedJustice**[2] is a vigilante website where volunteers run sting operations by posing as minors and luring paedophiles into volunteering identifying and

[2]perverted-justice.com

Fig. 2.3 Data type usage over publication years



incriminating information. They publish a large and growing corpus of chat logs involving attempted grooming, which have been used frequently in studies attempting to identify sexual predators or analyse stages of grooming attacks. The common use of this resource points to a common issue for researchers working on such topics – the lack of actual case data to work with means that researchers must work with proxy data. The degree to which predator-volunteer conversations accurately reflect predator-child conversations is unclear, in part also due to this lack of real case data to verify results. 15 papers make reference to this resource.

- **The Dark Web Forum Portal** is a search and summarisation interface to a collection of 28 fora which are linked to extremist or terrorist material. As a standard collection of online forums, it is of particular interest to researchers studying the organisation of terrorism online. 11 papers make reference to this resource.

How many studies are making use of multiple data sources?

Relatively few papers (21 or 10.2% of the total) combine different types of data or present methods which would apply to different types of data. Among those which did, the approach was typically either general monitoring of all network traffic such as in ECHELON and similar wiretapping, or else the use of NLP methods which could apply to online texts of many kinds. Slightly more papers (24 or 11.7%) were marked as 'partial' responses for this question, due to using a variety of data sources of the same type – for instance, using both English and Arabic text corpora to evaluate a hypothesis.

While in many cases a paper's contribution will be limited to one particular subject area, and thus would not be expected to apply to distinct data types, this result hints that more effort at synthesis of otherwise distinct forms of online data could well prove a fruitful area of research.

How many studies validate their contribution's utility to law enforcement practitioners?

Very few (10 or 4.9%) papers reported a positive evaluation of their tool's utility by a law enforcement practitioner or similar expert authority. It was possible to infer from the means of evaluation or similar references that a further (14 or 6.8%) of papers were written with co-operation of law enforcement, implicitly crediting the work with some level of practitioner support.

These figures do not necessarily reflect the true rate of interaction between researchers and practitioners, and it is possible that trials with law enforcement are only conducted after successful publication, or that law enforcement use of tools is not widely publicised in the name of reducing criminal awareness. Even with such qualifying scenarios in mind it seems problematic that papers specifically reporting themselves as supporting law enforcement and intelligence applications so rarely report on evaluations by the relevant professionals.

2.1.4 Discussion

Online criminality is often linked to the relative anonymity of electronic interaction, and in response to this the reviewed computer science literature, and particularly natural language processing, reveals a mature field of authorship analysis for online texts, with many rigorously evaluated methods for determining the author of a given text building on and referencing each other, with feature sets and reference corpora being shared between papers. Consideration has been given to the standards of evidence required for legal use. Other approaches to identification of online individuals for criminal matters show similar levels of evaluation.

The detection of sexual predators in online chat transcripts shows similar levels of interest, with multiple studies applying a range of methods to the same goal, and even a number of publications recording a specific competition, with methods using the exact same dataset so as best to be compared. It is notable, however, that the most common data source for these publications was a form of proxy data – the Perverted Justice website's transcripts between people outside of law enforcement agencies and sexual predators. This seems to indicate that a willing research community is having to work around legal or other restrictions on gaining access to actual criminal chat data.

Similar legal obstacles seem to be faced by researchers attempting to develop means to automatically detect child abuse media – many forced to use less-helpful forms of proxy data such as adult pornography – and even studies merely attempting to quantify the presence of child abuse media, where filename-only approaches are dominant.

Publications investigating terrorism or extremism also have access to a common data source in the form of the Dark Web Forum Portal, though it appears to be less uniformly drawn upon. With this topic, widely mentioned even in publications not directly addressing it, scarcity of ground-truth information about real-world threats appears to have diverted many efforts in the open literature into exploration of the networking and rhetorical properties of self-identified online extremist communities.

As is revealed in the breakdown of the quality analysis in the appendix to the full paper [69], a significant proportion of publications reviewed had deficiencies in evaluation

and indeed a quarter of publications had no evaluation. While in a minority of cases this may be because the paper proceeds via theoretical proof, or because the format of the publication does not allow sufficient space, in others poor adherence to scientific standards are evident. Especially when designing methods for use in law enforcement or intelligence deployments, where lives may directly be ruined by underperforming analysis tools, researchers must be focused on the best way to identify objective truth regarding their methods.

In some cases, with papers relying on social network analysis or information extraction methods, and particularly where the method designed was semi-automated or involved visualisation, evaluation sections presented only demonstrative case studies of application as support for their tool or method's utility. Case studies are sufficient for exploratory presentations, but where concrete benefit to law enforcement is promised, measurement should be made of these benefits. If the contribution is increased performance of the analyst interfacing with the software, for example, sufficiently defensible user trials must be presented, and the same can be said for tools which aim to help an analyst seek resources on the Web.

In other cases, laboratory evaluations of classifiers were presented, but insufficiently comparable to the real-world deployment scenarios. With the exception of copyright infringement, most fields of study in this review hold an inherent class imbalance problem – there are far fewer traces of criminals in the online world than there are traces of innocent netizens, and classifiers operating on a general population must thus overcome the likelihood of high rates of false alerts. Synthetic but unrealistic datasets may demonstrate a classifier's theoretical ability, but evaluations should always be linked to actual deployment scenarios.

A small number of long-term projects and toolsets were referenced in multiple papers gathered by this review. In some cases, these publications report on significant incremental improvements on developed approaches, with fresh evaluations (e.g., [19, 149, 150]). In more modular systems, such as the Email Mining Toolkit, the discussion is limited to brief description of individual components of a larger toolset (e.g., [209, 210]).

The lack of detail makes it difficult to ascertain the strengths and weaknesses of such extensions with respect to each other as well as other comparable approaches. Further work and more detailed evaluations are needed to fully understand the effectiveness of such extensions.

Impacting and underlying this survey's results is the rapid pace of development in online mediums. While certain technologies such as email have remained fundamentally consistent over the years, the same cannot be said for all online activity which draws law enforcement attention. Individual games and social networking platforms can become popular, draw law enforcement attention, and then become unpopular even as researchers devise the appropriate tools to analyse this content — some of the data sources in reviewed papers are from what might now be thought of as essentially dead communities. Certain methods, such as analysis of written text, can be generalised across a number of platforms and data sources, and are as such especially valuable.

The current volume of papers making use of multiple data sources and data types is low, with information extraction studies being the most likely to attempt this. The community should put greater focus on tools which generalise to different applications. Many methods may already be transferable, but studies attempting to replicate the performance of a method on a new type of data are very rare. The cross-examination of different data sets might also help standards of evaluation for researchers working in areas where accurate ground-truth is not readily available.

A number of papers on textual analysis and information extraction subjects demonstrate that their methods work with multiple languages, the most common being English and Arabic. English being dominant globally, online and in science makes it a clear target for analysis, whereas Arabic is clearly targeted by law enforcement and intelligence interest in counter-terrorism applications. The linguistic challenges behind textual analysis should not be forgotten or assumed solved when dealing with less-analysed tongues, but law enforcement should be aware that such technology exists outside of what is collected in this review, even if it does not advertise itself as applicable to law enforcement.

Finally, the extremely low overall level of engagement with law enforcement bodies or domain experts is problematic for a corpus of papers specifically selected for referencing their intended deployment with law enforcement. This is not necessarily a problem which may be overcome by the research community alone, but attempts should be made to involve relevant professionals in the evaluation of tools being designed for their use.

2.1.5 Conclusion

Many directions might be taken in moving forward from these results. This thesis focuses on the implications of two areas in particular. Firstly, the results indicate that tackling problems of online identity is one of the most fruitful areas for practically assisting security and law enforcement, addressing a unique and pressing aspect of the online environment. Particular focuses within this problem area, such as authorship attribution or avatar recognition, show signs of advanced and high-quality work.

Secondly, the relatively low rate of publications involving different types and sources of data suggests an area fruitful for methodological advancement. Given the multimedia nature of online social networks, their broad adoption, and their inherent connection to the problem of online identity, they would appear to make a natural ground for valuable research into *identity resolution* security tools.

As later sections will detail more closely, there is a coherent body of work on identity resolution methods for online social networks, but research in this area is being limited by a number of factors – ignorance of existing work in other fields, a lack of models for the comparable value of information attributes, hurdles to iterative improvement through lack of reliable and relevant ground-truth, and poor replication on 'real-world' datasets. It is to these theoretical and methodological issues that this thesis ultimately addresses itself.

2.2 Fundamentals of Identity Resolution

The topic of this thesis has roots which extend into an unfortunate terminological tangle. *Identity resolution* is the term by which the topic has been introduced, and for consistency's sake it will be adhered to where possible. Yet this is only one of many labels which have historically been applied to the same activity. It has also been known, variously, as *record linkage*, *data matching*, *entity resolution*, *data integration*, *object identification*, *duplicate detection*, *deduplication* and many combinations of the components of these terms. In addition, there are highly related fields such as *authorship attribution* and *information retrieval*, where relevant methodology has often been developed. As different fields use different terminology, it has not always been clear that similar problems are being addressed, and solutions found in one area may not have reached others. There are nuances of meaning and of context which colour each of these labels with regard to the core activity in question.

For clarity's sake, this many-named activity should be described. What is meant by *identity resolution* is a process whereby some investigator, armed with records regarding one or usually more persons or items, may match said records to a second set of records so that any matched records refer to the same person or item, thereby resolving all data regarding the same identities.

A classical example is presented in Figure 2.4. Given two databases, which might contain much of the same information, but in different formats and without a uniquely identifying key for both, an investigator wants to determine which of the two records refer to the same person.

We can easily identify that record A1 has a likely match in record B1. While the two databases store John Smith's information in different format, we can piece together his full name and address from the given components and identify the abbreviations of "St." and "Lancs". *Postcode* is not present in Database B, but aside from that all the attributes match, we just need to pre-process both datasets into the same format.

The cases of A2 and A3 are less clear. A sharp observer might note that there is a Mary in Database B with the same date of birth as the Mary in Database A. Given some

Database A

ID	*Surname*	*Firstname*	*Address 1*	*Address 2*	*Postcode*	*Birthdate*
A1	Smith	John	13 Broad St.	Lancaster	LA4 3HQ	07-11-1981
A2	Smith	James	12 Chapel St.	Blackburn	BB5 6ER	05-08-1984
A3	Smith	Mary	12 Chapel St.	Blackburn	BB5 6ER	04-03-1986

Database B

ID	*Name*	*Address*	*DOB*
B1	John Smith	13 Broad Street, Lancs	07/11/1981
B2	James Nearby	19 Parliament Rd., Clitheroe	06/08/1984
B3	James S. Short	14 Horton Way, Southampton	05/08/1984
B4	Mary Barlow	12 Parliament Rd., Clitheroe	04/03/1986
B5	Carl Barlow	12 Parliament Rd., Clitheroe	03/12/1974

Fig. 2.4 Two databases in a simplified identity-resolution example

knowledge about the world – that women in Western European cultures may change their surname when married – and noting the existence of a co-resident Carl Barlow, we might suspect that A3 should be matched to B4. Of course, even identifying A3 as a woman requires inference from the name field. If we were to construct a key just from first name and date of birth, we might find some systematic errors. It is possible, though less common, for men to change their names, so perhaps we should also consider the possibility that B2 or B3 are matches for A2. In both cases we have reason to be dubious: Southampton is quite far from Blackburn, and B2 has a different date of birth. Of course, A2's date of birth should not have changed, but it is possible that this reflects a data-entry error in Database B. There is no compelling evidence that either B2 or B3 refer to the same person as A2, and while we might suspect a connection, we cannot be sure that A3 refers to the same person as B4, rather than there just being a coincidental similarity in two individuals' records. The solution to identity resolution, then, must inherently be a *probabilistic* one, based not just on the attributes of the records presented but also on inferences made about the quality of the data and its real-world implications.

In the example given above, the records referred to people. Resolution of records referring to people is a particular focus of this thesis, and as such a contributing factor for the selection of *identity resolution* as the term of reference – 'identity' being more clearly

associated with people than the broader terms ‘entity’ or ‘object’. *Identity resolution* has also been the term typically used in investigative or security contexts, which are another focus of this thesis.

The earliest approaches to identity resolution came not from from a security background, but from the perspective of ordinary record-keeping. The seminal paper on the topic, given in 1946 by Howard Dunn of the National Office of Vital Statistics [65], uses the metaphor of a “Book of Life” written by each person on the archives of various private and governmental record-keeping bodies. He makes the case for assembling the pages of this book, in terms of benefits to the individual and to the state and other bodies. Among the questions considered in scope about a person are: “What sort of jobs do they hold?”, “How many children do they have?” and “What sort of illness do they suffer from?”. The solution he discusses focuses in the interim on the establishment of simple and effective record-keeping, drawing from his experiences with the Candian Dominion Bureau of Statistics, and in the longer term on the promotion of unique identifiers for citizens in the form of a Birth Card’s certificate number.

One of the first computational approaches to identity resolution was proposed by Newcombe et al. [159]. They identified the chief impediment to automatic record linkage as being *“the unreliability of the identifying information contained in successive records which have to do with the same individual or married pair”*. They gave examples such as altered spellings of surnames, inconsistent ordering of a person’s given names, and mistaken birthdates or ages. Their resolution was that improving record-keeping was not enough, and that records need to be linked *in spite* of these inconsistencies. They discuss an identity-resolution system for connecting marriage and birth records about the same family: using both family and female maiden surnames, phonetic encoding of these names, the birthplace, and the first initials of parents. While they considered no one of these features alone to be entirely reliable, they noted that, for instance, a member of the Canadian Records Bureau would put little weight on the information that parents in both birth and marriage were born in British Colombia, as this is surely the case for a large group of non-matched records. The same person might, however, put

weight on the information that in both records the parents were from New Zealand and Switzerland, this being an unusual combination in their population. Newcombe et al. move from this insight to a definitive model, connecting the value of an agreement or disagreement between records to the frequency of the values presented in those records. This probabilistic model was first expressed in terms of the binary logarithm of the observed frequencies, as

$$log_2 P_l - log_2 P_r \tag{2.1}$$

where P_l is the frequency with which an agreement (or disagreement) between these values occurs in *linked* records, and P_r is the frequency with which agreement appears at random between non-linked pairs of records. The value of the expression will be positive if the attribute match should be taken as evidence of a genuine match between records, and negative if it should be taken for disagreement. By calculating these values, and then summing these probabilities for all agreements and disagreements between a pair of records, a judgement can be reached about whether the balance of probability suggests these refer to the same family or not. Newcombe would later restate this model as

$$log_2(P_l/P_r) \tag{2.2}$$

which is the form most commonly referred to in later work, including this thesis [158]. Note that Newcombe defines these probabilities of agreement (or disagreement, there is no need to alter the model) in terms of specific values within fields, such as the New Zealand-Switzerland birthplace example, and not generically across the field (such as birthplace). Newcombe did however discuss understanding the value of particular fields for the process known as *blocking*.

Blocking is a necessary step in most identity-resolution systems, due to the vast quantity of records to be compared. Without blocking, identity resolution is an $O(n * m)$ process – in trying to resolve identities between two databases of 1000 records each, 1,000,000 pairs of records must be compared. This very quickly becomes an

unsustainable process, – consider the case of trying to match a few thousand records to a larger dataset that is already millions of records long – and in many cases much of the work is manifestly unnecessary, as simpler processes than full identity resolution calculations can tell one that the no match will be found. The solution used is to divide the pairs of records into *blocks* which are smaller than the entire search space, but which are highly likely to contain the genuine match. For example, a common method for identity resolution is to use some transformation of the surname of the individual as the *blocking key*, so that full profile comparison only occurs between a record and those records which share a similar surname. Of course, with any blocking system one runs the risk of accidentally excluding the true match. Newcombe expressed his blocking quality measure as a general merit ratio between this risk and the discriminative power gained by the blocking, or

$$M = D/I \tag{2.3}$$

Where I is the probability of introducing errors, and the discriminative power D is given by reference to a *coefficient of specificity*

$$C_s = \sum P_x^2 \tag{2.4}$$

with P_x being the proportion of a file in each block x, and the discriminative power is inversely related to C_s

$$D = log_2(1/C_s) \tag{2.5}$$

So that, intuitively, discrimination increases the more finely the blocking key divides the original records. This quality measure for a blocking system can be considered a precursor to the quality measures for identity resolution which will be presented later in this thesis.

Newcombe's methodology was formalised by Fellegi & Sunter [74] in 1969. They expressed what was then termed *record linkage* as a problem of deciding, for a set of

comparison vectors between pairs of records $\gamma \in \Gamma$, whether the vector referred to a link (A_1), a non-link (A_3) or whether there was insufficient evidence for a decision (A_2). All linkage rules are seen as producing conditional probabilities $P(A_1|\gamma)$, $P(A_2|\gamma)$ and $P(A_3|\gamma)$, which the authors constrain by

$$\sum_{i=1}^{3} P(A_i|\gamma) = 1 \tag{2.6}$$

They then define the possible errors in terms of the probability of a comparison vector γ given the true outcome, and define the optimal linkage rule as the one which has the minimal number of cases where no decision is made. They proceed to demonstrate that Newcombe's weighting method (Eq. 2.2) approaches the optimal method as the number of agreement measurements is extended, depending on certain conditions, and that this agrees with other formulations from the realm of hypothesis testing. For more detail on the development of the probabilistic record linkage techniques that underlie identity resolution, including the deployment of the EM algorithm, the reader is referred to the overview by Winkler [234].

2.3 Identity Resolution in Online Social Networks

Jumping ahead many years, and across disciplines, an interest in identity resolution has emerged amongst computer scientists studying online social networks and online profiles. For the most part, these authors do not appear directly aware of the existing approaches to identity resolution from the statistical literature.

2.3.1 Early clustering approaches

Early approaches to the problem viewed the issue through the lens of clustering together resources such as web-pages. Rekkerman & McCallum [24] tackle clustering web pages about the same person, using a variety of approaches to build a classifier for information about a specific person as opposed to individuals with the same name. Their approach highlights how unstructured or at best semi-structured data, from large and

difficult-to-index populations, frames the problem with different challenges, even if their methodology for collecting training data – hand-verification of matched pages – closely resembles that deployed in the statistical literature.

Bridging this work more specifically to social networks, Malin [140] discusses social network methods for disambiguating entities with the same name. He uses both hierarchical clustering of (for example) websites to form clusters of resources referring to the same entities, and a random-walk based method over a social network built from co-occurrence relationships weighted by the sparsity of such relationships. His evaluations on a dataset from the Internet Movie Database (IMDB) find that a surprisingly low threshold is sufficient for high F1 scores. The use of an external ground-truth is notable, but also the consideration of network properties as identifiers highlights how approaches not necessarily possible in the traditional statistical record linkage context can be highly effective at identity resolution.

This is further highlighted when Narayanan & Shmatikov [155] demonstrate a de-anonymisation attack against the Netflix Prize dataset. They focus on the identifiability of 'micro-data' – sparse ratings of content in a typical zipfian distribution, even where sanitisation and non-uniform subsampling have been applied to the original data. They present an approach which relies on the identification of improbable correlations between a small set of movie ratings in the Netflix dataset and public IMDB records. They find that 8 movie ratings with 14 day error in dating are enough to identify an individual with high probability.

Similarly, Szomszor et al. [214] focus on social tagging systems, and how user tagging tends to persist across different folksonomie focii. They find that salient interests are present in user tags from both `del.icio.us` and Flickr[3]. They use accounts they judge connected (based on exact profile name matching) to test the correlation of actual tagging patterns between the two networks. They filter tags, dealing with misspellings, pluralisation and shifts in terminology (including using Wikipedia to disambiguate proper

[3] `del.icio.us` is a now-defunct online bookmarking service allowing tags to be attached to links, Flickr is a popular photo-sharing service, permitting the same for images

nouns), and demonstrate that this improves the correlation between tags on the two social networks.

While these features are novel, and have unique properties of sparsity and distribution which are alien to the pre-digital statistical presentations, the approaches used by Narayanan & Shmatikov, and Szomszor et al. would fit comfortably within the general model of Fellegi & Sunter. Unfortunately, this connection is not acknowledged, and a partial theoretical re-derivation is achieved instead. The features being explored are very specific to the particular datasets, relying on a single attribute of the user (content ratings, tagging patterns).

2.3.2 Unique identifiers

Of course, non-probabilistic approaches have also been applied in online identity resolution. Golbeck & Rothstein [86] approach identity resolution via a Semantic Web project known as the Friend-of-a-Friend project, which merges together social connections from multiple social networking sites, using a common vocabulary and markup. They focus on enabling a reasoner to connect profiles via unique identifiers such as Chat IDs, and measure some of the properties of the cojoined social networks. This approach recalls that of Dunn, seeking to find or create a unique identifier for a person between records. Further developments in this area are pursued alongside and informing strictly probabilistic approaches to identity resolution. Bouquet & Bortoli [31] would later describe the "FOAF-O-matic" tool for creating FOAF profiles connected to a global resource identification system. Their system relies on the existence of globally unique identifiers and either the adoption of shared ontologies by social network providers or else the re-entry of data by users.

Zafarani & Liu [249] take a less strict approach to unique identification, focusing on the use of usernames as a discriminative feature for mapping identities across different online communities, starting with a blogging community. Their method is to perform a web search for the username of a profile, and parse the resulting URLs for candidate usernames, which they modify with common prefixes and suffixes. If any of these

usernames appear in the target domain, they consider this a match, and this method appears to hold for some 66% of cases based on the ground-truth they extracted from their blogging community's links. Many later works return to usernames as a roughly-unique identifying feature for validating other schemes, a practical but problematic approach given these figures.

2.3.3 Iterative filtering

In one of the earlier approaches to multi-attribute identity resolution in online data, Motoyama & Varghese [153] treat identity resolution as an information retrieval challenge for end-user support in social networks, focusing on the use-case of a user trying to locate their friend within a network. They use identities gathered from Myspace and Facebook in a multi-layered search process, using first one attribute (such as name) and then another, until they have a final set of candidate profiles. They build a classifier based on weighted values of individual threshold judgements for each feature. They find that name fields are the most important to matching.

A similar method is employed by Rowe [188], who focuses on disambiguating information about individuals with the same name within a semantic web context. His method relies on network graphs, based on the intuition that a person will appear on web pages which reference other people from their social networks. Using members of the local computing department, he extracts social network connections for each subject from their Facebook and other social network profiles to a FOAF ontological RDF structure connected by `sameAs` relations. He searches for the subjects' names on social networks and builds resource graphs of the entities extracted from these documents. He then merges these resource graphs based on string similarity for people and geographic proximity for locations, resolving social identities to resources using a graph traversal measure.

This iterative filtering approach used by both authors is distinct from the probabilistic combination of the Newcombe model, but begins to approach it in that it combines the judgement of individually unreliable attributes. The approaches could be understood as a

particularly aggressive form of blocking, such that relatively poor comparison vectors can be used, because the comparison set is already strongly preselected.

2.3.4 Data quality & availability

Carmagnola et al. [37] explore identity resolution as a task for online database holders to engage in to better manage user data, such as preferences, for personalisation purposes. They explore what they term the 'univocity' (how much a feature may assume the same value across different users) of various identity attributes including email address, last name, first name, and birth city. As part of this they record estimates for the number of values attributes take per user across some 25 services. They implement a hierarchical model for identification based on this, and test it on 80 users, 64 of which they cause to fill in registration forms for different services after some delays introduced to produce realistic 'copy errors'. They achieved 5/64 false negative and 2/16 false positives. Here again there is a partial reconstruction of Newcombe's approach, including comparison between the frequencies of values for the same user and for different users. They relate these terms in a manually-weighted model, however, and use a different thresholding system rather than inspect the sign of feature weights. Notable for the purposes of this thesis, though, Carmagnola et al. produce an overall identification value weight for a range of different online profile attributes, based on certain qualities of the attribute data.

A secondary aspect of data quality also sees some coverage. Irani et al. [109] crawl a social aggregator site in order to examine how many personal information fields are disclosed on average by a person using a social network, and relatedly how much of a person's total online footprint can be connected through attributes such as username and name. They discover that an active member has an average of 5.7 social networking profiles, and that connecting these profiles increases the number of identity attributes which can be collected. They test methods for connecting profiles based on a known pseudonym, revealing up to 40% of an individual's footprint, or real name, with more variable performance. They also discuss measuring the consistency of certain profile attributes as a means of confirming an identity. They find 'sex' to be the most consistent

marker, but discard it as not very discriminatory, and focus on last name, birth year and country as good fields. Though their treatment does not constitute a combined model, Irani et al are here exploring both the consistency and availability of certain profile attributes.

In similar availability-focused work, Krishnamurthy & Wills [126] focus on information leakage via HTTP headers and cookies associated with OSNs, looking at how third parties might access these details. This is of secondary importance to identity resolution, but as part of their discussion they include a list of personally identifiable information which is available across different social networking sites – an early contribution to understanding the scope of possible attributes in this domain.

Nosko et al. [162] describe a study into information disclosure in Facebook. 400 randomly selected Canadian Facebook profiles were selected, and their content was assessed according to a schema developed from multiple pass human annotation. They examine relationships between different demographic markers and willingness to reveal sensitive information. Abel et al. [5] set out to investigate general applications of interconnecting profile information. They collect a large sample of public profiles via Google's Social Graph API, downloading the publicly available attributes in order to study completeness of the attribute fields necessary to complete vCard or FOAF profiles on up to five different services. They examine tagging behaviour as well, including an exploration of tag prediction based on a user's tags from different services.

2.3.5 Security & privacy leak detection

The use of identity resolution to reveal security and privacy leaks in online social networks – as opposed to helping users or networks provide functionality – has gained serious interest. Narayanan & Shmatikov [156] examine how the anonymised data social networks share with advertisers and researchers can be de-anonymised with reference to external social networks. Their method proceeds by searching the anonymised graph using the degree of certain nodes in an auxiliary graph and the number of common connections between these nodes - early termination helping the process. They verify

this process on a dataset of Twitter, Flickr and Livejournal networks, using username matches as ground-truth. They hit a recall of 30% and a precision of 72%, impressive for a single feature.

Wondracek et al. [235] present a method that uses group membership information to de-anonymise users of social networks. They discuss how attackers can learn group memberships through examining the history and cookies associated with social networks. By learning the public membership list of groups, the attacker can create a candidate set of users which may be the target, which can be more quickly filtered by finding the intersection (though this can be fragile to misleading data in the browsing history). They analyse this attack against the Xing network, and check the feasibility for its application to Facebook and LinkedIn. They also detail the results of some crawling experiments, including the accessibility of group and member directories on a range of OSNs, showing an important recognition of the difference between a value being recorded in the social network and that value being available to an adversarial identity recognition system running against its web presence.

Iofciu et al. [103] explore whether data from social tagging systems such as Delicious can be used to identify individuals in other tagging systems such as Flickr. They suggest the use of tagging information as an accompaniment to username information, analysing performance across Delicious, StumbleUpon and Flickr. Their tag metric is BM25, a form of TFIDF where the IDF is tempered by site-specific features. They combine username and tag features through a parameter λ, which defines their relative importance. Their evaluation is on a dataset from the Social Graph API, which explicitly makes information about connections between profiles available. They found 1467 people with a Flickr and Delicious profile, and 321 of these with a profile on StumbleUpon. They note that very few tags are used in each system. They find usernames work for 55% of their data, and tags add roughly 9% to this.

Friedland et al. [78] create a general threat model of the privacy-invasion in identity resolution, distinguishing between targeted and easiest-K motives (trying to resolve specific identities, or trying to resolve some of the easiest-to-resolve identities) as well as

the capability of attackers, with focus on the heterogeneity of data which can be used by attackers. They discuss some motivating examples, and two specific attacks – one related to geotags and another related to the speakers in various audio samples.

2.3.6 Credibility & user support

In another area, authors start to compare the truth value of online attributes. Rowe & Ciravegna [190] address a number of issues regarding disambiguating resources about a person. They perform a comparison between the real world social network revealed by 50 participants and the network extracted from their Facebook accounts. They found that at least half of the real social network was duplicated in the digital social network, with an average of 77% coverage. They employ a rule-based decision process to identify web resources that refer to the identities from their original seed dataset. They go on to evaluate this process through the creation of a manually-identified gold standard of 50 people's Facebook and Twitter accounts. Further work by the same authors continues the theme of exploring the credibility of online information [189]. This same property will be explored later in this thesis under the title of *veracity*.

Related work by Cortis et al. [54] approaches identity resolution as a challenge for support of end users, focusing on the case of resolving the identities of contacts from multiple online accounts. They report percentages for users' self-reported rate of consistency between professional and personal online profiles. Using a common schema to represent profiles, they match accounts based on linguistic analysis – differentiated by the attribute type – as well as simple string matching of attributes and lookup of data such as addresses or job titles in knowledge bases to determine equivalence. They do not provide an evaluation of this approach's efficacy.

Raad et al. [184] also consider cross-network profile matching as an end-user challenge, with it being considered the basis of a number of user experience related functions. They use a weighted similarity vector approach for comparing profiles, validating their method against some simulated data based on real FOAF data. Significantly for the work carried out in this thesis, their method aims at generality across the domain of

online social networks, and lists a process to assign weights to attributes based on the consistency of values between matched profiles – one half of Newcombe's methodology.

Kontaxis et al. [121] propose an interesting application of identity resolution: to detect profiles which have been cloned across networks by attackers as part of a social engineering campaign. They extract key identifiable information from a legitimate social network profile, and search for profiles on other networks which contain this content. The identifiable information is identified by reference to the legitimate network's search system, in their demonstration on LinkedIn this became the person's title, current and previous employers and education history. Matching profiles are verified through exact string matching and profile image comparison. They provide a small-scale validation and then examine the possible extent of profile cloning within LinkedIn.

2.3.7 Integration of social network analysis

Further work at this time relied on developments in social network analysis, approaching identity resolution from a network edge construction perspective. The simpler approaches in this class include identity resolution using the overlap of names in friends lists, as presented by Labitzke et al. [128], and the work of Buccafurri et al. [35], who present identity resolution as the problem of finding 'me' edges to connect online social networks via 'bridge' users which have profiles in each network. Their method relies on the similarity of both usernames and a proximate contribution from the common neighbours of the two nodes, though they seem not to evaluate the efficacy of this second component in real classification attempts. Labitzke et al made note of the availability of attributes in their connection of online social networks, and acknowledge the impact that availability of information has on the possibility of linking a given profile.

In a notable development, Narayanan et al. [154] discuss the feasibility of authorship recognition when given large class sets: they work with texts from 100,000 possible authors, finding correct identification in 20% of cases, raising this to 80% when dropping recall to 50%. They strongly distinguish their work from previous authorship recognition, which typically deals with 100-300 possible authors. They evaluate their approach on a

large dataset of blog posts, using function words and single-character features along with word length distribution and captialisation, avoiding bag-of-words models[4] as part of an effort to avoid detecting only the commonality of topics between posts. The critical development here, however, is the acknowledgement of the difference between attribution in a bounded dataset, and attribution across the entire search space of online media, a distinction often brushed over in previous work.

Chen et al. [45] discuss the notion of 'complementarity' in aggregation of OSN profiles – the information being gained by resolution of the profiles, and the related finding that users with multiple social networks are more likely to share attributes within a network. They also measure the consistency of attributes versus a random model of data, and finally link online profiles to an Australian telephone directory as a demonstration of how linkage is not only a matter for online profiles.

2012 also saw important developments in multi-attribute identity resolution, with papers combining profile-based and network-based solutions to identity resolution. Malhotra et al [139] used social aggregators (FriendFeed, Profilactic and the Social Graph API) to gather information on users of both Twitter and LinkedIn. They used a similarity vector constructed of scores for the similarity between username, name, description, location, image and connections. They trialled four supervised classifiers: Naive Bayes, a decision tree, kNN and SVM. Strings are compared with Jaro-Winkler distance, self-descriptions with Jaccard distance between standardised term representations of the text, images as a greyscale vector of values in [0,255] which are compared with Levenshtein distance, locations via Euclidean distance and number of connections by an internal bin class of connection count. The most discriminative feature was the user's name. They test performance not only internally, but in the real data retrieved through searches for the display name of one profile on the other network. Their learning methods performed significantly better on their internal dataset (precision and recall of 0.99 and 0.96) than on the external one (accuracy of 0.64), suggesting that a classifier's reliance on names is dangerous for real-world performance of this sort.

[4]Which represent documents as a simple collection of all the words they contain, without more considered feature selection.

Similarly, Bartunov et al. [23] describe an approach to profile linkage across social networks using a Conditional Random Fields technique. Their approach combines network similarity information with a profile field comparison vector. Working with a dataset of Twitter and Facebook profiles, they train classifiers, revealing high performance, especially where node degree is high.

2.3.8 Recent developments

Work on identity resolution which was developed during the formation of this thesis has focused on more general approaches using the range of available profile attributes, with a focus on generalising to large user populations and resolution between multiple online social networks.

One exception to this general rule is the work of Chen et al. [44], which instead focused on the estimation of attribute uniqueness, leveraging a large dataset of Facebook profiles in order to identify the most-revelatory individual attributes when it comes to identification. Their thorough discussion of uniqueness estimation complements earlier studies of attribute consistency and availability.

Jain et al. [113] motivate identifying users across multiple online social networks from a security perspective, as a means for detecting malicious users. They describe user identities as being comprised of profile, content and network components. Within identity resolution, they distinguish 'identity search' from 'identity matching', with identity search being a means for creating a candidate set and identity matching being the identification of a true match in the candidate set based on the comparison vectors. Their conception of identity search is highly similar to the underlying principles of blocking, with the proviso that Jain et al are working with large and potentially incompletely-indexed databases. They describe some novel identity search methods to complement the usual name-based candidate generation, including methods which would not be possible in the traditional statistical setting. They then use some similarity measures between profile attributes and profile images to rank candidates, then presenting them to a manual verification in a semi-automated workflow.

Later work by the same authors [112] approaches identity resolution as both a security issue and a business concern regarding accurate audience estimation. They continue their earlier division of the subject into *identity search* and *identity matching*, and discuss a range of search methods, including profile, content, self-mention and network methods. They use the ground-truth from the Social Graph API dataset to evaluate their search approach between Twitter and Facebook users. They demonstrate a cascaded machine learning approach to using usernames as a discriminative feature for identity resolution, using Twitter self-mentions as the ground-truth.

Bennacer et al. [26] describe an algorithm to iteratively match profiles across numerous social networks, building on known connected profiles to inform ongoing identity resolution. They use a network topology method to select candidates for matching – adopting a similar process to that advocated by Jain et al above. The iterative nature of this process acknowledges the nature of identity resolution in its entire online context – this is not merely a process between two datasets, but a potentially unbounded exploration of the online profile space. Another identity-resolution system that focuses on multi-network applications is the Mypes tool presented by Abel et al. [6], who also used the tool as a means to explore information availability and revelation patterns across social networks.

Goga et al. [83] focus on the use of three specific features for identity resolution across social networking sites: geo-location, timestamps and writing style. Using a dataset of linked Flickr, Yelp and Twitter accounts gathered via a friend-finding function and an existing large list of emails, they demonstrate that the combination of these features is comparable in effectiveness to previous approaches based on usernames, so highlighting that users can be identified and connected despite adopting distinct pseudonyms. They also examine the improvements in accuracy from each feature, finding location and timing to be powerful features where available, combining them in a logistic regression classifier. At the same time, these authors [85] demonstrate the feasibility of large-scale account correlation attacks, building on previous efforts focused on the use of usernames alone by including real names, profile photos and locations as features, demonstrating significant

recall with a high precision. They work on a dataset of Twitter, Facebook, Google, Flickr and MySpace accounts. Usernames are compared using the Jaro string distance metric, images via a 'perceptual hash' which can be compared using Hamming distance, as well as a facial recognition module, and locations as a scaled geodesic distance between coordinates. They include a discussion of availability and discriminative ability of profile attributes.

The SuperIdentity project [28] is a recent large effort at understanding real and digital identities, and how different identity components from different scientific domains might relate to a core conceptual identity. As part of this project, Bruce et al. [33] have highlighted how identity fragments from different domains can be best visualised for law enforcement, including the chain of reasoning for inferred characteristics. As part of this work, they constructed a broad model of identity components, the *cyber persona* components of which could be construed as an alternative *matching schema* to the one developed within this thesis. Creese et al [55] elaborate on how this model can be used to understand the reachability of information across social networks, though the process is best understood as semi-automatic, with tools and domain knowledge guiding investigators to resolve identities and build a consolidated profile, rather than enabling the resolution of multiple identities across large datasets.

The availability of ground-truth data is a significant limitation on developments in the application of machine-learning to identity resolution. This is highlighted by the impressive work of Liu et al. [133], enabled by a large governmental database linking millions of Chinese online accounts to a unique identifier. This high population legibility allows for the development of highly accurate models based on user text and facial recognition. Similarly, Wilder et al. [233] deploy identity-resolution techniques on a 10TB collection of roughly 6 million entities' various social media accounts. They note with some confusion the lack of more developed general approaches to identity resolution.

Most recently, the area begins to see publications which are critical of the real performance of previous identity-resolution solutions for online profiles. Vosoughi et

al. [222] noted that many previous publications have relied on unreliable ground-truth – matched usernames – for the mapping between profiles, despite known problems with this data. Goga et al. [84] use a number of criteria to evaluate the reliability of identity-resolution systems, with model which closely resembles that presented in this thesis [5], and commented critically on the data collection methodology of many previous studies.

Related work has also addressed the topic of identity *verification*, where the focus is on gauging the trustworthiness of attributes drawn from online profiles, or entire profiles. Bahri et al. [16] provide a solution to this problem based on crowdsourcing information about validity from online communities, having raters rate the validity of individual attributes in order to detect Sybil profiles. This is an intriguing potential source of information about the *veracity* of profile attributes.

2.4 Summary

The survey of online data mining technologies targeted at law enforcement revealed several issues with the quality and evaluation of such work, which are worrying for such an important discipline. This motivates work which enables security researchers to raise the bar in these areas. The review also revealed that methods for the identification of criminals in online contexts are highly important to law enforcement, and that methods fusing multiple data-types are under-employed.

Returning to the foundations of identity-resolution procedures reveals that a strongly defensible probabilistic framework exists for understanding identity resolution at the level of individual traits (e.g. the name Smith, the birthplace Ireland) but this is only extended in a limited way to understanding the value of attributes which might present those traits (e.g. names, addresses, photos).

Looking at the history of recent work which attempts identity resolution in online social networks, in a computer science context, it can be seen that the foundation statistical literature is not well referenced or integrated, with publications mostly focusing on identity resolution as a machine-learning problem, and individual studies reconstructing

[5] See Chapter 4.1 for a detailed discussion of the similarities and differences between the approaches.

portions of the general understanding given by Newcombe. Novel methodologies are advanced which deal well with new types of data, but comparison between these types is scant, and recent papers call into question the reliability and reproducibility of published approaches, and trace this in part to issues with data collection strategies.

It is this context, of the unreliable ground-truth provenance and a lack of well-grounded domain-general data quality measures for identity resolution, that the following pages of this thesis is situated, and it is these challenges which this thesis will seek to address.

Chapter 3

Sampling Labelled Profile Data for Identity Resolution

The advent of the internet, and in particular online social networks (OSNs), has brought a fresh wave of voluntarily-provided profile information on individuals, the majority of it available to the general public. These profiles contain detailed information about aspects of peoples' lives which were previously unrecorded, and as such the value of linking profiles has exploded for everyone from advertisers to sociologists to criminal investigators.

Particularly relevant for gathering additional information about a person is the case where identities need to be resolved across different OSNs due to the way specific OSNs record different categories of information. For example, a profile on an image-sharing site may reveal a person's visual record of their day, while a microblogging platform profile presents written report and commentary. Tying the two together provides a more complete picture of events. Consider the motivating example given in Chapter 1, of the police officer checking other social networking sites for corroborating evidence about an alibi – it may be that a microblogging post attests a suspect's location, but an image shared on another platform clearly captures them elsewhere.

The social web could almost be viewed as a commercial implementation of the DARPA Lifelog project [57], providing an online database for nearly every aspect

of a person's life, except that the tables lack proper index keys to connect records. Demonstrating how profiles may be linked across services despite this has been a method for privacy researchers to alert the public to the value of what they are publicly revealing.

A number of solutions have been proposed specifically for identity resolution tasks across OSNs, each making use of some part of the diverse feature set available in social network profiles [85, 156]. Yet without a common frame of reference to work against, these various approaches and results are difficult to compare, which hinders identification of the best-performing methods and the direction of future research.

In many machine learning domains, research is advanced by the sharing of labelled datasets for purposes of replication, validation and incremental improvement on methodology. However, ethical constraints can prevent the dissemination of such datasets when they contain significant personal information, such as is always the case with profile data from OSNs [258]. While this profile data is nominally public information, as accessible as newspapers, it would be irresponsible to assume that personal information embedded in a public profile dataset is safe to preserve forever, and allowing members to later excise their data would pose significant obstacles to maintenance and consistency of instances of the dataset. Attempts have been made to anonymise these resources, but numerous de-anonymisation attacks have been demonstrated against such ostensibly anonymised datasets [61, 156, 258].

Rather than provide a single common dataset, this chapter proposes a sampling method which should allow researchers to independently gather *comparable* datasets. This approach is taken to overcome the tension between the research need for replication and ethical handling of personal information. The following sections propose, implement, and evaluate a sampling tool for gathering labelled connections between online instances of profiles, and also for gathering suitable negative data – real profiles which a classifier may be realistically asked to discriminate from the actual linked target. The output of this tool is a labelled dataset of profiles suitable for training and evaluating systems aimed at resolving identities across different OSNs.

Providing a tool rather than a dataset allows for comparable samples of linked profiles to be independently harvested by researchers from publicly available data on OSNs, without need for public release of actual profile data snapshots.

The aim of this chapter is to demonstrate that data collected by different researchers using this tool will be sufficiently comparable that their methods and results can be contrasted with some confidence, while at the same time they are working with data realistically reflecting the current social networking landscape. This approach also allows individuals and OSNs to determine between them what information is to be revealed to the public and does not presume upon any improper access on the part of researchers acting as part of that public.

The chapter is structured as follows. Section 3.1, surveys historic and existing sources of ground-truth data as used in previous studies, identifying issues with these sources. Section 3.2, outlines the sampling method proposed, along with some requirements for implementing it. A demonstration of one such implementation is provided in Section 3.3, and two large samples are gathered via this implementation for use in Section 3.4 and 3.5 to validate that samples drawn through this method are comparable. Section 3.6 concludes by discussing the results and some outstanding issues in this area.

3.1 Ground-Truth Data Sources

In aid of identifying suitable methodology, this section surveys the data sources employed in existing literature on identity resolution across social networking sites.

Malhotra et al. [139] in 2012 made use of three separate sources: Google's Social Graph API, and two social aggregators, FriendFeed and Profilactic. Of these three sources, none are still operational. This is a recurring pattern with social aggregation services similar to FriendFeed. Many exist or have existed, marketing themselves to users on the basis of consolidated access to multiple social networks, but they commonly go out of operation or are bought up by dominant social media organisations which repurpose their assets. This is disappointing, because as Malhotra et al. and also Jain et al. [113] with their small Social Graph API dataset and Irani et al. [109] with their unnamed single

aggregator site all demonstrate, these sites can be a rich source of ground-truth data whilst they exist.

One of these services, Plaxo [180] (which now operates as an online address book service, with mostly private profiles), has released a tool which highlights how user annotation of links might be utilised by researchers to gather labelled profile linkage data, relying on `rel='me'` annotation within the anchor tags for links as part of a crawler. To make suitable use of this annotation, researchers would first have to gather a large random sample of profiles which contain annotated links. Though they do not explicitly state their collection method, Buccafurri et. al [35] appear to have made use of such `rel='me'` annotations and/or Friend-of-a-Friend (FOAF) data (see below) in identifying cross-links between profiles on LiveJournal, Flickr, Twitter and Youtube, a dataset which was later enriched by Bennacer et. al [26]. In this dataset of 93,169 nodes, only 462 unique cross-links are identified, suggesting such annotations are not in widespread adoption.

Golbeck and Rothstein [86] used FOAF semantic data obtained from a number of social networking sites, looking for specific shared traits in FOAF files such as chat IDs or homepages in order to identify profiles of the same person. The FOAF format – being a common format for description of profiles and their interconnections – would be theoretically ideal for gathering linked profiles, if it were widely supported by large OSNs. However this does not appear to be the case, with LiveJournal the lone popular exception amongst a largely niche set of small OSNs which support it.

Goga et al. [83] made use of the Friend Finder functionality which was formerly common on many social networks, using an existing list of 10 million email addresses to find users' accounts present on multiple social media platforms. Due to several privacy concerns raised by the feature, many social networks no longer allow email-based search for profiles, most notably Facebook [18]. Even were the functionality still available, the email addresses required in order to utilise it to gather linked profiles are typically more closely guarded than other profile information.

Narayanan et al. [156] take a somewhat different approach in their de-anonymisation study, basing their ground-truth mappings between profiles on exact matches in the username or name fields, attempting to verify such matches with a score generated from a small number of heuristics – the length and rarity of the name, and overlap in location information. As their method (topographical identification) did not rely on any of these features, this linkage method retains validity within their study, but it cannot easily be generalised as a means for other researchers to go about acquiring ground-truth mappings for identity resolution.

Based on an exploit discovered by Kaafar et. al [116], some researchers make use of the optional 'other profiles' feature of Google Buzz profiles to identify cross-links between profiles from different networks. They gather a large dataset of some 4 million profile identifiers from Buzz, a predecessor of the Google+ social network, using a graph-based crawler which collects lists of Follower/Following users from each profile. A large proportion of these profiles made use of the 'other profiles' feature, and as such this dataset has gone on to be reused in several other studies on identity resolution across OSNs [45, 85, 178]. However, Google Buzz was discontinued in 2011, and its successor Google+ does not make a profile's Circles (the Follower/Following relationship being abandoned with Buzz) easily accessible for scraping.

Based on this survey, it appears that the majority of previously employed datasets in this area of identity resolution come from sources which are no longer available for re-sampling. Those datasets which may theoretically be re-sampled in the same manner are of limited value, covering only small user populations.

3.2 Sampling Method

If researchers are to avoid making assumptions based on usernames, and cannot rely on the availability of unique identifiers persisting across OSNs (such as email addresses), then the search for ground-truth data is effectively a search for instances where a user has stated a connection between two or more of their own profiles. Social aggregation services are one means by which such information may be collected. However, they

appear to be an unpredictable source, not suitable for the basis of long-term research. If social aggregation services cannot be relied upon as indexes, then it may be better to examine the social networks themselves for users' revelation of connections to other networks.

This is similar to the approach used in the tool released by Plaxo [180], which examines the `rel='me'` property of links to find links which a user identifies as being another profile of theirs. This annotation does not appear to be in widespread adoption, but it may be possible to find alternative indications that a link is intended to represent another profile of the user.

Presuming for the moment one such OSN where one might expect to find this ground-truth link data, terming this the *primary study network* or *primary network*, the problems can be stated as follows:

1. Gathering a representative random sample of profiles from the *primary network*. Notably, one is not interested in identifying the most connected users or in sampling a connected subgraph of the *primary network*, only in a random selection of profiles (or in graph terms, nodes).

   Previous efforts focused on crawling large graphs of OSN users through the application of breadth-first search or random walks [34] are unable to reach disconnected components of the overall graph and are usually biased towards popular nodes by early stopping.

   Most desirable would be methods which can directly sample from the network, such as the ability to randomly select from assigned unique identifiers, but these indexing mechanisms are usually not publicly available.

   As an alternative, the network search functionality provided by many OSNs can be used to gather unbiased samples of profiles. This functionality is provided to users to enable them to find other users based upon their name or other information. Given a random selection of search attributes (such as can be constructed based on population data such as census records), these search systems can provide a random index into the OSN's profiles.

2. Identifying in randomly selected profiles those linked profiles which belong to networks of interest. While links act as identifiers for a profile, extracting the profile content is an involved process highly dependent on the network being targeted. As such, it is prudent to focus on a few such networks – *secondary study networks* – and discard links to other networks.

3. Gathering plausible negative examples for a 'realistically challenging' dataset. A sample consisting of only those profiles which are known to be matched would be of little use for training and evaluating a classifier. As well as positive examples of profiles which should be matched, an appropriate sample should be made of those profiles which are not matched in other networks, for both primary and secondary study networks.

   For any profile, it would be possible to use other profiles in the same network as negative examples, but these profiles would make for a poor candidate set, being mostly easily distinguishable from the true results. Instead, researchers should opt for a candidate set which more reasonably reflects real disambiguation tasks with public social network data – search results in the *secondary study network*, with the query being constructed based on attributes of the *primary network* profile from which a link was found. Such a dataset better reflects a core issue of identity resolution: given a particular individual profile, how does one find out which of many profiles with the same name are the ones to be connected?

Note that users voluntarily complete these fields in their profiles, and so as with previously discussed datasets, the datasets this method aims to generate may not be valid for *adversarial* profile linkage tasks, where the emphasis is on detecting a link between a user who is attempting to mask any connection between their two profiles. Nor should the sampling method be taken to enumerate all matching profiles in the *primary network*, or any similar property which assumes an exhaustive exploration of any of the study networks. The dataset should remain relevant for purposes such as estimating the privacy impact of revealing certain profile attributes, testing existing identity resolution methods and comparing behaviours between the same individuals on different social networks.

Considering these issues, the requirements for this method are:

1. A *primary study network* in which users provide links which can be understood as statements that the link refers to another profile of theirs. This network must have a network search system which can be used for random sampling of profiles.

2. A set of *secondary study networks* which are linked to from the *primary network*. These networks must have an index suitable for selecting negative examples.

3.3 Implementation

One of the most promising data sources as a *primary network* for implementing this sampling of ground-truth data would appear to be Google+. As previously mentioned, Google+ provides an "other profiles" field on a person's profile page where users can provide links to their profiles elsewhere on the web. This field is accessible via the Google+ API and so it is possible to automatically examine the Google+ network to find profiles which link to other profiles of the same person.

There are other reasons to favour the selection of Google+: while it is difficult to predict the shifting landscape of OSNs, Google as an organisation seems unlikely to disappear in the short-term, and it seems reasonably likely to maintain the Google+ service or an equivalent network for the next few years. At the same time, a number of influential studies referenced above have historically made use of a dataset drawn from Google profiles.

The *primary network* must also have a search system which can be used to perform random sampling from the network. This is drawn from the approach of Gonzalez et. al. [87], whereby a random sample of names from a large list of uncommon surnames are used as input into Google+'s profile search API, and those result sets numbering less than Google+'s cap on responses are taken as an unbiased sample of profiles. The aim of using uncommon surnames is to increase the likelihood of retrieving result sets numbering less than the results cap. Because the Google+ search API limits the number

of returned profiles to a maximum of 300 per query[1], and these results are ordered by popularity, a sample which includes all search results would be biased towards more popular users. Therefore, the implementation accepts only those profiles returned by queries which have fewer than 300 results in total.

In detail, the method proceeds as follows.

1. Initial search terms are randomly selected from a list of 128,000 uncommon US surnames. Following Gonzalez et al. [87] this list was drawn from those surnames which occurred more than 100 times and less than 1000 times in the US Census 2000[2]. The US makes up a majority (55%) of the Google+ userbase, and so is the best national census for this purpose [207].

2. The Google+ search API is queried for these terms. Those result sets with < 300 items are taken as unbiased.

3. The search phase completed, all publicly available data on the accepted profiles is downloaded via the Google+ API. Two formats are used to store the data – one which records the exact queries and the raw responses, and another which standardises the data into a Profile object.

4. The 'other profiles' sections of the Google+ profiles gathered are examined to establish the ground-truth true links. Where a link is made to one of the *secondary networks*, that link is queued for download and a record is made of the connection between the two profiles.

5. The full name attributes of the Google+ profiles are then gathered to create a second set of search terms.

6. This second search term list is then entered into the search functionality for each of the *secondary networks*, and the resulting profiles are queued for later download. These results form the realistic candidate set for attempted identity resolution from the seed profile.

[1] At the time of publication for Gonzalez et al.[87], this limit was 1000

[2] Data on surnames occurring less than 100 times was not available

7. The profiles indicated by the true links and the candidate sets from the name-based searches are then downloaded from their respective networks' APIs, and stored in the same manner as the Google+ profiles.

There are a few implications of this method which should be borne in mind. Firstly, surnames of profiles will be unusually distinctive as compared with a population average, though the procedure for selection of negative results given above should mitigate this impact. Secondly, these names are those which are uncommon in the United States. As previously addressed by Gonzalez et al, the diverse immigrant history of the United States combined with the US bias in Google+ membership would mitigate the US-centric aspect of this concern, but there are possible correlates of low-incidence surnames with recent immigration and thus socio-economic status and perhaps in turn lower digital literacy. Next, it should be noted that the sample mechanism used has only 128,000 different search possibilities, with a proportional chance of collision, and also a maximum theoretical result size of 38,272,000 Google+ profiles (though in practice there are likely to be far fewer than this). Finally, there will be at most 299 Google+ profiles with the same name, so for a method attempting specifically to discriminate between such profiles, its capability cannot be demonstrated as greater than this limit.

3.3.1 Secondary study networks

Only nodes from a specific set of OSNs should be selected, these OSNs are termed the *secondary study networks*. Extracting structured information from profile pages involves API queries for the content of profile pages identified by URLs, analysis of which must be specific to the social network in question. Additionally, name-based search functionality must be implemented for each social network being sampled, in order to furnish negative examples. Therefore it is important to find the social networking sites which form a strong initial set of *study networks* from which to draw samples.

A number of constraints exist, including that the network in question must make profiles public (to members of the network, if not the wider internet) and allow for name-based search. The main deciding factor for including a network will be whether a

Network	Links Counted	Percent of Linked
youtube.com	322	22.5%
picasaweb.google.com	214	14.9%
facebook.com	195	13.7%
twitter.com	186	13.0%
linkedin.com	65	4.5%
blogspot.com	59	4.1%
google.com/reader/	26	1.8%
profile.live.com	24	1.7%
flickr.com	22	1.5%
yahoo.com	16	1.1%
instagram.com	15	1.1%
blogger.com	12	0.8%
tumblr.com	11	0.8%
soundcloud.com	10	0.7%

Table 3.1 Most commonly linked profile networks.

significant number of Google+ profiles link to profiles in the network, as this furnishes researchers with a greater number of positive examples to analyse, and focuses efforts on linkage tasks likely to be of more value in application scenarios. The method could easily be extended to include less-frequently-linked networks, though researchers may need to select larger initial samples from Google+ to get representative sets of linked profiles. Using the proposed sampling procedure for Google+ profiles and examining those profiles with links to other networks, counts were made of links to other networks.

As shown in Table 3.1, the most common networks which were not other services owned by Google (which one might expect to be overrepresented, and are increasingly integrated into Google+) were Facebook, Twitter and LinkedIn. These top three networks would appear to be mostly suitable as *secondary networks*, with some minor caveats regarding their accessibility: for example, LinkedIn does not offer a global name-based search feature within its ordinary public API, but this functionality can be obtained through web-scraping calls.

3.4 Evaluation

The implementation is realised in a Python tool[3] capable of sampling ground-truth data from the *primary* and *secondary* networks given in this paper. The primary evaluation of the sampling method is to compare the distribution of certain node attributes in different samples gathered by the implementation. The node attributes that are the simplest to compare in this manner are numerical, so the distribution of certain numeric properties of nodes – such as counts of followers and posts – is examined in different samples gathered from the Google+ and Twitter profile networks via the implemented sampling method.

Using the methodology described above, two large independent samples are gathered from both the Google+ and Twitter networks. The two samples of the Google+ network had respective sizes of 4,986 and 11,719 nodes, while the samples of the Twitter network had 8,259 and 17,862 nodes. These samples (henceforth Datasets 1 & 2) were gathered over Oct-Nov 2015 and Dec-Jan 2016 respectively. A number of numeric properties were recorded reflecting attributes of interest to identity-resolution research.

One could attempt to demonstrate a lack of statistically significant differences between these samples by aiming to *fail* a statistical test such as the two-sample Kolmogorov-Smirnov test. However, the large sample sizes mean that a direct test for statistically significant differences between the two samples would likely be overpowered for the usual critical values, with a high chance of committing a Type I error and finding a false difference between the groups.

Rather than focusing on statistical significance, it is possible to test whether there are important differences between the samples by comparing the effect sizes between the two samples. Table 3.2 shows comparisons between counts of attributes for each node. Cohen's d is the typical measure of effect size, but its calculation relies on assumptions of normality which are violated in social network data, which tends to follow power-law distributions. Instead, a nonparametric measure of effect size is applied. This measure is known as Cliff's δ, and has been recommended specifically for such situations [145].

[3] https://github.com/Betawolf/identity-sampler

Other properties of the two samples may impact their comparability for research purposes. This can be more directly examined by reference to the Kullback-Leibler divergence, also known as *information gain* when using one sample in order to approximate the other. This measure directly relates to the intended use of the sampling mechanism – as a means for researchers to compare results obtained on one sample with existing results obtained on a similarly collected sample. Table 3.2 reports the Kullback-Leibler (KL) divergence between the two samples, with measures discretised into 15 bins for computation[4]. As the KL divergence is non-symmetrical between distributions, the figures reported are the average of both directions of the measure.

Property	G+ $\lvert\delta\rvert$	Tw $\lvert\delta\rvert$	G+ KL	Tw KL
Age*	<0.01	–	0.01	–
NumFollowers	0.05	0.04	0.01	0.00
NumFollowing	–	0.03	–	0.00
NumInteracted	0.05	0.04	0.05	0.01
NumLocations	<0.01	0.02	0.01	0.00
NumTexts	0.03	0.05	0.00	0.01
NumDescribes	0.04	0.02	0.01	0.02
NumLinks	0.04	0.05	0.02	0.04
NumPics	<0.01	0.02	0.00	0.00
NumTimes	0.04	0.05	0.13	0.01

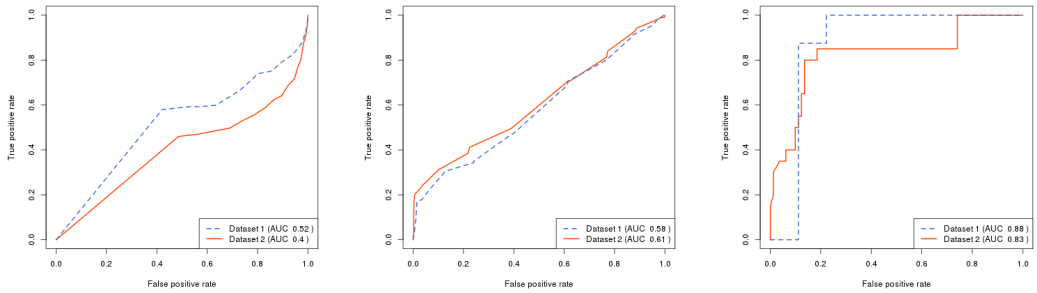
Table 3.2 Nonparametric effect sizes and average KL-divergence for comparison of the two samples from the Google+ and Twitter networks. *Age where available.

The KL results show that very little divergence is present between the two samples, or, alternately, that very little information is lost when using one to approximate the other. Similarly, The average of all δ for Google+ comparisons is <0.03 and for Twitter is <0.04, indicating a very low practical difference overall between properties in the two samples – the usual standard for a 'small' effect size is 0.2. The large sample size increases confidence that this result is not due to a failure to detect larger effects.

[4]Based on Sturge's formula, $k = \lceil log_2 n + 1 \rceil$

3.5 Application of Existing Identity Resolution Approach

As a secondary evaluation of the proposed approach, an existing identity resolution method is applied to both of the datasets. This serves to illustrate a possible use of these samples and further validates the comparability of results drawn from different samples. The aim here is not to provide a novel and competitive classifier, but to demonstrate the viability of the suggested replication method.



(a) Name-based classification (b) Image-based classification (c) Location-based classification

Fig. 3.1 ROC plots for individual feature classifiers

Following Goga et al. [85], the three features they used for identity resolution are investigated: the *name*, *profile image* and *location* of each pair of profiles.

3.5.1 Username

Usernames have often been considered a useful feature in identity resolution. Perito et al. [178] provide a full treatment of this topic. However, facets of the sampling method make names unlikely to be effective features: the 'display name' feature was used to generate the negative examples, so all comparisons are between profiles with highly similar names.

The effect is that names are not highly discriminative features in the comparisons made in the datasets, as shown in Figure 3.1a. In fact, the average Levenshtein distance between matched pairs of profiles was actually greater than the distance between unmatched pairs (5.82 and 4.01 for matched vs 2.75 and 3.24 for unmatched). This is the reverse of the normally expected direction in broader comparisons.

3.5.2 Image

A *perceptual hashing* technique is used to identify the key features of all profile images. The Hamming distance between two hashes [120] is then used to test for superficial adjustments to the same avatar image. This feature showed some small but consistent discrimination, with the average Hamming distance between matched pairs being 27.72 and 27.02 for Datasets 1 and 2 respectively, and 31.66 and 31.94 between unmatched pairs. Just as Goga et al. discovered, simple threshold-based classification using this image feature has poor *recall*, but high *precision* – not many users do use the same profile image, but when they do they are very likely to be the same person. As Figure 3.1b shows, this means this type of image similarity performs poorly as a classifier by itself.

3.5.3 Location

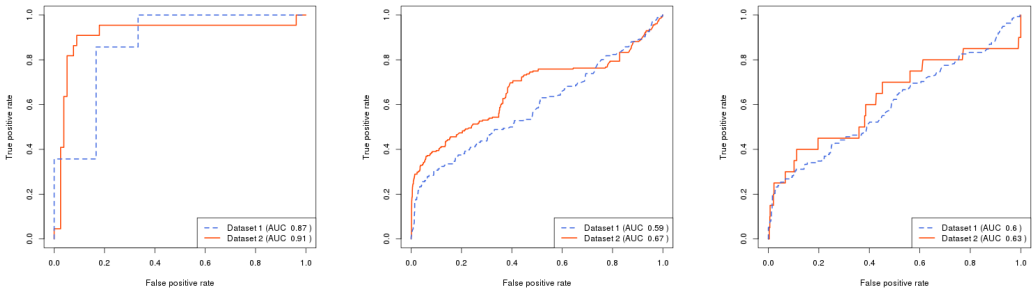
Location data such as geolocated status updates or persistent 'hometown' or 'location' fields can be a good feature when it is available. However, location data is quite rare in the datasets, and this rarity is compounded by location comparisons only being possible where both profiles have location data: only 72 of 9558 comparisons in Dataset 2 and 17 of 1309 comparisons in Dataset 1 could use geodesic distance as a feature, even where any available location information was used (i.e., both tagged status updates and stated profile locations).

As Figure 3.1c shows, however, within this small ($< 1\%$) subset, location distance was highly predictive.

3.5.4 Combined

The identifiability of these features was investigated jointly as part of a binary logistic regression classifier combining all three features, using a ten-fold cross-validation approach.

An important issue for classification tasks such as this is the handling of missing data. The majority of comparisons lack a location distance component, so how this is handled



(a) Omitting records with missing features

(b) Missing features replaced with means

(c) Subsampling 1000 of each dataset

Fig. 3.2 ROC plots for combined classifiers

has a significant impact on model performance. Naively omitting records with missing data produces good-looking performance, as shown in Figure 3.2a, but tells us little about performance for the majority of cases. Imputing missing data with feature averages produces a more muted performance across more examples, shown in Figure 3.2b.

Performance in general was quite poor where location information was not available, unlike the findings of Goga et al. [85]. This can be attributed largely to the differences in the discriminative ability of the *username* feature, as this has poor performance within the dataset due to the manner in which negative examples are gathered and comparisons are made.

The aim here was not to provide a competitive identity-resolution approach, but to demonstrate the comparability of results obtained through different samples via our methodology. It can be seen from the ROC plots that this is validated, with curves following the same trajectories with only minor deviations. Dataset 2 does tend to produce marginally better performance, but this is due to training benefiting from a larger sample size. Randomly subsampling 1000 data points from both samples produces a much closer match, as illustrated in Figure 3.2c.

3.6 Discussion

3.6.1 Implications for identity-resolution research

This chapter presented a sampling mechanism for gathering ground-truth links between profile networks and appropriate negative examples, in proportion to their appearance in real-world data. The evaluations confirm that samples being drawn in this manner are sufficiently comparable that methods developed against one sample should transfer to other samples drawn in the same manner with minimal impact – based on this initial analysis one could expect even small effect sizes to be replicated between experiments performed on different samples. It can also be expected that ROC curves from a method trialled on one dataset to closely track those from another.

A common reference point for experimentation is necessary for researchers to compare their methodologies, and sampling mechanisms which reflect their population are necessary for properly grounding results. Both comparison and reference to the true population are necessary for advancing the state of the art. It is hoped that this sampling method will be used by researchers in identity resolution as a basis for reproducing each others' results and comparing identity-resolution systems which make use of the heterogeneous data available in OSN profiles, something which has been hindered by the difficulties in obtaining and sharing such privacy-sensitive data.

The implementation presented focuses on the Google+ profile network as its *primary network*. However, the method is not restricted to application on just this network. Any OSN which provides a similar field to the 'other profiles' field within Google+ and makes this field publicly accessible would prove a suitable replacement. Indeed, recent work in identity resolution has started to recognise the identification use of URLs included in Twitter profiles [112]. While this field is less well-designated than the 'other profiles' field on Google+, and its utility as a source of ground-truth must be investigated, it would provisionally appear to be a candidate replacement for the Google+ 'other profiles' attribute which would allow samples to be drawn with Twitter as the *primary network*. This could improve the throughput of the system. It is also possible that

conventional blogging platforms might provide a long-lasting index which is open to more conventional web-scraping approaches, which might allow a future implementation to sidestep API limitations.

Similarly, the implementation presented suggests that blocking – the generation of candidate record pairs for identity resolution – be based on the name of one or more profiles, as this is the search mechanism used for collecting negative examples. This is not necessarily problematic, as name fields are often used as blocking keys, but it should be noted that alternative search systems can be used for finding candidate profiles, including searches based on content and network properties as described by Jain et al. [113]. Generally speaking, any property which can be used to generate negative examples from search of *secondary study networks* can also be used for blocking.

This may be particularly important when considering the performance of classifiers which include profile name similarity as a key feature, as sampling negative results based on name necessarily reduces its utility as a distinguishing feature. Such a task, however, realistically reflects real-world challenges in disambiguating users with the same or similar names.

Finally, note that while this approach is particularly tailored to research for identifying links between profiles on OSNs, the generation of accurate ground-truth data is a recognised problem for identity resolution in general [124], and it is possible that this sampling approach could be informative for researchers working within similar constraints, such as in bibliographical or medical record linkage.

3.6.2 Selection bias & limitations

As noted, the tool does not generate data which is appropriate to *adversarial* identity resolution – where the owner of the profile has anticipated a potential attempt to connect their identity, and worked to frustrate such efforts. The users for which positive matches are available are those who have volunteered the connection between their profiles, and could hypothetically be systematically more consistent in their presentation than profiles linked through some other means. A valuable angle for future work would be to explore

how adversarial, voluntary and otherwise inferred true connections differ in their profile consistency, to estimate effect size reduction from datasets such as are generated by this tool.

There are other selection concerns connected to the current implementation's use of Google+ and a US population index. Google+ itself may have particular social biases (e.g., towards early-adopters within the tech community) which influence the matching rate with other OSNs, and the consistency of contained profiles. Similarly, the sampling mechanism implemented may unduly weight results towards US profiles, which could be less representative for identity resolution in other domains. Finally, it must be acknowledged that the trends of social media are fast-moving, and even using OSNs themselves to identify links could prove fragile, as the population moves on to new OSNs or even more novel platforms, wherein new sampling mechanisms may have to be designed.

3.6.3 Limitations of the tool

The potential for one or more OSNs to alter or close their public API is a partial threat to continued functionality of the sampling tool. While the tool has been designed in a modular manner, so that secondary study network APIs which no longer work need not impair the general operation of the tool, it is likely that maintenance will be necessary to keep these modules functional. Policy changes on the part of the OSN may similarly affect the data this tool is able to provide to researchers. Note also that potential improvements in the speed and reliability of the tool could be achieved through sustained development.

This work has concentrated on development of a sampling tool which uses the APIs provided by the OSNs, using only the access rights granted to any app developer. This is ethically necessary: the sampling position as members of the public ensures no improper access is gained to the profile content of users by e.g. befriending them, or paying for profile information as an advertiser. Authentication with the OSN means their release of the data being sampled is tracked and recorded. However, use of the APIs for these

services can be limiting – in some cases, content which a member of the public may view on the web is not available within the API.

A possible solution to these limits would be to apply web-scraping technology to enrich profile data. This would bypass many hurdles with API limitations. However, this is not a straightforward proposition: modern OSNs make extensive use of asynchronously-loaded content, with little profile information accessible at the initial page load. Scraping technology has advanced in step, but a scraper intent on accessing large numbers of profiles may also have to contend with accounts and IP addresses being blacklisted, necessitating greater infrastructural requirements – such as a cooperating network of machines – for any sampling tool, which would hinder replication. Overcoming these issues may require centralisation of the sampling tool as a service for researchers, which re-opens questions about sharing profile data.

3.6.4 Privacy and ethics

The issue underlying the design of this sampling mechanism can be described as an ethical tension. It is easy for scientists to identify that making their results replicable is ethically necessary, this having long been a guiding principle of science. A direct approach to satisfying this replication requirement would be to release all the data used in an experiment, and in most areas this is still appropriate. At the same time, however, there is an increasing recognition of the paramount ethical obligations to protect the privacy of data subjects [143]. Even where, as in psychology or the social sciences, waivers can be be gathered to permit the release of some personal information, only relevant data is collected and communicated, to reduce the risk of a subject being identified. In large-scale studies of social networks, contacting profile owners for approval would be impractical, and in the field of identity resolution in particular it is not sensible to talk of removing personally identifiable information from a data release (except perhaps as a research challenge). Researchers are presented with a difficult choice: either they never release their data, protecting their subjects but hindering the development of their field,

or else release it, and risk harm to their many subjects and perhaps also personal legal consequences.

This chapter's contribution has been to identify a means for researchers in identity resolution – and related fields – to fulfil their ethical duties to their profession and colleagues without revealing the personal information of their subjects, drawing upon the reachability of a common population for sampling purposes. However, the solution cannot be said to entirely remove the underlying tension. For one, researchers must remain cautious about how they store and present data from these samples. For another, the scraping countermeasures discussed above require a careful response: the decreasing availability of useful ground-truth data about the identities of social media users may be a barrier for research in this field, but it could also be more positively viewed as an indication that social networking sites are becoming more protective of their users' privacy.

3.7 Summary

A review of the data sources from previous identity-resolution literature reveals that the majority of sources are no longer available for reuse or re-sampling, and those which are available are of limited value due to their constrained scope.

On the basis that mining the social networks themselves may be a source of ground truth information, a sampling methodology is described for producing realistically challenging datasets based on the visible cross-links from randomly sampled nodes, using the search features provided by social networks to collect both the initial population samples and negative examples that complement cross-links with realistic candidate sets.

An implementation of such a system is given for the Google+ social network, linking to profiles on Facebook, Twitter and LinkedIn. In an evaluation of this method, profile characteristics from two samples are compared to show their low deviation, and identity-resolution methods from previous literature are applied to demonstrate comparable results across samples.

Finally, a discussion is made of the limitations of this approach and implementation, along with areas where improvements could be made in future work.

Chapter 4

Modelling and Valuing Online Profile Information

This chapter presents the ACU model for understanding the identification value of attributes in profiles. Return to the example from Chapter 1, of the police officer searching amongst profiles with the same name as one offered up in evidence. How can he know which pieces of these profiles are useful to compare?

The first stage in answering this question would be to have a consistent theoretical system for understanding in general the identification value of a profile attribute. Such a system is described in Section 4.2 below. The next step would be to obtain a schema for understanding which attributes on profiles from different networks are functionally similar, so that they could be identified for comparison. This schema-building activity is detailed in Section 4.3. Finally, empirical measurements to inform the model must be made, for the components of the theoretical model, using the schema from the previous stage. Measurements in this regard are presented in Sections 4.4, 4.5 and 4.6. Finally, these values can be combined according to the model and a general identification value can be extracted for each schema item. The officer, or another user of the method, can then understand which attributes are useful to their application of identity resolution.

4.1 Background

We begin in identity resolution with a selection of datasets $D = [d_1, d_2, \ldots d_j, \ldots d_n]$. Each dataset contains some number of records $R = [r_1, r_2, \ldots r_l, \ldots r_o]$ relating to an individual, with fields $F = [f_1, f_2, \ldots f_i, \ldots f_m]$ relating to their different features within the dataset, such as name, age, etc. We can construct a *matching schema* to understand which fields, speaking across all records, are defined in a manner comparable to other fields in different datasets so that F is consistent across all datasets, even if for some datasets there are no corresponding values in records for a particular field f_i that is shared between other datasets, because that dataset did not contain a comparable piece of information.

The goal of identity resolution is to understand when a record $d_j r_l$ can be treated as equivalent to another $d_j r_l \equiv d_k r_m$. The $\equiv$ operator here means that the records are equivalent as a result of referring to the same individual. The $\equiv$ operator will also be used between fields for particular records to indicate successful comparison between the values – the particular age or name given for that record. For example, $d_j r_l f_i \equiv d_k r_m f_i$ indicates equivalence between the value given for a field f_i in two records $d_j r_l$ and $d_k r_m$, which happen to be from different datasets. The similarity of values held in a field f_i lends support to the hypothesis that the two records should be considered equivalent. The term 'profiles' is used interchangeably with that of 'records', and 'attribute' is used interchangeably with 'field'.

In prior work, Goga et al. [84] proposed the ACID framework for understanding the reliability of identity-resolution schemes in online social networks. This framework suggests that matching scheme reliability be understood as depending on four properties of fields (so each measure is defined for a field f_i). Below, that framework is restated in the terms defined.

Availability The probability that profile attributes for individual profiles $d_j r_l$ and $d_k r_m$ which are matched across distinct datasets have values $d_j r_l f_i$ and $d_k r_m f_i$ available in both datasets.

$$A = P(\exists d_j r_l f_i \cap \exists d_k r_m f_i | d_j r_l \equiv d_k r_m)_{\forall l,m \in 1\ldots o \forall j,k \in 1\ldots n} \tag{4.1}$$

Consistency The probability that the similarity function s for the two values $d_j r_l f_i$ and $d_k r_m f_i$ produces a result greater than some threshold value th, given that the profiles $d_j r_l$ and $d_k r_m$ do refer to the same entity and both values are available.

$$C = P(s(d_j r_l f_i, d_k r_m f_i) > th | d_j r_l \equiv d_k r_m, \exists d_j r_l f_i \cap \exists d_k r_m f_i)_{\forall l,m \in 1...o \forall j,k \in 1...n} \tag{4.2}$$

non-Impersonability The probability that the value $d_k r_m f_i$ has not been intentionally duplicated from the value $d_j r_l f_i$ by an attacker, so that the maximum similarity between two values $d_j r_l f_i$ and $d_k r_m f_i$ (where $d_k r_m f_i$ is drawn from a set of possible impersonator values VI) is less than a threshold th.

$$nI = P(\max_{d_k r_m f_i \in VI} s(d_j r_l f_i, d_k r_m f_i) < th)_{\forall l,m \in 1...o \forall j,k \in 1...n} \tag{4.3}$$

Discriminability The probability that an attribute value $d_k r_x f_i$ from the set of non-matching profiles $d_k r_x \ni \{d_k r_m, d_j r_l\}$ is less similar to $d_j r_l$ than some threshold th.

$$D = P(\max_{d_k r_x \ni \{d_k r_m, d_j r_l\}} s(d_j r_l f_i, d_k r_x f_i) < th \mid nI(d_j r_l, d_k r_x))_{\forall l,m \in 1...o \forall j,k \in 1...n} \tag{4.4}$$

They attempt to measure some of these probabilities with reference to a small set of possible attributes, and use them to explore the practical limits of some existing identity resolution approaches [84].

This chapter presents ACU, a refined, expanded and repurposed version of the ACID framework, suitable for application to *understanding the identification value of individual profile attributes* in both identity resolution and record linkage generally. The model rests on the *availability*, *consistency* and *uniqueness* of profile information. Alongside this revised model a grounded schema is presented for profile attributes in the domain of online social networking. These two components are then combined in the collection

of *domain-general* estimates of the identification value of each of the attributes in the schema, according to the components of the ACU framework.

Chapter 5 will further explore the application of these *domain-general* identification value estimates to improving general-purpose classifiers, demonstrating an application in feature selection under missing data conditions.

Chapter 6 will discuss possible extensions to the ACU framework, including re-integration of the concept of *non-Impersonability* under the label of *Veracity*.

4.2 The ACU Framework

This section will describe three properties which are necessary for understanding the identification value of the fields of records for record linkage purposes. The *identification value* of fields can be contextually sensitive, so here two applications should be distinguished: the *task-specific* identification value of fields and their *domain-general* identification value.

The *task-specific* identification value of a field is the expected utility of that field in a specific record linkage task between two or more datasets[1]. This necessarily implies a restricted domain (for example, joining two bibliography datasets), and where ground-truth data is available, an appropriate statistical analysis will reveal a direct measurement of the variance in match prediction being explained by each field. Of course, ground-truth data is often not available *a priori* for real-world record linkage tasks, and so alternative means of identifying valuable information are desirable.

The *domain-general* identification value of a field is a generalisation of the expected utility of that field in a certain class of record linkage tasks. This class should be defined by a matching schema which covers the major fields which are available in any of the datasets for which linkage might be attempted. This ontologically grounds the domain. For example, *domain-general* identification values might be estimated for the fields which are available in most bibliography datasets, for the purpose of understanding which fields are generally useful in linking bibliographical records.

[1] Or indeed in a de-duplication task within a single dataset.

The *domain-general* identification value is thus the upper limit case on the *task-specific* identification value. If the expectation of the task-specific identification value of some field f_i in a linkage task between two datasets is some function

$$E(f_i) = h(d_1, d_2)$$

of the two datasets d_1 and d_2, and in general for n datasets is an average

$$E(f_i) = \sum_{\forall i,j \in 1 \ldots n} \frac{h(d_i, d_j)}{n}$$

it can also be helpful to think of the relationship between *task-specific* identification value and *domain-general* value in terms of sampling theory, where the *task-specific* value is a property of a sample, and the *domain-general* value is the related population statistic. As will be examined in Chapter 5, an estimate of the *domain-general* value can also be used as an *a priori* estimate of the *task-specific* identification value of a field.

The ACU framework consists of three properties: *availability*, *consistency* and *uniqueness*. These properties are orthogonal dimensions for understanding the identification value of any of the fields $F = [f_1, f_2, \ldots f_i, \ldots f_m]$ in a *matching schema* of n fields. The selection of fields in a matching schema is based on on the common and corresponding fields which are usually visible in the application domain $D = [d_1, d_2, \ldots d_j, \ldots d_n]$ of datasets. In the *task-specific* case, D is the datasets across which records should be linked. In the *domain-general* case, D is the set of possible datasets which typify the area. In both cases, availability, consistency and uniqueness measure different properties which combine to determine the expected utility of any field f_i in record linkage.

4.2.1 Availability

The availability definition given in Eq. 4.1 is made conditional on the known true match status of two records. Within the authors' context of the reliability of a linkage method, this definition limits consideration to only the records most relevant to method reliability. However, this definition severely limits the property in the context of attribute

identification value, by leaving it undefined for non-matched or unknown match status records.

The alternative is to remove this restriction, and define availability as the probability that for a field f_i from the matching scheme there are at least some non-null values $d_j f_i$ and $d_k f_i$ in records from datasets d_j and d_k respectively

$$A(f_i) = P(\exists d_j f_i \neq \emptyset \cap \exists d_k f_i \neq \emptyset)_{\forall j,k \in 1...n} \tag{4.5}$$

This has the desirable property of being calculable without the provision of ground-truth data on match statuses, and thus usable by record linkage practitioners without the need for manual classification or model training.

Additionally, in many cases the *task-specific* identification value might be approximated by

$$A(f_i) = {A_{gf_i}}^2 \tag{4.6}$$

where A_{gf_i} is a *domain-general* estimate of the internal availability of f_i

$$A_{gf_i} = P(\exists d_j f_i \neq \emptyset)_{\forall j \in 1...n} \tag{4.7}$$

and so with suitable *domain-general* estimates, approximations of A_{f_i} can be used to inform model-building even before access is granted to the target datasets[2].

Furthermore, in the *domain-general* case, this definition can be usefully decomposed into two components: the *structural support* for a field within a domain

$$SS(f_i) = P(\exists d_j f_i)_{\forall j \in 1...n} \tag{4.8}$$

which measures the probability that a field exists within a dataset from the domain D, and the complementary component of *completeness*

$$CM(f_i) = P(d_j f_i \neq \emptyset | \exists d_j f_i)_{\forall j \in 1...n} \tag{4.9}$$

[2]Much in the same way domain expertise enables the same thing.

which measures the probability that that field contains a non-null value for any given record in a dataset d_j. Thus decomposed, availability spans two distinct explanations for missing data for a field f_i and dataset d_j, that either

1. the field in question was not intended to be recorded in this dataset; or
2. the field in question was intended to be recorded in this dataset, but individual records are missing values for this field.

The ability to distinguish between these cases is important, as remedies for one (reviewing data entry methods, making fields mandatory) are not applicable to the other.

This decomposition is still valid in the *task-specific* case, however, in such cases the structural support is resolved to a binary variable (either the fields exist or they do not) and non-existent fields would be omitted by sensible *matching schema* design, so considerations of completeness alone will dominate the measurement of availability.

4.2.2 Consistency

The definition of consistency given in Eq. 4.2 and its extended discussion [84] identifies consistency as being a probability of the output of some field similarity function $s()$ being above a certain threshold th. However, the introduction of a threshold into this definition is unnecessary, as $s()$ itself should be an appropriate probability for the definition of consistency between two attribute values.

Thus, the definition of consistency is the probability that two values are *equivalent*, given that a) The records containing these fields are in fact referring to the same entity; b) both fields exist and are non-null in the given records.

$$C(f_i) = P(d_j r_l f_i \equiv d_k r_m f_i | d_j r_l \equiv d_k r_m, A(f_i) = 1)_{\forall l,m \in 1\ldots o \forall j,k \in 1\ldots n} \tag{4.10}$$

This probability of equivalence can then be given by the appropriate similarity function s for the field f_i of known matched records. Definition of the similarity function is domain-specific.

One benefit of this definition is that it permits consistency information to be used in a Bayesian analysis, giving *consistency* a known role in Bayesian probabilistic approaches to record linkage. Bayes' rule establishes the manner in which one conditional probability might be constructed from its inverse and the non-conditional probabilities of events

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4.11}$$

If one defines

$$P(B|A) = C(f_i)$$

Then once values can be obtained for

$$P(A) = P(d_j r_l \equiv d_k r_m \cap A(f_i))$$

$$P(B) = P(d_j r_l f_i \equiv d_k r_m f_i)$$

a probability can be obtained for the likelihood that two records match, given that two fields match (expressed in terms of $A(f_i)$, but as noted before this is easily calculable for a given task).

$$P(A|B) = P(d_j r_l \equiv d_k r_m \cap A(f_i))|d_j r_l f_i \equiv d_k r_m f_i)$$

The value of $P(A)$, the expected overlap rate between datasets, will be task-dependent. The value of $P(B)$, the base rate of equivalences, is the inverse of the *uniqueness* of field's values, as discussed below.

4.2.3 Uniqueness

Goga et al. acknowledge in their initial statement of the ACID framework that the version of 'discriminability' given by Eq. 4.4 is practically impossible to estimate without

knowledge of impersonating profiles [84]. They instead work with a proxy definition:

$$\hat{D} = P(\max_{d_k r_x \ni \{d_k r_m, d_j r_l\}} s(d_j r_l f_i, d_k r_x f_i) < th)_{\forall l,m \in 1...o \forall j,k \in 1...n} \tag{4.12}$$

Taking this revised equation 4.12 as a basis, but also remaining consistent with the approach to the similarity function $s()$ outlined in the definition of Eq. 4.10, allows for a revised definition of:

$$U(f_i) = P(d_j r_l f_i \neq d_k r_m f_i | d_j r_l \neq d_k r_m)_{\forall l,m \in 1...o \forall j,k \in 1...n} \tag{4.13}$$

That is, uniqueness is the probability that two non-matching records' values for the field f_i will not match by chance. This probability, and its complement, have a number of desirable properties. Firstly, as mentioned above, $\hat{U}(f_i) = 1 - U(f_i)$ fulfills the denominator in Eq. 4.11, thus integrating all three terms of the framework into Bayesian approaches to probabilistic record linkage.

Second, U maps almost directly to the concept of *entropy* in information theory

$$H(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i) \tag{4.14}$$

most critically in that it rises in proportion to the number of distinct values which a field may hold, and thus existing solutions for measuring entropy can be utilised (however, entropy itself is not expressed as a probability).

Third, in the case of identity resolution and other areas where very low match-rates are to be expected, uniqueness can be approximated as

$$U(f_i) \simeq P(d_j r_l f_i \neq d_k r_m f_i)$$

which is calculable without known match states, or alternatively as

$$U(f_i) \simeq P(d_j r_l f_i \neq d_j r_m f_i | l \neq m)$$

within a single dataset d_j if match rates cannot be assumed to be low, but duplication within a single dataset is known not to exist.

4.2.4 Combination

Returning to equation Eq. 4.11, all the necessary components are in place to construct, for a given field and similarity function, from a dataset in a particular domain, $I(f_i)$, the conditional probability of a match in terms of a similarity function, weighting the function by both its prior performance on this kind of task and the availability of the information it relies upon.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(match|similarity) = \frac{P(similarity|match)P(match)}{P(similarity)}$$

By the earlier justified substitutions of $P(similarity)$ with $1-U(f_i)$, and $P(similarity|match)$ with $C(f_i)$ which is dependent on $A(f_i)$

$$I(f_i) = \frac{P(match) \times C(f_i)|A(f_i)}{1-U(f_i)}$$

As the value of $P(match)$ is a constant factor within a particular dataset, it has no bearing on the comparative weighting of field/similarity functions and can be omitted. For comparisons of feature weights between different datasets it should be retained, but this is a less likely application.

$$I(f_i) = \frac{C(f_i)|A(f_i)}{1-U(f_i)}$$

The conditional probability can then be simplified under assumption of independence[3].

$$I(f_i) = \frac{C(f_i)A(f_i)}{1 - U(f_i)} \tag{4.15}$$

This formula agrees with a secondary derivation from the classical statistical treatment. Recall the value-specific frequency ratio given in Eq. 2.2. The terms in that model record, for a specific value of a field, the frequency of agreement between matched records and between unmatched records. This relates to the definitions above for $C(f_i)|A(f_i)$ (As frequency of agreement weights agreement where complete by possibility of agreement) and $1 - U(f_i)$. A simple generalisation of Newcombe's method to the case of attributes, then would set the identification value of an attribute as

$$I(f_i) = log_2(\frac{C(f_i)A(f_i)}{1 - U(f_i)}) \tag{4.16}$$

Which is a simple and useful transformation of the weights given by 4.15.

4.2.5 Multiple fields

The previous discussion focuses on finding the identification value of a singular field f_i through the three measures of *availability*, *consistency* and *uniqueness*. However, it might be asked how this model extends to measuring these properties for the combination of n fields from F, which can be termed F_n.

Values of *availability* extended across multiple fields will tend to decrease as the number of fields included in the definition increases. Take for example a set of fields such as a username, a location and a profile picture in a domain. The availability of the field f_1, username, may be 1. The availability of the field f_2, location may be lower, at 0.4, and for f_3 it may be 0.2. The availability of the combination $\{f_1, f_2\}$ must be at most that of the least available component, 0.4. The availability of $\{f_1, f_2, f_3\}$ similarly must be at most 0.2. The lower bound on the availability will depend on the relationship

[3]Given very low empirical correlations observed later in this chapter, this assumption would seem to be justified

between the availability of the individual pairs. If we assume the two are independent, then the availability would be $1 \times 0.4 \times 0.2 = 0.08$. However, this availability could be correlated (users with locations always give pictures), in which case availability would be 0.2, or anti-correlated (users with locations never give pictures) in which case availability would be 0.

A similar process should hold for *consistency*. The consistency of any two fields in concert is, in the best case, the product of the consistency of each component field, and any combination of fields cannot be more consistent than the least consistent field amongst the components.

For *uniqueness*, the combination produces a different result. For a given record r_l the number of records in d_k (a dataset of m records) which accidentally share with r_l a comparable value for f_i may be I, and for f_j may be J. The number of records which by chance share comparable values for both f_1 and f_2 is, under independence $\frac{I}{m} \times \frac{J}{m}$, or under correlation of uniqueness of f_1 and f_2 is $min(I,J)$. In either case, the number of chance matches tends to decrease as the number of fields increases, thereby *raising* the uniqueness as more fields are added, as there are fewer possible accidental matches.

Combining multiple fields then becomes a tradeoff between the falling values of consistency and availability (which, it must be stressed, are multiplicative in the general identification value definition given in Eq. 4.15) and the rising value of uniqueness. Where adding additional fields fails to raise combined uniqueness of F_n more than it decreases the product of consistency and availability (and whether this is the case will be dependent on these measurement figures for the field in question), the additional fields are harmful to the identification value of F_n.

4.2.6 Summary

This section has defined the ACU framework for establishing the identity value of fields in a record linkage task, extending and refining the ACID framework presented in recent literature.

The ACID framework includes one term, *non-Impersonability* of attributes, which is not represented under the ACU framework. There are two primary reasons for this. First, it is not clear how values for non-Impersonability can be reliably established from a dataset. Such measurements might be accomplished by identifying actual impersonating accounts and examining the attributes held in common with the original profiles compared to baseline or matched profiles, but on a large scale network it may be difficult to gather appropriate ground-truth to support such an analysis. Goga et al. themselves only identify *potential* impersonators [84], and it is not clear that these are well distinguished from profiles with similar attributes occurring by chance.

Secondly, *non-Impersonability* is a property which is only meaningful in identity resolution, and perhaps primarily in the domain of online social networks. It does not extend to record linkage tasks in other domains, and as such it limits the applicability of the ACID framework to a narrow domain. By removing this property, the ACU simultaneously broadens the scope for its application and increases its ease of use.

In the following sections, this chapter proceeds by first describing how a matching schema of online profiles was grounded and built. Next, each of the ACU properties is described in detail with reference to this domain, and estimates of the domain-general value of each property are made for each of the profile attributes from the schema. These estimates are derived based on data gathered via the method described in Chapter 3, and on previous literature. Finally, the chapter concludes by examining these empirically grounded estimates of identification value, and how well such values correlate with each other and the existing understanding of the identification value of profile elements in identity-resolution literature.

4.3 Building a Matching Schema for User Profile Information

In a record linkage system, it is important to build a *matching schema* which can represent information from different data sources in a common format, so that attributes can be compared for equivalency. For identity resolution across online social networks, this would mean creating a schema to map the essential attributes of online user profiles.

Some previous work has been attempted in this area, but the schemas so derived are of questionable relevance. Chandler [38] provides a detailed discussion of identity construction on personal web pages, along with a structured list of their key features, but this feature set from 1998 does not fit well with the modern, service-defined user profile pages and the updated set of media they make available. More modern efforts such as the FOAF ontology [60] do better at reflecting certain aspects of modern profile pages, but are both too domain-specific (identifying attributes for particular online services) and fail to capture several important aspects of profile pages such as popularity information.

In a typical record linkage setting, a matching schema would be constructed between two or more databases which are intended to be connected. The application area of this chapter is not resolution between any particular set of online profile datasets, but understanding identifiability across the range of possible online services. To support such a goal, the model of user profiles presented in this section is constructed based on a sample of online profile datasets such as online social networks.

The broadest possible support for this model would be a review of the structure of all sites hosting public user profiles, but this standard is impractical for a manual review. One alternative approach would be to focus on a selection of websites which contain the greatest number of user profiles, but comparable information on the size of online communities is difficult to obtain, perhaps in part because such information is increasingly considered commercially sensitive.

Instead, this section draws on websites selected for a high overall volume of web traffic, based upon figures provided by Alexa[4]. The top 100 such domains were manually examined (in 2014) for public user profile pages, with 65 unique domains found to contain appropriate profile information. The Alexa rankings included many highly-ranked sites which were effective duplicates of each other[5], so such duplicates were resolved, the domains representing a total of 39 distinct services, four of which proved difficult to translate and were omitted, leaving a total of 35. It is important to note the limitations of this selection procedure: in the case of social sites like Twitter, Facebook or Reddit, traffic volume will usually correlate with user population, but this does not necessarily hold for sites such as Wikipedia or the BBC, where typical traffic is consumption-oriented and few visitors create public profiles.

Figure 4.1 lists the different sites examined, with categorisation according to their function and the number of different information items found for each site. Different site categories appear to carry profiles with similar levels of detail.

Clear leaders in information content are typical social networking sites such as Google and Facebook, as might be expected. Pornographic sites also rank highly, with detailed biographical pages for members, who seem to use these pages in a similar manner to a dating site. Question-and-answer forums such as StackOverflow also seem slightly above-average in the number of information items present.

Knowledge-building sites like IMDB and Wikipedia ranked low for user profile content, as did news sites, blogs and video-sharing sites. Aside from blogs, these categories are largely those where user profiles are less central to site functionality. Blogs and Wikipedia user pages are notable in that both may carry more structured information about the user in optional widgets, which have uneven adoption, and so were not considered integral to the profile pages.

Following is a model of user profiles, based on the information which was structurally and publicly available in the surveyed user profiles across these high-traffic sites [68].

[4]http://alexa.com (2014)

[5]For example, google.com and google.co.uk may differ in their user population, but the profile service offered is the same at both domains

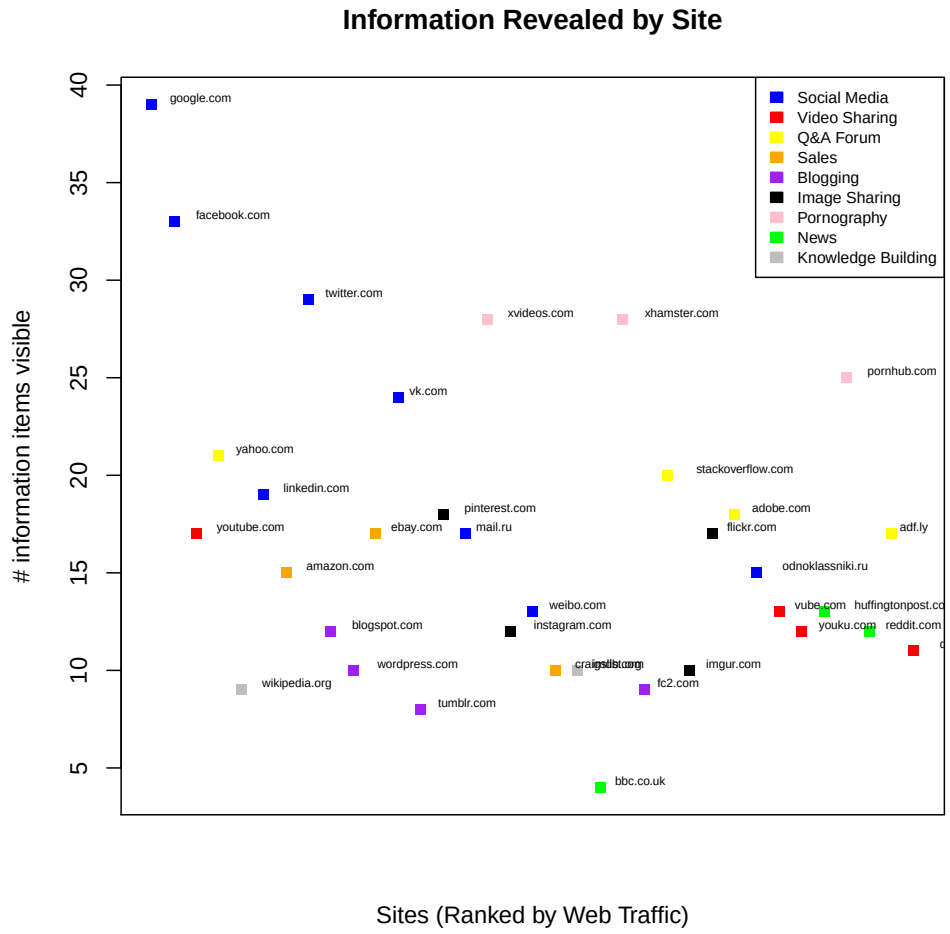


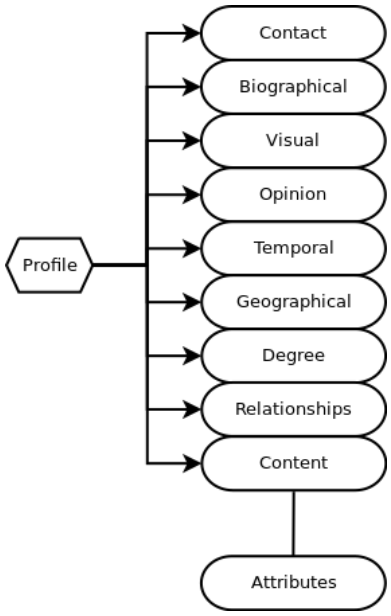
Fig. 4.1 Sites plotted against number of public information items

The 50 profile attributes so derived are grouped into 10 categories. Figure 4.2 outlines these categories, and the following subsections detail the contents of each category.

4.3.1 Contact

Contact information refers to profile attributes which are primarily of use for communicating with the profile owner, expressed as some form of technical address. Types of contact information are noted in Table 4.1, along with a short description of each attribute and the number of services surveyed which were visibly reporting this attribute on user profile pages. Not included are means of contacting the user such as platform-integrated messaging services, as these details cannot persist between nor aid in identification across different platforms.

Fig. 4.2 Outline of the categories of attributes in the schema



Web links – often titled as a 'homepage' – were the most broadly-available contact identifier, perhaps due to being a particularly general and well-known method for expressing identity online. Links to profiles on other networks were also common, and could be considered highly useful for re-identification purposes. Both email addresses and phone numbers were found on relatively few platforms. This is likely due to the sensitivity of these information items: while users can provide them to social networks for identification and contact purposes, they are not revealed to the general public due to privacy and security concerns.

Table 4.1 Contact information

Name	Description	#Services
Web links	URL field designated for the user's web site or homepage.	11
Profile links	Direct links to this person's profile on other services, explicitly particular to those services.	9
Email address	An email address for the user. Potentially service-provided or partially anonymised.	5
Phone number	A visible phone number for the user.	3

4.3.2 Biographical

Various fields from services provide biographical information about users. This includes the only feature common to all examined services — the username. In referring to `Bio→Username`, no distinction is made between fields supposedly for 'real' names and more typical online handles, any verified status being handled by the `Bio→Verified` attribute. As Table 4.2 shows, usernames were rarely verified as the user's actual name, but in practice they may well be a useful identifying feature by themselves. The most common field after `Bio→Username` were the `Self-description` areas, which contain free text about the user, followed by fields for age or birthdate and fields for a number of general-purpose descriptive tags.

Table 4.2 Biographical attributes

Name	Description	#Services
Username	A human-readable identifier for the user.	35
Self-description	A free text field for user self-identification.	21
Age	The age or birthdate of the user.	10
Tags	Short textual labels self-applied by the user.	10
Education	The user's educational history, marked by school names. (Sometimes dated).	8
Occupation	The user's current employment.	8
Gender	An explicit marker for the user's gender.	6
Relationship status	A field denoting the user's relationship status.	6
Sexual orientation	A field denoting the user's sexual orientation.	4
Verified	Whether or not the platform indicates if the username given matches the user's real name.	3
Religion	A field for the user's faith.	3
Physical description	An physical self-description of the user.	2
Habits	Fields denoting other personal habits.	2

4.3.3 Visual

Visual identification information is defined here as any image data useful for identifying the user via a representation of themself (as opposed to e.g. interpolation from a visual record of their surroundings). As outlined in Table 4.3, the use of `Visual→Avatar` images (also known as profile images) is widespread, but they are not always photographic representations of the user. Nonetheless, a consistently-adopted avatar may be identifiable even if it does not reflect the user's actual appearance – there is further discussion of the (lack of a) relationship between truthful and identifiably useful information in Chapter 6.

Supplementary to avatars are `Visual→Banner` images, which are used as a heading or background to a profile page. These again may reflect a consistent visual presentation of the user across different platforms, perhaps even a photographic representation.

Images uploaded by the user (Section 4.3.9) may include photographs containing their physical appearance, but this is not assured except where specific tagging is available. This sort of specific identification of photographic subjects is as yet not widespread, concentrated in only a few popular services.

Table 4.3 Visual identifiers

Name	Description	#Services
Avatar	A user's chosen photo or avatar representation.	23
Banner image	A background or banner image, usually chosen by the user to complement their profile image.	13
Tagged photos	Photographs labelled as containing the user, usually as identified by the service or other users, but accessible via the user's profile.	3

4.3.4 Opinion

Opinion markers (e.g. favourites, star-ratings) are expressions of a user's interests and judgement about either other users' content on a platform, brands represented within the platform in some fashion, or off-platform content represented by URLs which a user may express opinions on. Note that this is profile information about what a user has expressed opinions on, rather than the networks' opinions on content produced by the user themselves (this popularity information is covered in Section 4.3.7). While some opinion codification may be platform-specific, opinion markers may be combined to identify a user by tabulating their expressed opinions about the same entities on different platforms.

Textually expressed opinion is not included in this attribute set, as extracting opinion from text requires additional inference, and this schema is only interested in structurally supported attributes. As Table 4.4 shows, opinion markers are most common for within-platform content, either that produced by other users or the social presence of various brands. Brands themselves are less commonly available for approval information, and off-platform content is least likely of all to have popularity with a user tracked on their profile page. However, all opinion markers were comparatively uncommon, present across at most a third of the surveyed sites.

Table 4.4 Opinion markers

Name	Description	#Services
Content	This user's opinion of on-platform user content.	12
Brand	This user's opinion of a product, company or item on the platform.	7
Other	The user's opinion of off-platform items.	2

4.3.5 Temporal

As Table 4.5 shows, activity timestamps are common, with only two services not revealing any of these. The two other information items in this category – join date and last activity – are not as common but can be partially inferred from these activity logs. Not recorded here is the specificity of time reporting – some sites present detailed timestamps, while others provide vaguer information as items age (i.e. 'a minute ago', '3 hours ago', 'a day ago', '1 week ago').

Table 4.5 Temporal information

Name	Description	#Services
Activity timestamps	Date and time stamps for user posts and activity, accessible via the profile page.	33
Membership date	When a user created their account.	17
Last seen	The date and time at which the user was last recognised by the service.	9

4.3.6 Geographical

Geographical information seems more closely guarded by services than temporal information. As Table 4.6 shows, the most common form is an indicator of the user's location in a city or country, with varying levels of specificity. In interpreting this table, note that location history simply extends a location set with temporal information, so both are counted where the first is available.

Table 4.6 Geographical Information

Name	Description	#Services
Current Location	The user's address or advertised location.	12
Location Set	A set of locations associated with the user.	6
Location History	A timestamped set of user locations.	3

4.3.7 Degree

Degree information is information about the user's popularity, dedication or influence within a platform. Most commonly this is linked to social connectedness within a network's population, particularly `Degree→Subscribers`, the number of other users following[6] the profile owner. However, other measures track different kinds of integration with the network.

The second most common degree information is the reflection of subscriber information, `Degree→Subscribed`, which tracks the number of users the profile owner is following. Equally prevalent is information about the number of contributions the profile owner has made to the platform, in terms of e.g. posts made.

Trailing these markers are, jointly, `Degree→Visibility`, `Degree→Reputation` and `Degree→Trophies`. The first reflects the number of times other users have viewed the profile, the second aggregates opinion about the profile owner, and the third indicates specific notable events or accomplishments. The final item, `Degree→Rank`, highlights official membership level in a platform.

Table 4.7 Degree metrics

Name	Description	#Services
Subscribers	The number of other users following this user.	21
Subscribed	The number of other users that this user follows	16
Contributions	The number of contributions that this user has made to the network (links, posts, videos, etc.)	16
Visibility	The number of views of this user's profile.	9
Reputation	This user's reputation, as rated by other users or other performance metrics.	9
Trophies	Markers for a user's special achievements.	9
Rank	The user's rank within the community, as one of a number of possible tiered levels.	5

[6]Following being where a user is publicly listed as receiving updates about the subject.

4.3.8 Relationships

Relationship information is that which explains a connection to another entity within a platform. The relationship information presented in Table 4.8 is broken down into two components:

1. relationships the user has with other people in a platform's network and

2. relationships that the user has with brands within the network;

where brands are either companies, products, or media items which are not uploaded by other users.

Most available relations were simple interaction relationships – information that reveals that a particular pair of users interacted in some way within the platform[7]. The `Relationships→FollowsUser` was also widespread, with the `Relationship→FollowedBy` and `Relationship→FollowsBrand` relationship information significantly less common.

In gathering data, where relationships in a platform's network are bilateral friendships, both `FollowsUser` and `FollowedBy` are considered fulfilled.

4.3.9 Content

Content consists of the available record of a user's submissions or publications via the service in question. Table 4.9 outlines the different categories of content which may be published, whereas Table 4.10 shows the attributes which each item of content – of whatever type – may have.

As can be discerned from these two tables, the most common attributes were temporal and impact, and the most common content type was plain text. Structural support for textual content is by far the most flexible, as it can easily carry URLs referencing other content types within a text body. While temporal information is quite common, geographic annotation on content was the least commonly available.

[7]For example, commenting on another's content.

Table 4.8 Relationships

Name	Description	#Services
	People	
Interacted	A user this user has interacted with.	26
Follows User	A user following this user.	20
Followed By	A user this user follows.	14
Grouped	A user this user shares a group with.	10
	Brands	
Follows Brand	A product, company or item this user follows.	13
Contributor	A product, company or item this user has interacted with or is marked as affiliated with.	6

Table 4.9 Content types

Name	Description	#Services
Text	Typed text content, not necessarily exclusive of images or other media	29
Image	Still images.	17
Video	Video content.	11
Links	Links to content from outside the platform	5

Table 4.10 Content attributes

Name	Description	#Services
Temporal	Timestamp for a post.	31
Impact	The content item's popularity, in terms of any of various vote or viewing measures.	31
Category	A categorisation or tag for the content item.	12
Spatial	Geographical locator for the content	3

The following sections examine collected data and existing literature to calculate various measures for each of the attributes listed in this schema.

4.4 Availability

The *availability* of information attributes is the probability of finding an information attribute in a record. In the field of identity resolution, this is the probability of finding a profile which contains a particular information attribute.

Availability is highly important for understanding an attribute's practical value for identification. When a piece of information known from one profile is likely to be present in a different online social networking profile, this provides comparable profile attributes which can be used to establish a probability that the two profiles are in fact of the same person. In the extreme case, if two social networks shared no overlap in the type of profile content made available, it would be impossible to establish a direct connection between profiles drawn from either of them. Proportionally, the greater the overlap in profile content, the more opportunities for comparison between profile attributes, and a greater wealth of data for any re-identification algorithm[8].

There are two components to availability:

Structural Support for the attribute. The estimated proportion of a class of datasets (such as social networks) which contain a given profile attribute. For example, if an online social network does not include a gender field for users to fill out, it does not have structural support for the gender attribute. If many networks do not support gender fields, then this decreases the identification value of explicit gender declarations.

Completeness of this attribute within profiles. The proportion of profiles which report non-null values for this attribute within datasets which permit it. While structural support indicates the possibility of finding a match, this is also affected by the probability of users actually providing a value for this attribute. In many online profiles, there are optional attributes which may or may not be given values. If gender is structurally supported, but in practice few users report their gender, then

[8]Though the utility of this information also depends on how consistent and unique the attributes are, as discussed in later sections.

the availability of gender information is lower and thus the identification value of gender information is lower.

These properties are derived independently in Sections 4.4.1 and 4.4.2 for each of the profile attributes in the schema. The estimates are then combined and evaluated in Section 4.4.3.

4.4.1 Support for profile fields

The property called *structural support* refers to how widely an attribute is likely to be found within the reference class of datasets. For any class of datasets P and an attribute $a \in A$, which an investigator is interested in understanding the identification value of, structural support is $E(P_a)$, the estimated proportion of P which contain the attribute a. Within the realm of social networking profiles, this would be an estimate of the number of online social networks P which can reveal the profile attribute a.

This structural support value is to be kept distinct from considerations of how likely it is that any particular profile page within a network P_i will contain a, which is addressed separately in Section 4.4.2.

The degree of structural support for various identity attributes is a topic which appears to have attracted little specific research attention, though some studies [18, 37, 109] have included data encompassing it as part of investigations into completeness or consistency of profile attributes.

In gauging the degree of structural support for profile attributes, it is important to consider the limitations within which the work takes place. It is not possible to anticipate the internal structure of all the private datasets which may be combined with public profile attributes, which could include datasets as varied as medical records, insurance databases or job applications. Instead, support for the model is grounded in observations of online social networks themselves.

For the most part, the measurement of structural support is grounded in the same survey of 35 high-traffic unique sites which was used in constructing the schema of

profile attributes itself. However, where possible this data can be reinforced with figures from previous work.

Carmagnola et al. [37] report a core list constructed from the registration form data of 25 user-adaptive systems from a number of domains. They identified `Bio→Username` as the most re-occurring feature, closely followed by first and last name fields[9], then `Contact→Email`, birth date (`Bio→Age`) and finally birthplace, which can be considered an item of `Geographical→Locations`.

Irani et al. [109] include some information on structural support in their study of attribute completeness across 10 popular online social networks. The attributes they report on are represented in our schema as `Contact→Web`, `Bio→Username`, `Bio→Relationship`, `Bio→Relationship`, `Bio→Age` and `Geographical→Current`.

Balduzzi et al. [18] reported on the vulnerability of online social networks to profile-harvesting attacks based on large lists of email addresses. As part of their reporting, they covered the availability of a number of different profile attributes across the eight social networks targeted (Facebook, MySpace, Twitter, LinkedIn, Friendster, Badoo, Netlog and XING).

While the methodologies and ontologies are not directly comparable, and many of the profile attributes of interest are not considered, these additional studies do provide some verification for the structural support figures collected in forming the schema. In both, `Bio→Username` was nigh-universal, and less than half of sites reported `Bio→Age`. Carmagnola et al. did report a notably higher proportion of services using `Contact→Email`, no doubt a result of their method of examining forms rather than publicly-visible information. Email addresses can be used to locate profiles within services even where the address is not made publicly visible, so adjusting structural support here to reflect Carmagnola's count would be justified.

Table 4.11 lists the profile attributes of the schema, along with the raw counts of these attributes from the 35 top-ranked dataset from Section 4.3 and the three supplementary counts from the literature [18, 37, 109]. For each attribute, a structural support score

[9]The schema used in this thesis considers these fragments of an alternative username

Table 4.11 Quantified structural support for profile attributes

Attribute	*Bdzzi. (8)*	*Carmag. (25)*	*Irani (10)*	*Edwards (35)*	*Score*
Contact→Web	5	-	5	11	0.40
Contact→Profile	-	-	-	9	0.26
Contact→Email	-	16	-	5	0.35
Contact→Phone	3	-	-	3	0.14
Bio→Username	8	22	9	35	0.95
Bio→Description	-	-	-	21	0.60
Bio→Age	5	8	4	10	0.35
Bio→Tags	-	-	-	10	0.29
Bio→Education	7	-	-	8	0.35
Bio→Occupation	7	-	-	8	0.35
Bio→Gender	5	-	4	6	0.28
Bio→Relationship	6	-	2	6	0.27
Bio→Sexuality	4	-	-	4	0.19
Bio→Verified	-	-	-	3	0.09
Bio→Religion	-	-	-	3	0.09
Bio→Physical	2	-	-	2	0.06
Bio→Habits	8	-	-	2	0.23
Visual→Avatar	8	-	-	23	0.72
Visual→Banner	-	-	-	13	0.37
Visual→Photos	-	-	-	3	0.09
Opinion→Content	-	-	-	12	0.34
Opinion→Brand	-	-	-	7	0.20
Opinion→Other	-	-	-	2	0.05
Temporal→Activity	-	-	-	33	0.94
Temporal→Membership	-	-	-	17	0.49
Temporal→Seen	3	-	-	9	0.28
Geographical→Current	8	-	6	12	0.50
Geographical→Locations	-	6	8	6	0.29
Geographical→History	-	-	-	3	0.09
Degree→Subscribers	-	-	-	21	0.60
Degree→Subscribed	-	-	-	16	0.46
Degree→Contributions	-	-	-	16	0.46
Degree→Visibility	2	-	-	9	0.26
Degree→Reputation	-	-	-	9	0.26
Degree→Trophies	-	-	-	9	0.26
Degree→Rank	-	-	-	5	0.14
Relationships→Interacted	-	-	-	26	0.74
Relationships→FollowsUser	7	-	-	20	0.63
Relationships→FollowedBy	7	-	-	14	0.49
Relationships→Grouped	-	-	-	10	0.29
Relationships→FollowsBrand	-	-	-	13	0.37
Relationships→Contributes	-	-	-	6	0.17
Content→Text	-	-	-	29	0.83
Content→Image	-	-	-	17	0.49
Content→Video	-	-	-	11	0.31
Content→Links	-	-	-	5	0.14
Attributes→Temporal	-	-	-	31	0.89
Attributes→Impact	-	-	-	31	0.89
Attributes→Category	-	-	-	12	0.34
Attributes→Spatial	-	-	-	3	0.06

is calculated from the proportion of sites containing it out of all those examined for it. Those attributes with a structural support score greater than 0.6 are highlighted in bold font.

The most widely supported attribute was `Bio→Username`, which may have been expected given its criticality to the concept of a profile page. This was followed closely by `Temporal→Activity`, indicating a surprisingly widespread support for timestamp information, and then temporal and popularity-related attributes of user-generated content, both attributes being more widely supported than any particular form of user-generated content, even `Content→Text`, which was itself widely supported. Some interesting distinctions can be drawn between apparently related measures: `Degree→Subscribers` corresponds to `Relationships→FollowedBy`, but the degree information is more widely supported than actual user connections, and the reverse is true of `Degree→Subscribed` and `Relationships→FollowsUser`.

The attributes most typically considered in database studies, such as `Bio→Age` and `Bio→Gender` were not very widely supported, with only the biographical attribute `Bio→Description`, the freeform self-descriptive text, being supported on a significant proportion of examined platforms. In contrast, visual information in the form of `Visual→Avatar` is widely supported in social networks. This highlights the domain-specific differences between identity resolution in online social networks and record linkage in different domains.

4.4.2 Completeness of profile information

The second component of availability is *completeness*. Complementary to *structural support*, this measures the proportion of records which report non-null values for an attribute, given that the dataset in question contains the attributes. In the field of identity resolution, this means the proportion of user profiles where users have filled out (and made visible) the information attribute in question. Certain profile items might be made mandatory by a network which structurally supports them, but more often than not these profile items are either optional or can be hidden from the public.

Completeness of profile information is a component of availability which has been given some attention in previous literature, including literature on identity resolution. These prior estimates are gathered in Table 4.12.

One of the earlier studies, Dyer et al. [66] tabulates self-reports of disclosure from 69 Facebook members and 48 Myspace members, sampled on an availability basis. These 2007 figures show notably higher availability of information than later studies, reflecting either sampling bias or a general trend to less disclosure as privacy risks are popularised. Notably, these figures are derived from questionnaire responses rather than direct observation.

In later studies based on observation:

- Irani et al. [109] study the completeness of a few attributes across a range of online social networks, including a total of 21,764 profiles across 10 of the most common social networking sites, though some of their attributes were not structurally supported across all 10 sites.

- Nosko et al. [162] present a study of information disclosure in 400 Facebook profiles, reporting the number of profiles which filled in each profile attribute. Their profile schema maps easily to the one presented in Section 4.3, and they study the broadest range of attributes outside of this thesis.

- Balduzzi et al. [18] present attribute completeness information for a number of attributes, based on a set of 1,228,644 profiles drawn unevenly from eight different social networks. The proportions reported in Table 4.12 are derived by recombining the figures per-attribute across networks where the attributes were supported.

- Chen et al. [45] present attribute completeness information for some optional attributes within the English YouTube social network, based on a sample of some 26,562 profiles. Their presentation divides this into profiles which use real names and pseudonyms based on identification from connected LinkedIn accounts – these figures are merged by a 1:4 ratio. The figures reported in Table 4.12 are extracted from a graph and may contain error in the 0.01 to 0.02 range.

Table 4.12 Prior measurements of profile attribute completeness

Attribute	*Goga15*	*Bdzzi*	*Dwyer*	*Chen*	*Goga13*	*Irani*	*Noska*	*Score*
Contact→Web	-	-	-	-	-	-	0.08	0.08
Contact→Profile	-	-	-	-	-	-	-	-
Contact→Email	-	-	0.72	-	-	-	0.43	0.58
Contact→Phone	-	-	0.25	-	-	-	0.05	0.15
Bio→Username	1.00	-	-	-	1.00	0.99	-	1.00
Bio→Description	-	-	-	0.30	-	-	0.31	0.31
Bio→Age	-	0.34	-	-	-	0.75	0.73	0.61
Bio→Tags	-	-	-	0.17	-	-	0.47	0.25
Bio→Education	-	0.15	-	0.15	-	-	0.83	0.38
Bio→Occupation	-	0.27	-	0.29	-	-	0.35	0.30
Bio→Gender	-	0.30	-	-	-	0.84	0.82	0.65
Bio→Relationship	-	0.06	0.81	-	-	0.63	0.63	0.53
Bio→Sexuality	-	0.04	0.77	-	-	-	0.55	0.45
Bio→Verified	-	-	-	-	-	-	-	-
Bio→Religion	-	-	-	-	-	-	0.32	0.32
Bio→Physical	-	-	-	-	-	-	-	-
Bio→Habits	-	-	-	-	-	-	0.37	0.37
Visual→Avatar	0.63	0.54	0.99	-	0.93	-	0.98	0.81
Visual→Banner	-	-	-	-	-	-	-	-
Visual→Photos	-	-	-	-	-	-	0.76	0.76
Opinion→Content	-	-	-	-	-	-	-	-
Opinion→Brand	-	-	-	0.10	-	-	0.51	0.31
Opinion→Other	-	-	-	0.10	-	-	-	0.10
Temporal→Activity	-	-	-	-	-	-	0.98	0.98
Temporal→Membership	-	-	-	-	-	-	-	-
Temporal→Seen	-	0.81	-	-	-	-	-	0.81
Geographical→Current	0.54	0.43	-	-	0.46	0.77	0.08	0.46
Geographical→Locations	-	-	0.93	0.38	-	0.72	-	0.68
Geographical→History	-	-	-	-	-	-	-	-
Degree→Subscribers	-	-	-	-	-	-	0.88	0.88
Degree→Subscribed	-	-	-	-	-	-	0.88	0.88
Degree→Contributions	-	-	-	-	-	-	0.25	0.25
Degree→Visibility	-	0.75	-	-	-	-	-	0.75
Degree→Reputation	-	-	-	-	-	-	-	-
Degree→Trophies	-	-	-	-	-	-	0	0
Degree→Rank	-	-	-	-	-	-	-	-
Relationships→Interacted	-	-	-	-	-	-	0.83	0.83
Relationships→FollowsUser	0.62	0.81	-	-	-	-	0.88	0.77
Relationships→FollowedBy	0.62	0.81	-	-	-	-	0.88	0.77
Relationships→Grouped	-	-	-	-	-	-	0.79	0.79
Relationships→FollowsBrand	-	-	-	-	-	-	-	-
Relationships→Contributes	-	-	-	-	-	-	-	-
Content→Text	-	-	-	-	-	-	0.26	0.26
Content→Image	-	-	-	-	-	-	0.70	0.70
Content→Video	-	-	-	-	-	-	0.13	0.13
Content→Links	-	-	-	-	-	-	-	-
Attributes→Temporal	-	-	-	-	-	-	-	-
Attributes→Impact	-	-	-	-	-	-	-	-
Attributes→Category	-	-	-	-	-	-	-	-
Attributes→Spatial	-	-	-	-	-	-	-	-

- Goga et al. [85] also report on the completeness of three attributes of interest (`Bio→Username`, `Visual→Avatar` and `Geographical→Current`) across Twitter, Facebook, Google+, Myspace and Flickr.

- Goga et al. [84] later report on the completeness of `Bio→Username`, `Visual→Avatar`, `Geographical→Current` and `Relationships→FollowsUser` across a collection Twitter, LinkedIn, Flickr and Facebook profiles. Their analysis focused on paired network availability and a comparison of different collection methods, so per-network totals were not available for merging and simple averages were used to produce the summary figures reported here.

Table 4.13 complements previous research with novel measurements of availability. There are two components to these original measurements. The first is an automatic appraisal of profile content drawn from a sample of some 52,641 profiles across Facebook, Google+, Twitter and LinkedIn, sampled according to the methodology explained in Section 3. API access to user profiles can be incomplete, however, and as a result certain attributes cannot be retrieved through this method. As a result, a manual appraisal was made of a subset of 100 of these profiles, with 25 profiles drawn randomly from each network, forming a second measurement.

The manual appraisal was conducted with the developer profile which authorised API access for the automated data collection, which has no friendship relation to any of the targeted profiles. The profile page for each member was loaded and examined for all the profile attributes supported by the network. The proportions reported reflect completeness *within networks supporting the attribute*.

These automatic and manual appraisal values are then combined with the prior measurement information from the other studies detailed in Table 4.12 to provide a final estimate of attribute completeness.

Table 4.13 highlights in bold font the attributes with an overall *completeness* estimate exceeding 0.60. Notably, all `Temporal→*` attributes were highly complete, perhaps due to being reported by default in networks, often without the user having to fill anything out. This interpretation would be supported by the high *completeness* values

Table 4.13 Original measurements of profile attribute completeness

Attribute	*Automated*	*Manual*	*Priors*	*Score*
Contact→Web	0.09	0.14	0.08	0.10
Contact→Profile	0.05	0.20	-	0.13
Contact→Email	0.00	0.01	0.58	0.20
Contact→Phone	0.00	0.01	0.15	0.05
Bio→Username	1.00	1.00	1.00	1.00
Bio→Description	0.54	0.32	0.31	0.39
Bio→Age	$>$ 0.00	0.05	0.61	0.22
Bio→Tags	0.05	0.30	0.25	0.20
Bio→Education	0.12	0.53	0.38	0.34
Bio→Occupation	0.09	0.49	0.30	0.29
Bio→Gender	0.59	0.96	0.65	0.73
Bio→Relationship	0.02	0.10	0.53	0.22
Bio→Sexuality	-	0.28	0.45	0.34
Bio→Verified	0.01	0.02	-	0.02
Bio→Religion	-	0.03	0.32	0.16
Bio→Physical	-	0.00	-	0.00
Bio→Habits	-	0.15	0.37	0.22
Visual→Avatar	0.88	0.66	0.81	0.78
Visual→Banner	0.06	0.37	-	0.22
Visual→Photos	-	0.04	0.76	0.40
Opinion→Content	-	0.31	-	0.31
Opinion→Brand	-	0.19	0.31	0.23
Opinion→Other	-	0.05	0.10	0.07
Temporal→Activity	0.40	0.54	0.98	0.64
Temporal→Membership	0.99	1.00	-	1.00
Temporal→Seen	-	-	0.81	0.81
Geographical→Current	0.51	0.62	0.46	0.53
Geographical→Locations	0.51	0.65	0.68	0.61
Geographical→History	0.04	0.25	-	0.15
Degree→Subscribers	0.55	0.78	0.88	0.74
Degree→Subscribed	0.43	0.75	0.88	0.69
Degree→Contributions	0.66	0.80	0.25	0.57
Degree→Visibility	-	1.00	0.75	0.92
Degree→Reputation	0.49	0.64	-	0.57
Degree→Trophies	-	0.10	0.00	0.05
Degree→Rank	-	0.01	-	0.01
Relationships→Interacted	0.28	0.53	0.83	0.55
Relationships→FollowsUser	0.39	0.67	0.77	0.61
Relationships→FollowedBy	0.34	0.62	0.77	0.58
Relationships→Grouped	-	0.25	0.79	0.43
Relationships→FollowsBrand	-	0.42	-	0.42
Relationships→Contributes	-	0.30	-	0.30
Content→Text	0.32	0.46	0.26	0.35
Content→Image	0.05	0.38	0.70	0.38
Content→Video	0.04	0.19	0.13	0.12
Content→Links	0.23	0.29	-	0.26
Attributes→Temporal	0.39	0.52	-	0.46
Attributes→Impact	0.38	0.41	-	0.40
Attributes→Category	0.19	0.06	-	0.13
Attributes→Spatial	0.03	0.22	-	0.13

for `Degree→Visibility`, which is also tracked by the platform rather than the user. In contrast, `Contact→*` had poor completeness, with most users keeping private contact details such as email addresses, phone numbers and other social media profiles as well as many `Bio→*` attributes (though not `Bio→Gender`). However, `Visual→Avatar` is highly complete, despite being user-provided, indicating that certain personalisations are commonly completed.

No `Content→*` or `Attributes→*` were highly complete, suggesting a high proportion of social media users who either keep private or do not submit content to their social networks. The prior measurements from related work [162] do back up these unexpected results.

4.4.3 Estimates of availability

The combination of the *structural support* and *completeness* components allows us to attach an overall availability rating to the information items in our schema. Table 4.14 combines the results columns from Table 4.11 and Table 4.13 to produce the overall availability scores for each of the measured attributes. This estimation assumes that equal weight should be given to considerations of completeness and support, but in certain situations this assumption should be reviewed. For example, in an application to identity resolution in a specific target network, *support* becomes a binary variable and so considerations of *completeness* should dominate the availability valuation of information attributes.

There are important differences between the completeness and support information for which analysis may be fruitful. To facilitate this, the terms *very low*, *low*, *average* and *high* in the following discussion are mapped to the interquartile ranges of the data as presented in Table 4.15.

Contact information is not very often available on social networks. The most available attribute was `Contact→Email`, which still had *low* availability, while `Contact→Profile` and `Contact→Phone` had *very low* availability. `Contact→Email` was the only contact detail to have *low* rather than *very low* completeness, and this measurement can be

Table 4.14 Estimates of profile attribute availability

Attribute	*Completeness*	*Support*	*Availability*
Contact→Web	0.10	0.40	0.25
Contact→Profile	0.13	0.26	0.19
Contact→Email	0.20	0.35	0.28
Contact→Phone	0.05	0.14	0.10
Bio→Username	1.00	0.95	0.98
Bio→Description	0.39	0.60	0.50
Bio→Age	0.22	0.35	0.29
Bio→Tags	0.20	0.29	0.25
Bio→Education	0.34	0.35	0.35
Bio→Occupation	0.29	0.35	0.32
Bio→Gender	0.73	0.28	0.51
Bio→Relationship	0.22	0.27	0.25
Bio→Sexuality	0.34	0.19	0.27
Bio→Verified	0.02	0.09	0.06
Bio→Religion	0.16	0.09	0.13
Bio→Physical	0.00	0.06	0.03
Bio→Habits	0.22	0.23	0.23
Visual→Avatar	0.78	0.72	0.75
Visual→Banner	0.22	0.37	0.29
Visual→Photos	0.40	0.09	0.25
Opinion→Content	0.31	0.34	0.33
Opinion→Brand	0.23	0.20	0.22
Opinion→Other	0.07	0.05	0.06
Temporal→Activity	0.64	0.94	0.79
Temporal→Membership	1.00	0.49	0.75
Temporal→Seen	0.81	0.28	0.55
Geographical→Current	0.53	0.50	0.52
Geographical→Locations	0.61	0.29	0.45
Geographical→History	0.15	0.09	0.12
Degree→Subscribers	0.74	0.60	0.67
Degree→Subscribed	0.69	0.46	0.58
Degree→Contributions	0.57	0.46	0.52
Degree→Visibility	0.92	0.26	0.59
Degree→Reputation	0.57	0.26	0.42
Degree→Trophies	0.05	0.26	0.16
Degree→Rank	0.01	0.14	0.08
Relationships→Interacted	0.55	0.74	0.65
Relationships→FollowsUser	0.61	0.63	0.62
Relationships→FollowedBy	0.58	0.49	0.54
Relationships→Grouped	0.43	0.29	0.36
Relationships→FollowsBrand	0.42	0.37	0.40
Relationships→Contributes	0.30	0.17	0.24
Content→Text	0.35	0.83	0.59
Content→Image	0.38	0.49	0.44
Content→Video	0.12	0.31	0.22
Content→Links	0.26	0.14	0.22
Attributes→Temporal	0.46	0.89	0.68
Attributes→Impact	0.40	0.89	0.65
Attributes→Category	0.13	0.34	0.24
Attributes→Spatial	0.13	0.06	0.10

Table 4.15 Term mappings for discussion of completeness, structural support and availability

Term	"very low"	"low"	"average"	"high"
Range	0 - 1st Q	1st - 2nd Q	2nd - 3rd Q	3rd Q - 1
Support	0 - 0.20	0.21 - 0.37	0.37 - 0.49	0.50 - 1
Completeness	0 - 0.17	0.17 - 0.37	0.38 - 0.57	0.58 - 1
Availability	0 - 0.21	0.22 - 0.37	0.38 - 0.54	0.55 - 1

traced back to prior research into profile completeness – the original measurements by themselves would indicate a *very low* completeness. `Contact→Web` also had *low* availability, but in this case the attribute's *average* structural support is counteracting a *very low* completeness – the option to link to a personal webpage is reasonably well-supported, but users rarely take advantage of the opportunity.

`Bio→Username` is easily the most available profile attribute, with some form of name field supported by a *high* proportion of all social networking sites and universally completed within our observations. Other biographical attributes performed less robustly. `Bio→Description` and `Bio→Gender` both managed *average* totals by combining one *high* score with a *low* or *average* one, whilst the remaining attributes were had *low* or *very low* availability, with the same for their support and completeness scores. Unless the information carried by these latter attributes can be otherwise inferred from profile content, identity-resolution methods which rely on them are not going to have broad applicability. Most notably, these figures indicate a *low* availability of tag information, an area where some previous identity resolution studies have focused [103, 214].

Amongst visual information, `Visual→Avatar` was the stand-out performer, with *high* availability backed by *high* scores in both completeness and support, making it a desirable feature. `Visual→Banner` images had *low* completeness and *low* support, whereas `Visual→Photos` were estimated to have *average* completeness but *very low* structural support. This suggests that the visual information most available to identity resolution classifiers is un-annotated and potentially non-photographic.

Structural information about user opinion is not widely available in social networks, with *low* availability, completeness and support for `Opinion→Content` and `Opinion→Brand`, dropping to *very low* when it comes to `Opinion→Other`. Given that the former

pair refer to entities that only exist within the same online social network, finding equivalences between sets of opinions in different networks seems unlikely to be profitable as an identity-resolution method.

Temporal information as a class has a very strong showing, with uniformly *high* completeness backing a range of support values. `Temporal→Activity` benefited from the highest support score, rendering it a very available feature, whilst `Temporal→Seen` suffered from *low* support which might hinder the general application of methods relying on it.

`Geographical→Current`, a single location associated with the user, has *average* availability based on *average* completeness and *high* support. In contrast, `Geographical→Locations`, a set of associated locations, has *average* availability based on *high* completeness and *low* support. Both are plausibly available enough to be of use in identity resolution. However, `Geographical→History`, locations associated with timestamps, are of *very low* availability, and so methods which rely on spatio-temporal rather than purely spatial methods will struggle to apply to many use cases.

Degree information recovers from a mostly *low* or *average* support by several times reaching *high* completeness. As a result, three degree attributes have *high* availability: `Degree→Subscribers`, `Degree→Subscribed` and `Degree→Visibility`, supported by strong *average* availability for `Degree→Reputation` and `Degree→Visibility`. Data on trophies or in-network rankings were least available, being *very low*.

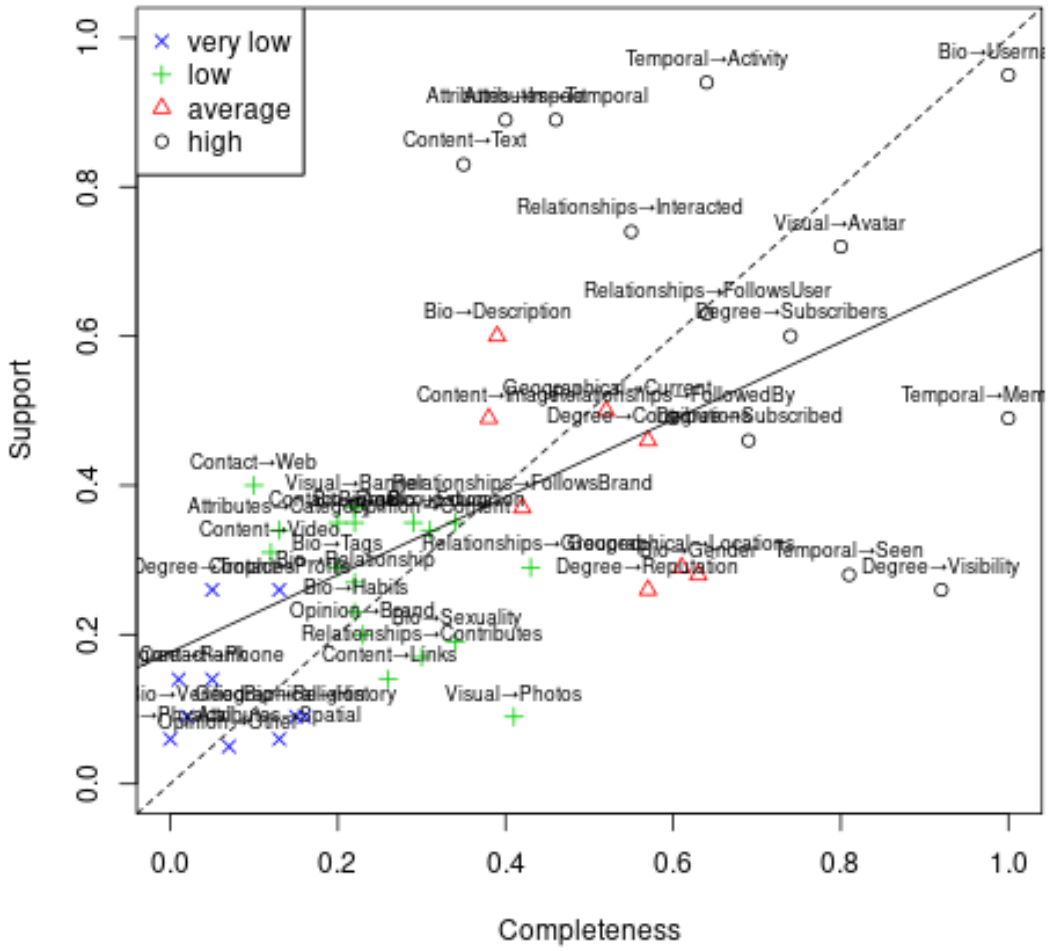
Amongst relationship information, the most available was `Relationships→Interacted`, with a *high* availability backed by *high* support and completeness. This could be attributed to online interactions being an inherent component of the usage of OSNs. The same follows for the two following-related relationships, `Relationships→FollowsUser` and `Relationships→FollowedBy`. Connections to brands were of a more *average* availability, with *low* support, but relationships where the user reported an active contribution to the brand were of *low* availability. Information tying users together via groups was also of *low* availability, despite an *average* completeness.

Content items suffer from *low* or *very low* completeness across all types, pointing to users either keeping submissions private or not making any submissions to their social networks. In the case of the very highly supported `Content→Text`, this is remedied sufficiently to reach a *high* availability, while an *average* structural support score pitches `Content→Image` at an equivalently *average* level of availability, whilst `Content→Video` and `Content→Link` are of only low availability.

Counter-intuitively, some content attributes were more available than particular content types, due to being able to cross the content type boundaries. `Attributes→Temporal` and `Attributes→Impact` had *high* availability due to *high* support and *average* completeness. However, this was not a universal rule. `Attributes→Category` was only of *low* availability, with *very low* completeness, and `Attributes→Spatial` was of *very low* overall availability. That temporal annotations are widely available is also reflected within `Temporal→Activity` and previous time-based profiling work, but the high availability of popularity information for user content might suggest a viable new method for identity resolution.

Figure 4.3 shows a comparison of *completeness* and *support* scores for all attributes. There was a moderate positive correlation (r=0.57) between the two measurements, indicating that more broadly-supported attributes are also more likely to be completed on user profiles within networks which support them.

Fig. 4.3 Comparison of completeness and stuctural support scores



4.5 Consistency

Consistency, as given by Eq. 4.10, can be understood as a measure of how well an information attribute remains the same across different datasets about the same entity. In the case of online social networks, this becomes a measure of how closely matched attributes (such as username or textual content) are across a user's profiles on different networks. Consistency is highly important for identity resolution, as commonly-mismatched information items are less valuable and thus should be given less weight, even if they are highly available and can contain many unique values.

It may be worth asking why the same field, about the same user, should be expected to be anything other than consistent across datasets. Consistency can vary for different reasons. In referent-defined fields, consistency may vary due to any of intentional misreporting (as when a false identity is constructed), novelty in self-expression (as when asked to describe themselves), considerations of format (a Tweet may necessarily be shorter than a blog post) or even a changing truth value (as when the user ages, or changes address, and does not update an old profile). In fields which are not completed by the referred individual, values may be entered incorrectly (as when a name is misheard or misspelt) or simply stored in a manner which is lossy compared to another instance of the field (a dataset which records only year of birth compared to one which records a full birth date).

4.5.1 Measuring consistency

In measuring consistency of profile attributes or record fields, domain-specific similarity functions need to be defined to complement the matching schema. Below, such similarity functions are presented for each of the attributes in the schema developed in Section 4.3. Where appropriate, similarity functions which have been used in previous literature are reused here, to allow previous measurements to be integrated into the estimates.

Contact comparison

The consistency of contact information does not appear to have been addressed in previous work. All contact information consists of a set of unique identifiers, and so a natural approach in this case would be to standardise each contact attribute value to a standard format and then compare the set overlap between contact information on two profiles using the Jaccard similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{4.17}$$

In the case of `Contact→Phone`, phone numbers were standardised by removing all non-numeric characters and replacing any international country code prefixes with a single "0" (the standard national access code) except in the case of US and CA numbers.

For `Contact→Web` addresses, values are standardised by stripping the protocol (`http://` or `https://`) and any query parameters from addresses, leaving only the `domain/path` information. The same applies for `Contact→Profile` addresses. Finally, `Contact→Email` addresses are standardised by reducing all characters to lowercase and stripping any protocol information (`mailto://`).

Biographical comparison

Of all biographical information, `Bio→Username` is the most well-studied. There are a number of suitable string similarity functions which may be applied to comparing names and usernames, but previous studies [84, 139, 178] have mostly used Jaro-Winkler string similarity to compare both username and real-name similarity, based in part on experimental evidence that it is the better performing method [52], and so this is the measure adopted in this study. The best name similarity rating is calculated as the average Jaro-Winkler similarity over the best pairs between the different name tokens in each profile. The same approach is applied to `Bio→Education` and `Bio→Occupation`, as both

attributes are also fields with (potentially multiple) names. String similarity is also used to compare `Bio→Religion` fields, after standardising case.

For measuring the consistency of `Bio→Description`, this study follows the *tf-idf vector space model* approach of Malhotra et al. [139]. All description fields from one profile are merged, punctuation and words from the NLTK list of stop words are removed[10], and terms are lemmatised using the Lancaster stemming algorithm [167]. The Jaccard similarity is then computed between the two resulting token sets. The same could be done for `Bio→Physical`, but data is not available for this attribute. A similar approach is however taken for the tokens `Bio→Tags` and `Bio→Habits`, without the need to strip stopwords.

`Bio→Verified` is standardised to a binary variable, similarity also being binary. `Bio→Gender` is standardised as a trinary value (male, female, other), and `Bio→Sexuality` is standardised to one of a combination of genders (male-male, male-female, male-other, etc.) or 'any' and 'none' values (male-any, male-none, etc.), leading to a total of 15 possible values. Both attributes can also be compared in a binary manner.

Finally, the similarity between two `Bio→Age` values a_i and a_j is calculated as $1 - \frac{|a_i - a_j|}{max(a_i, aj)}$, giving the proportion of the difference in reported ages.

Visual comparison

Previous work has often examined the consistency of `Visual→Avatar` images. The approaches trialled by Malhotra et al. [139] were all based on detecting simple image equality based on an edit distance between a normalised thumbnail. Goga et al. [84] combine two more advanced algorithms in the same general perceptual hashing approach. This study makes use of the fixed-length Discrete Cosine Transform hash employed by default in the *pHash* library of perceptual hashing techniques used by Goga et al. Like all perceptual hashes, this is designed to identify and compare essentially identical images which have been cropped, resized or superficially altered, as may happen between different social media presentations of the same base image. The same method is

[10] http://www.nltk.org/book/ch02.html

used for calculating the similarity of `Visual→Banner` and aggregated across the set of `Visual→Photos`.

Opinion comparison

Due to the difficulties in automatic harvesting and ontological mapping between different social networks, the consistency of `Opinion→*` information is not measured in this study. Theoretically, the problem can be approached as the Jaccard distance between the two sets of positively-rated items (negative opinions being less commonly expressed). For `Opinion→Brand` this might be managed by finding equality between set members based on a string comparison of their names. In the case of `Opinion→Content`, however, there is no necessary mapping to be found between user content on two separate networks. It may be possible to instead aggregate opinion of content into opinion of content owners, using this additional layer to augment similarity metrics based on topological comparison of friend networks. It may be possible to directly compare overlap in the network-independent `Opinion→Other` ratings between profiles, but its low availability makes this ineffective.

Temporal comparison

As `Temporal→Membership` and `Temporal→Seen` are both single dates, both can be compared as a function of the difference δ between two values. In this study, comparisons are made at the day level, and differences of greater than 366 days are rated as a similarity of 0. For values of δ between 1 and 366, the following logarithmic ratio is used:

$$sim(\delta) = 1 - \frac{log(\delta)}{log(366)} \tag{4.18}$$

For `Temporal→Activity`, where multiple dates can be expected, a temporal fingerprinting method developed by Atig et al. [13] is used. This method converts each timestamp to its UTC time value, and divides each day into six activity periods. Timestamps are then allotted to the activity period into which they fall and threshold values are used to identify each activity period as *high* or *low* activity based on the percentage of all

activity allotted to it. The comparison of which activity periods are *high* or *low* between profiles is then used to gauge the similarity of temporal activity.

Geographic comparison

Following the approaches in previous literature [45, 84, 139], this study identifies geographic similarity by the geodesic distance between latitude and longitude coordinates returned by Google's Geocoding API[11], normalised by the maximum possible distance ($sim = 1 - \frac{dist}{max}$), which can be called the *geodesic similarity*. For `Geographic→Current` this measure is sufficient. For `Geographic→History`, distance is calculated as the average geodesic similarity between pairs of locations, where pairs are constructed according to the highest temporal similarity as defined in 4.18. For `Geographic→Locations`, the same approach is used, but pairs are constructed based on the highest geodesic similarity.

Degree comparison

Information about node degrees is susceptible to a range of problems in comparison. Raw counts are dependant in part on the size and nature of network in which they were generated, so it may be, for example, that numbers of friends on Facebook are an order of magnitude greater than numbers of friends on Twitter. This can be addressed by scaling all counts from a network into the [0,1] range according to the highest value observed for that network, but under this method unusually high values can compress the scale, losing important information for most values. This study follows the solution applied by Malhotra et al. [139] in dividing the sorted vector of observed values from a network into k equally-sized bins and then comparing the distance in terms of bins traversed between values, such that similarity is:

$$sim = 1 = \frac{|b_i - b_j|}{k} \tag{4.19}$$

where b_i and b_j are the bin indices for degree values i and j. This measurement is applied for `Degree→Subscribers`, `Degree→Subscribed`, `Degree→Contributions`,

[11] https://developers.google.com/maps/documentation/geocoding/

`Degree→Visibility` and `Degree→Reputation`, as all are measured as a numeric quantity. `Degree→Trophies` and `Degree→Rank` are too rarely available and difficult to map across networks, and so consistency was not calculated for these attributes.

Relationships comparison

For each of the `Relationships→*` attributes, comparisons are made according to the method used by Goga et al. [84] and other related work [156]. Each relationship's similarity is measured as the Jaccard similarity of so-related users between the two profiles, where friend equality across networks is determined by the string similarity of their usernames at a threshold.

Content comparison

The approach for determining consistency between user content is twofold. Firstly, `Content→*` similarity is determined by the Jaccard similarity of each category of content, where equality between content is determined differently for each content type:

`Content→Text` items are reduced to the same standardised word-vector form as is used for `Bio→Description`, and equality determined by a similarity threshold between the two vectors.

`Content→Image` items are compared according to the perceptual hashing approach also used for `Visual→*` comparisons.

`Content→Video` items are compared according the perceptual hashes of initial frames.

`Content→Link` items are compared by first un-shortening any shortened URLs and then comparing the `domain/path` information of the two links to determine equality.

Following the equivalence construction between `Content→*` items, the `Attribute→*` similarity is determined as the average similarity for each paired content item which has the attribute in question in both profiles, where:

- `Attributes→Temporal` similarity is determined by the logarithmic ratio given in Eq. 4.18.

- `Attributes→Impact` similarity is calculated similarly to the `Degree→*` similarity method, with k equally-sized bins for the popularity of each user's posts on each network, and with similarity for each pair quantified as given in Eq. 4.19.

- `Attributes→Category` similarity is determined as the average Jaro-Winkler string similarity between category labels.

- `Attributes→Spatial` similarity is the *geodesic similarity* measure used for `Geographical→*` comparisons.

Additionally, there are a range of stylometric techniques which may be applied to the comparison of `Content→Text` corpora to identify and compare the writing style of authors. Consistency of expressed writing style is examined through comparison of the patterns of function-word usage across the two corpora, following previous work finding function words highly indicative of writing style [125]. The proportional usage of each of a set of N function words is calculated for each profile based on the total collected corpus, forming a fingerprint, and similarity is quantified by the Euclidean distance between the two fingerprints, proportional to the greatest possible difference.

4.5.2 Estimates of consistency

Table 4.16 summarises known estimates of attribute consistency from identity-resolution literature addressing online social networks. In general, information examining the consistency of `Bio→Username` and `Geographical→Current` was most prevalent, with other attributes being studied more sparsely. In more detail:

- Irani et al. [109] investigate the consistency of a range of profile fields within 21,764 matched profiles crawled from an identity management service. Their coverage includes `Contact→Web` `Bio→Age`, `Bio→Relationship` and `Bio→Gender` as well as `Bio→Username` and `Geographical→Current`. Figures reported are extracted from a graph and may contain error in the 0.1 to 0.2 range. In addition, some figures are aggregates of categories reported.

Table 4.16 Prior measurements of profile attribute consistency

Attribute	*Irani*	*Goga13*	*Goga15*	*Malhotra*	*Score*
Contact→Web	0.53	-	-	-	0.53
Contact→Profile	-	-	-	-	-
Contact→Email	-	-	-	-	-
Contact→Phone	-	-	-	-	-
Bio→Username	0.63	0.42	0.84	0.81	0.68
Bio→Description	-	-	-	0.32	0.32
Bio→Age	0.74	-	-	-	0.74
Bio→Tags	-	-	-	-	-
Bio→Education	-	-	-	-	-
Bio→Occupation	-	-	-	-	-
Bio→Gender	0.98	-	-	-	0.98
Bio→Relationship	0.43	-	-	-	0.43
Bio→Sexuality	-	-	-	-	-
Bio→Verified	-	-	-	-	-
Bio→Religion	-	-	-	-	-
Bio→Physical	-	-	-	-	-
Bio→Habits	-	-	-	-	-
Visual→Avatar	-	-	0.21	0.22	0.22
Visual→Banner	-	-	-	-	-
Visual→Photos	-	-	-	-	-
Opinion→Content	-	-	-	-	-
Opinion→Brand	-	-	-	-	-
Opinion→Other	-	-	-	-	-
Temporal→Activity	-	0.13	-	-	0.13
Temporal→Membership	-	-	-	-	-
Temporal→Seen	-	-	-	-	-
Geographical→Current	0.33	0.52	0.70	0.52	0.52
Geographical→Locations	-	-	-	-	-
Geographical→History	-	-	-	-	-
Degree→Subscribers	-	-	-	-	-
Degree→Subscribed	-	-	-	0.01	0.01
Degree→Contributions	-	-	-	-	-
Degree→Visibility	-	-	-	-	-
Degree→Reputation	-	-	-	-	-
Degree→Trophies	-	-	-	-	-
Degree→Rank	-	-	-	-	-
Relationships→Interacted	-	-	-	-	-
Relationships→FollowsUser	-	-	0.49	-	0.49
Relationships→FollowedBy	-	-	0.49	-	0.49
Relationships→Grouped	-	-	-	-	-
Relationships→FollowsBrand	-	-	-	-	-
Relationships→Contributes	-	-	-	-	-
Content→Text	-	0.08	-	-	0.08
Content→Image	-	-	-	-	-
Content→Video	-	-	-	-	-
Content→Links	-	-	-	-	-
Attributes→Temporal	-	-	-	-	-
Attributes→Impact	-	-	-	-	-
Attributes→Category	-	-	-	-	-
Attributes→Spatial	-	-	-	-	-

- Malhotra et al. [139] report on the consistency of `Bio→Username`, `Bio→Description`, `Geographical→Current`, `Visual→Avatar` and `Degree→Subscribed` between 29,129 paired Twitter and LinkedIn accounts. They report consistency as measured via *information gain* between the distributions.

- Chen et al. [45] investigate profile consistency between 179,188 profiles across ten social networks. As well as measuring the consistency of information revelation patterns (the availability of attributes in a user's profile on different networks), they also investigate the consistency attribute values for several attributes, and report on the consistency of the `Geographical→Current` attribute across profiles of the same user, based on the different levels of representation available in Google's Geocoding API.

- Goga et al [83] investigate the consistency of `Bio→Username`, `Temporal→Activity`, `Geographical→Current` and `Content→Test` in the form of accuracy at a 1% FPR between Twitter, Flickr and Yelp profiles.

- Goga et al. [84] investigate profile consistency between 28,211 profiles across five different social networks, looking at `Bio→Username`, `Visual→Avatar`, `Geographical→Current` and `Relationships→FollowsUser`. These figures were calculated using similarity thresholds based on the evaluation of human annotators (i.e. figures are the proportion of matches with both attributes which pass a manually-defined threshold)

Table 4.18 complements the aggregated prior estimates of consistency with original measurement. An automated analysis of consistency was performed on a dataset of 476 pairs of profiles, these being the ground-truth pairings from the same dataset used in Section 4.4.2. For each attribute, the appropriate similarity function was selected and the similarity averaged over the matching profiles.

The attributes with a consistency value > 0.6 are highlighted in bold text in Table 4.18. Note that the measures reported here are very dependent on the similarity functions

Table 4.17 Term mappings for discussion of consistency values

Term	"very low"	"low"	"average"	"high"
Range	0 - 1st Q	1st - 2nd Q	2nd - 3rd Q	3rd Q - 1
Completeness	0 - 0.28	0.29 - 0.50	0.51 - 0.75	0.76 - 1

selected in Section 4.5.1. Following the system from the previous section, terms are once again mapped to interquartile ranges, as delineated in Table 4.17.

Regarding `Contact→*` items, only `Contact→Web` was available enough in the dataset to provide an estimate of consistency, and this consistency being quite low, pulled up from *very low* status by a more middling prior in the literature. This perhaps surprising result suggests that different contact information may be found even on declared connections, perhaps highlighting different social strategies for different networks.

`Bio→*` attributes were comparatively well-represented in previous literature. Of these, `Bio→Gender` was by far the most consistent, with `Bio→Age` trailing at an average rating and `Bio→Relationship` having only low consistency. Our own automated measurements failed to find enough overlap to independently measure the consistency of all these attributes, but these were achieved for `Bio→Username`, which had high consistency, and `Bio→Description` which had very low consistency.

Of `Visual→*` attributes, only `Visual→Avatar` was measurable. The consistency here was low, suggesting that the reuse of the same or highly similar images across profiles is not all that widespread, though the original measurements place a more average weight, suggesting some value which may not be captured in prior work. `Opinion→*` consistency was not measured via automated means or in prior literature.

`Temporal→Activity` registered a low consistency, somewhat greater than the very low level reported in previous observations. `Temporal→Membership` showed very low consistency, indicating that date of membership is not a strong predictor of identity. `Temporal→Seen` was not available enough for comparison in the sampled data.

`Geographical→*` showed very high similarity values across the board. The degree to which these exceed prior measurements is likely due to the scale by which the geodesic distances are normalised. In this data, the scale is set by the maximum distance between antipodes on Earth, whereas it has been more common to use the maximum distance

Table 4.18 Original measurements of profile attribute consistency

Attribute	*Automated*	*Manual*	*Priors*	$C(f_i)$
Contact→Web	0.28	0.18	0.53	0.33
Contact→Profile	-	-	-	-
Contact→Email	-	-	-	-
Contact→Phone	-	-	-	-
Bio→Username	0.84	0.88	0.68	0.80
Bio→Description	0.11	0.07	0.32	0.17
Bio→Age	-	-	0.74	0.74
Bio→Tags	-	-	-	-
Bio→Education	-	0.57	-	0.57
Bio→Occupation	-	0.58	-	0.58
Bio→Gender	-	-	0.98	0.98
Bio→Relationship	-	-	0.43	0.43
Bio→Sexuality	-	-	-	-
Bio→Verified	-	-	-	-
Bio→Religion	-	-	-	-
Bio→Physical	-	-	-	-
Bio→Habits	-	-	-	-
Visual→Avatar	0.58	0.57	0.22	0.46
Visual→Banner	-	0.44	-	0.44
Visual→Photos	-	-	-	-
Opinion→Content	-	-	-	-
Opinion→Brand	-	-	-	-
Opinion→Other	-	-	-	-
Temporal→Activity	0.45	0.60	0.13	0.39
Temporal→Membership	0.20	-	-	0.20
Temporal→Seen	-	-	-	-
Geographical→Current	0.96	0.98	0.52	0.82
Geographical→Locations	0.91	0.62	-	0.77
Geographical→History	0.99	-	-	0.99
Degree→Subscribers	0.74	0.74	-	0.74
Degree→Subscribed	1.00	0.83	0.01	0.61
Degree→Contributions	1.00	-	-	1.00
Degree→Visibility	-	-	-	-
Degree→Reputation	-	-	-	-
Degree→Trophies	-	-	-	-
Degree→Rank	-	-	-	-
Relationships→Interacted	0.02	-	-	0.02
Relationships→FollowsUser	0.12	-	0.49	0.31
Relationships→FollowedBy	0.12	-	0.49	0.31
Relationships→Grouped	-	-	-	-
Relationships→FollowsBrand	-	-	-	-
Relationships→Contributes	-	-	-	-
Content→Text	0.02	0.00	0.08	0.03
Content→Image	0.01	-	-	0.01
Content→Video	-	-	-	-
Content→Links	0.02	0.02	-	0.02
Attributes→Temporal	0.72	1.00	-	0.86
Attributes→Impact	0.86	0.80	-	0.83
Attributes→Category	1.00	-	-	1.00
Attributes→Spatial	0.67	-	-	0.67

between any two points in the dataset. Nonetheless, there appears to be high similarity in locations of matched profiles, a similarity which increases as locations are matched by associated time values.

`Degree→*` information, where measurable, was highly consistent between matched profiles, indicating that a user is likely to have proportionally similar numbers of friends for each network on which they have a profile, and the same for the number of contributions they make.

`Relationships→*` information was of either low or very low consistency. Much of a person's social graph in one network is not present in another according to name-based profile matching, and measurements for some other relationships were not available.

The consistency of `Content→*` and `Attribute→*` items presents an interesting pattern. Although very few of any content items were found to have exact matches in the corresponding profile, the attributes of those content items which were matched were of high or upper average consistency. This is intuitive, as already-matched content should be mostly expected to share attributes – when a user posts the same message on two networks, you expect them to do so at roughly the same time and from roughly the same location, to use similar categories for it and for it to garner proportionally similar levels of attention.

4.6 Uniqueness

The *uniqueness* of information items explores what is typically considered their entropy – how much variation do you see between different users' profile values for a given information attribute? Uniqueness has direct value for linking profiles – consider the example of a profile attribute like gender, which can be highly available and accurate, yet not very identifiable due to the expected proportion of the population which will share a gender value with a particular individual.

4.6.1 Measuring uniqueness

There are two approaches which can be taken when measuring the uniqueness of information carried by an attribute. The first is empirical: to compare the similarity of the attribute between non-matched profiles within a large dataset, and use this internal uniqueness measure as an approximation of this attribute's uniqueness in the domain-general case. The second approach is to approach the attributes theoretically: to identify the expected variation in attribute values, and thus how much information can be represented by this attribute in terms of bits. In this section a combined approach will be taken.

Firstly, an automated comparison is made of the similarity of attribute values across 100,000 non-matched profiles drawn from real-world social network profiles, using the similarity metrics defined in Section 4.5.1. These profiles are drawn from the sampling process that found true matches in the previous section, and described fully in Chapter 3, but comparison is now made between a profile and the specifically-sampled plausible candidate set, excluding the correct match. Secondly, a random subset of 50 non-matching profile pairs are compared in a semi-automated fashion, to include uniqueness estimates for attributes which were not automatically retrieved. Finally, theoretical values are backed by previous literature or understanding of the attribute.

Contact uniqueness

`Contact→*` details are all intended to uniquely identify an individual, and as such should be expected to have the maximum possible uniqueness.

Biographical uniqueness

Biographical details carry varying levels of information, often heavily influenced by the sample size under consideration and the means of gathering said sample. The most well-studied attribute for uniqueness purposes is `Bio→Username`. Perito et al. [178] give a thorough analysis of the identifiability of usernames, and find high precision for the Jaro similarity measure in a dataset of some 100,000 non-matched Google profiles, with graphs indicating precision as high as 0.95 at the threshold level of recall they find

optimum for other methods. However, one can expect collisions to be much more likely within our study, due to the sampling method relying on a relatively small number of surnames.

`Bio→Gender` information should be binary with a expected equal distribution, so the probability of matches across the population would be 0.5, making the likelihood of a correct match only $1/\frac{N}{2}$, which quickly approaches 0 for datasets of any size.

The range of options for `Bio→Religion` could be highly variable if considering free-form selection from all possible religious doctrines, but in practice, four major world religions (Christianity, Islam, Hinduism and Buddhism) cover 77% of the world population. Leaving an option for atheism, and a catch-all category for the remainder, a very rough estimate would suggest a 1-in-6 chance of a false match, though this is unadjusted for the relative populations of these religions in online accounts, and is likely an overestimate.

The number of `Bio→Verified` users is very low compared to non-verified users, but verified status is still common enough that it is not particularly identifying: estimates from Twitter alone suggest some 187,000 distinct identities (though this is still below 1% of Twitter's userbase).

Visual uniqueness

The erroneous similarity level for the perceptual hashing methods used is expected to be below at most 0.2[12]. Accepting this as a bound, a uniqueness estimate can be determined by its complement for `Visual→Avatar` and `Visual→Banner`. Where multiple `Visual→Photos` are available, this tends towards highly uniquely identifying.

Opinion uniqueness

As a set of ratings for potentially highly distinct sets of names, all opinion information can be expected to be highly unique in aggregate.

[12]Based on observations recorded at http://phash.org/docs/design.html

Temporal uniqueness

Under the similarity function defined for `Temporal→Activity`, the possible values are 6 different binary options (high or low activity). Under a uniform assumption, this would result in a 2^{-6} probability that profiles coincidentally match. This is almost certainly an underestimate, as user activity patterns will be heavily influenced by common timeframes. `Temporal→Membership` can be expected to be relatively unique, as there is no particular reason to expect many users to share an exact date of joining a site, whereas `Temporal→Seen` can be expected to have a higher probability of collisions, with active users likely interacting with the site between a certain period of activity.

Geographic uniqueness

The similarity of randomly chosen users' locations should be very low overall, with `Geographical→Current` the most likely to overlap, followed by the other attributes, which are more unique due to their multiplicity.[13]

Degree uniqueness

The method for comparing most `Degree→*` information, given in Section 4.5.1, divides the dataset into n equal portions, so the probability of a false match should in each case be approximately $\frac{1}{n}$, where n in this case is 6.

Relationship uniqueness

The pattern of relationships to similar names, which is expressed by all `Relationships→*` attributes, is highly uniquely identifying.

Content uniqueness

The rate of exact duplication of user content items (`Content→*`) should be expected to be near-nil, with some small allowance for the replication of popular images or slogans. Content attribute similarities can be expected to be more variable. `Attributes→Impact`

[13] This is complicated in measurements by the way the values are normalised.

follows the expected uniqueness of Degree→* attributes due to the shared similarity measurement approach, as Attributes→Temporal follows Temporal→Membership and Attributes→Spatial follows Geographical→Current. Attributes→Category, as a single-word free-text field, could be expected to follow high-uniqueness templates such as Bio→Tags but at a lower degree.

4.6.2 Estimates of uniqueness

Based on the dataset from Section 3 and the methods discussed in the previous section, Table 4.19 presents estimates of uniqueness which can be made for the information attributes in the schema.

Focusing on the results which have been empirically validated, it can be seen that Contact→*, Relationships→* and Content→* items are highly unique, as is the singular biographical item Bio→Description, which uses a similarity function which is related to that of Content→Text. These qualities are all notable in that they are a combination of multiple exact comparisons between profile attributes – contact details, network connections, and content items. The timestamps associated with matched content (Attributes→Temporal) can also be established to be highly unique.

It is worth remarking on the comparably low uniqueness of username data, when considering to the theoretical values grounded by Perito et. al. The sampling methodology used makes particular use of the username feature, so that randomly-matched profiles might be expected to be unusually non-unique in their name values. For usage where a different sampling methodology is being applied, Perito's value should be preferred, but as is argued in Chapter 3, the task solved in the present case more realistically reflects an important identity-resolution challenge.

Other attributes of middling uniqueness values include a smattering of measurable biographical variables (Bio→Age, Bio→Education, and Bio→Occupation). The observed values for occupation and education were less unique than predicted, suggesting that there is a notable bias in reporting of education and occupation, or that the string comparison function used is unduly influenced by similarity of institutional titles. Visual information

Table 4.19 Original measurements of profile attribute uniqueness

Attribute	*Automated*	*Manual*	*Theory*	$U(f_i)$
Contact→Web	>0.99	1.00	1.00	1.00
Contact→Profile	1.00	1.00	1.00	1.00
Contact→Email	-	-	1.00	1.00
Contact→Phone	-	-	1.00	1.00
Bio→Username	0.47*	0.50	0.95	0.64
Bio→Description	>0.99	0.99	0.99	0.99
Bio→Age	-	-	0.80	0.80
Bio→Tags	-	-	0.99	0.99
Bio→Education	0.56	0.48	0.95	0.66
Bio→Occupation	0.54	0.46	0.95	0.65
Bio→Gender	0.00	0.00	0.00	0.00
Bio→Relationship	-	-	0.33	0.33
Bio→Sexuality	-	-	0.45	0.45
Bio→Verified	-	-	0.01	0.01
Bio→Religion	-	-	0.17	0.17
Bio→Physical	-	-	0.99	0.99
Bio→Habits	-	-	0.95	0.95
Visual→Avatar	0.47	0.46	0.80	0.58
Visual→Banner	-	0.50	0.80	0.65
Visual→Photos	-	-	0.99	0.99
Opinion→Content	-	-	0.99	0.99
Opinion→Brand	-	-	0.99	0.99
Opinion→Other	-	-	0.99	0.99
Temporal→Activity	0.59	0.53	0.98	0.70
Temporal→Membership	0.95	1.00	0.99	0.98
Temporal→Seen	-	-	0.60	0.60
Geographical→Current	0.41	0.42	0.90	0.58
Geographical→Locations	0.40	0.82	0.95	0.72
Geographical→History	0.44	-	0.99	0.72
Degree→Subscribers	0.24	0.29	0.17	0.23
Degree→Subscribed	0.00*	0.07	0.17	0.08
Degree→Contributions	-	-	0.17	0.17
Degree→Visibility	-	-	0.17	0.17
Degree→Reputation	0.17	-	0.17	0.17
Degree→Trophies	-	-	0.60	0.60
Degree→Rank	-	-	0.46	0.46
Relationships→Interacted	>0.99	1.00	0.99	0.99
Relationships→FollowsUser	>0.99	1.00	0.99	0.99
Relationships→FollowedBy	>0.99	1.00	0.99	0.99
Relationships→Grouped	-	-	0.99	0.99
Relationships→FollowsBrand	-	-	0.99	0.99
Relationships→Contributes	-	-	0.99	0.99
Content→Text	>0.99	1.00	0.99	0.99
Content→Image	1.00	1.00	0.99	1.00
Content→Video	1.00	-	0.99	0.99
Content→Links	>0.99	1.00	0.99	0.99
Attributes→Temporal	0.88	-	0.99	0.94
Attributes→Impact	0.16	-	0.17	0.17
Attributes→Category	0.67	-	0.90	0.79
Attributes→Spatial	-	-	0.90	0.90

also falls into this bracket, with some notable low uniqueness of avatar images. Similarly, `Geographical→*` information produces low uniqueness, perhaps due to the similarity function embedding an overbroad scope of similarity relative to the typical locations of social network users.

Least unique of all were the `Degree→*` attributes and the `Attributes→Impact` attribute. These all use the same grounding similarity function, which sorts degrees such that they are assigned to a small number of ranked bins covering the spread of observed values, and as such it is unsurprising that there is a high chance of matching to randomly selected users in the empirical measures.

4.7 Interrelation of Metrics

This chapter has outlined a schema for online user profile information, and calculated estimates for the availability, consistency and uniqueness of profile attributes from that schema. This final section will explore the options for combining these measurements, and discuss how the resultant ranking of information attributes corresponds to how attributes are being used in the literature.

Firstly, it is important to note that the measurements in the above sections were unable to provide some consistency estimates for certain attributes, due to the impact of low availability of those attributes. Similarly, some values for uniqueness have been only theoretically estimated. As such, in what follows a curtailed selection of attributes will be considered, covering only those for which all three of availability, consistency and uniqueness are measured. Table 4.20 below lists these attributes and their values. The estimated availability of the removed items was an average of 0.25, compared to 0.49 for the remaining.

It is of interest whether attributes generally perform better across categories, or whether they are orthogonal dimensions. The correlations between the measures, given in Table 4.21, seem to indicate the latter conclusion. No notable correlation exists between an attribute's availability and its uniqueness or consistency. A middling-strength inverse correlation does appear to exist, however, between uniqueness and consistency.

Table 4.20 Attributes for which all measurements are available.

Attribute	Availability	Uniqueness	Consistency
Contact→Web	0.25	1.00	0.33
Bio→Username	0.98	0.64	0.80
Bio→Description	0.50	0.99	0.17
Bio→Age	0.29	0.80	0.74
Bio→Education	0.35	0.66	0.57
Bio→Occupation	0.32	0.65	0.58
Bio→Gender	0.51	0.00	0.98
Bio→Relationship	0.25	0.33	0.43
Visual→Avatar	0.76	0.58	0.46
Visual→Banner	0.29	0.65	0.44
Temporal→Activity	0.79	0.70	0.39
Temporal→Membership	0.75	0.98	0.20
Geographical→Current	0.51	0.58	0.82
Geographical→Locations	0.45	0.72	0.77
Geographical→History	0.12	0.72	0.99
Degree→Subscribers	0.67	0.23	0.74
Degree→Subscribed	0.58	0.08	0.61
Degree→Contributions	0.52	0.17	1.00
Relationships→Interacted	0.65	0.99	0.02
Relationships→FollowsUser	0.64	0.99	0.31
Relationships→FollowedBy	0.55	0.99	0.31
Content→Text	0.59	0.99	0.03
Content→Image	0.44	1.00	0.01
Content→Links	0.22	0.99	0.02
Attributes→Temporal	0.68	0.94	0.86
Attributes→Impact	0.65	0.17	0.83
Attributes→Category	0.24	0.79	1.00
Attributes→Spatial	0.10	0.90	0.67

As both measures make use of the same similarity function on an attribute-by-attribute basis, one can connect this correlation to the discriminative ability of these similarity functions.

This correlation is explored visually in Figure 4.4. Attributes clustering at high consistency and low uniqueness are from the `Degree→*` and `Attribute→Impact` grounding. The similarity function used for these is a six-bin system based on the observed ranges of possible values. This function would appear to be poorly discriminative, matching easily to non-matched and matched profiles alike. At the other end of the range, one sees attributes which rarely match to either matched or non-matched profiles – i.e. they pose too high a bar for similarity. These include exact matching methods for profile content.

Table 4.21 Correlations between estimates

	Availability	Uniqueness	Consistency
Availability	-	-0.13	-0.09
Uniqueness	-0.13	-	-0.56
Consistency	-0.09	-0.56	-

Notably, though, when content is matched, the similarity of attributes of that content is both unique and highly consistent.

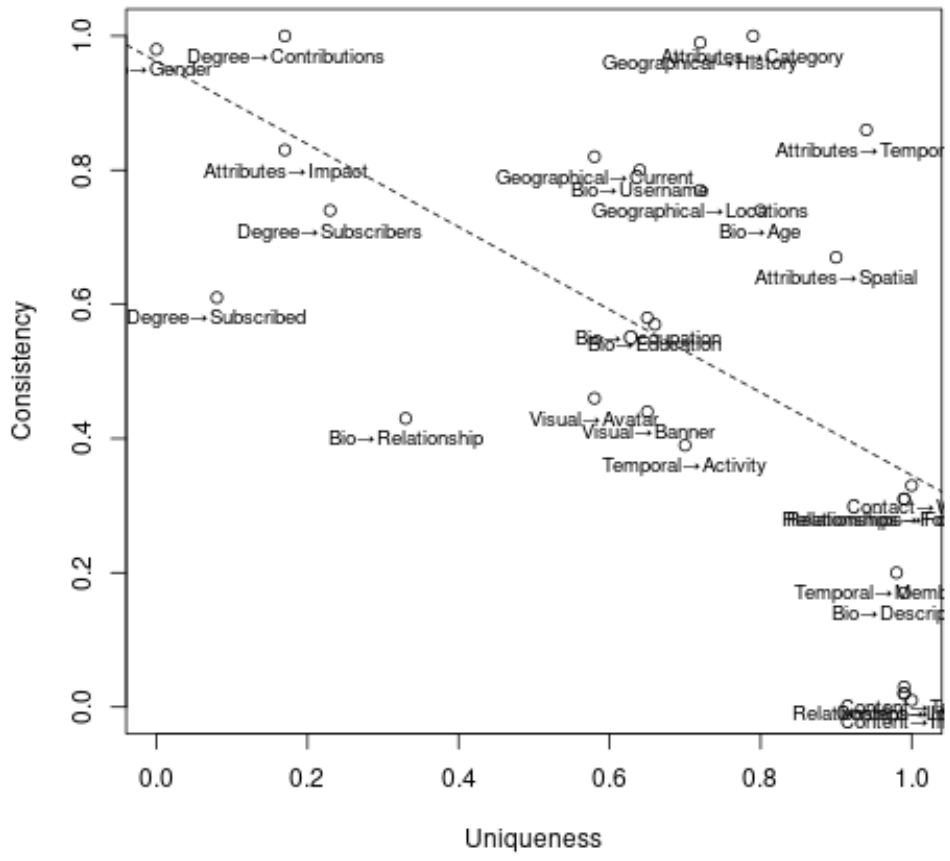


Fig. 4.4 Attribute consistency plotted against attribute uniqueness.

4.8 Summary

The ACID framework of Goga et al. [84] provides a basis for understanding the value of profile information, but parts of the framework are not necessary or difficult to measure. The ACU model is described, relating the properties of *availability*, *consistency* and

uniqueness to the overall identification value of a profile attribute, and to Bayesian and classical statistical approaches to identity resolution.

A survey of online user profiles is carried out in order to define an appropriate matching schema for online profiles from the top-ranked websites by traffic measurement rates. A schema of 35 profile attributes is defined as a reference point.

Using previous literature, and ground truth data sampled according to the method presented in Chapter 3, estimates of *availability*, *consistency* and *uniqueness* are calculated for these 35 profile attributes, and the relative rankings of attributes under each measure are discussed, along with potential implications and interrelations of measures.

Chapter 5

Feature Selection under Missing Data Conditions

The ACU model derived and grounded in Chapter 4 provides a means for understanding the quality of data available for identity resolution. In Eq. 4.16 a formula is presented for the conditional probability of matches based on similarity functions for particular fields, in terms of estimates of the availability, consistency, and uniqueness components of the model. This chapter covers how this model can be deployed within identity resolution to select features which are likely to be both identifying and available.

Identity-resolution attacks are a range of different methods for automatically linking online social network (OSN) profiles and other personal information datasets. Security and privacy researchers use identity-resolution attacks to understand the privacy impact of information-disclosure on the part of social networks, users or third-parties, helping calibrate risk in online social networks.

Some previous attacks in the security and privacy literature have focused on the identification value of individual features, including: usernames [109, 178]; network properties [128, 156]; user annotation information [103, 214]; group membership [235]; writing style [154, 222]; and time activity profiles [13]. Other attacks have taken joint approaches which merge multiple such features [23, 85, 139], using a selection of classifiers such as binomial linear regression, support vector machines (SVMs), decision-

tree classifiers and Naive Bayesian approaches, as well as more probabilistic methods drawn from statistical literature [50].

More recently, the reliability of these methods in practice has been called into question [84]. Problems arise from discrepancies between the filtered profile datasets used in initial evaluations and the sparse data revealed in profiles from more representative samples of OSNs. The most important issue after sampling bias is the handling of missing data. Improper treatment of missing data can transform an unbiased representation of social networking data into a highly skewed sample, analysis of which can give misleading results about the practical application and effectiveness of an identity-resolution attack.

Recall that in a typical identity-resolution situation, there are some identities represented by profiles $I = [i_1 \dots i_n]$ each of which contain a set of attributes $F = [f_1 \dots f_m]$. When two profiles are compared, the result is a comparison vector which relates how similar each of the attributes in the two profiles are, according to appropriate predefined similarity functions:

$$sim(i_1, i_2) = [sim(i_1 f_1, i_2 f_1) \dots sim(i_1 f_m, i_2 f_m)]$$

The case of interest is when the data for some field f_i is missing for i_1 or i_2 (or both), and as a result the similarity function for this field is also undefined. How should this missing data be handled? Take for example the first motivating scenario from Chapter 1, in which an officer is attempting to match online profiles of an individual of interest to an investigation. Inevitably, the known profile and many of the candidate profiles will be missing some of the attributes which the host OSN allows them to display. How the officer works with this partial data will be important to determining whether they manage to locate matched profiles for their target. In a one-to-many case, the missing data might entirely prevent certain comparison schemes, as the singular starting example does not contain the necessary values to employ them. Consider in comparison the second motivating example from Chapter 1, of a community administrator attempting to match profiles of known disruptive elements. Here it may be possible, if not desirable, to

use a broader range of comparisons and accept that there may be missing data on both the source and target end.

A common approach to handling missing data is *listwise deletion*, whereby any profile with any missing attribute would be removed from consideration. This approach appears to have been used in the identity-resolution literature several times, in different ways [128, 139, 154, 222]. While this approach can be appropriate for demonstrating performance of a technique with particular requirements, there are two major issues with applying this process to identity-resolution data. Firstly, most missing data in identity resolution is *missing not at random* (MNAR), as missing data comes from optional profile attributes which were either never provided by the user, or which are intentionally restricted from public access. Treating MNAR data as *missing at random* (MAR), or the even more unreasonable *missing completely at random* (MCAR), creates a bias in the resulting analysis. As listwise deletion is only valid for MAR data, this is a serious violation.

Secondly, as was explored in more detail in Section 4.4, most attributes used in identity resolution are missing for a significant proportion of real-world profile data. As such, listwise deletion combined over all attributes will result in deletion of a majority of the gathered data. The combined effect will be that a dataset gathered to be representative of a real-world identity-resolution problem is instead constrained to a small subset of the most willing-to-share individuals, whose revealed information can be expected to be significantly different from the population as a whole. A different approach known to statisticians, *pairwise deletion*, is equivalent to listwise deletion for most identity-resolution purposes, and assumes missing data is MCAR [88].

An alternative to deleting rows of data (removing profiles) would be to delete columns (remove attributes from profiles). This approach is not without merit in some cases, such as where a low-value attribute presents difficulty due to a high proportion of missing values. However, it does not solve the modal case, where a highly identifying attribute such as geographic location can be expected to be missing for a significant proportion ($>5\%$) of the dataset. Removing all attributes with a significant proportion of missing

data would limit classification to only a few mandatory attributes such as usernames, and prevent intelligent use of incidental profile data such as is most valuable in adversarial conditions.

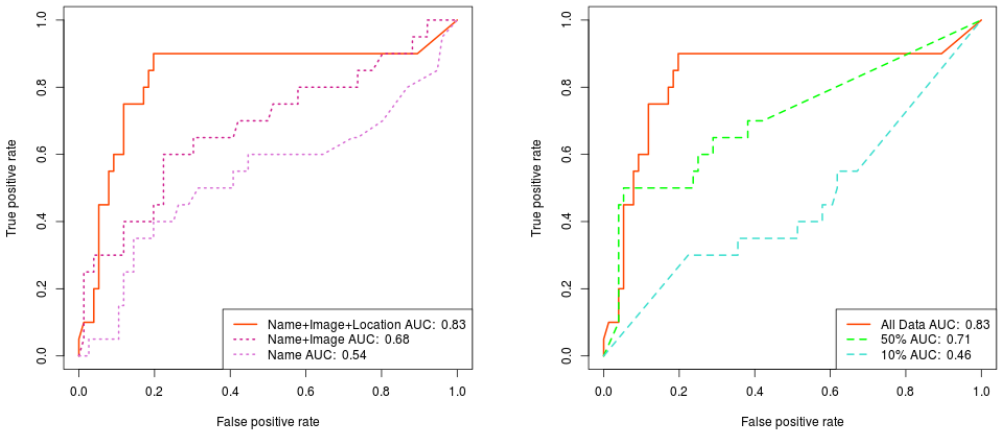
5.1 The Missing Data Problem in Identity Resolution

A dataset drawn from the method in Chapter 3 is used here to demonstrate the issues raised by missing data in identity resolution.

The entire dataset consists of 9,447 comparisons between potentially matching profiles, where potential matches are blocked based on the profile display name, and there are 187 correct matches (2%). A wide range of profile features can be extracted for these profiles, but the present discussion focuses on three features used in previous literature. Three standard similarity metrics are used to compare the username (Levenshtein distance), profile picture (Hamming distance of the perceptual hash) and location (geodesic distance between inferred coordinates) of profiles – a solid scheme drawn from a recent approach to identity resolution [85].

This dataset is a clear example of the explosive costs of deletion. For username alone, all comparisons are possible. Adding in the profile image feature, however, would reduce the dataset to 4,276 comparisons (45%) under listwise deletion, as profiles without images are removed from consideration. None of the correct matches are removed at this stage, so a false base-rate is already being established (from 2% to 4.4%). Adding in location similarity even more dramatically shrinks the dataset to just 96 comparisons (1% of the original data), with 20 positive examples (21% of the revised dataset) establishing an even more distorted base-rate.

The implications here for performance on test data are obvious: for large populations of real data, classifiers which have no means of coping with missing data can make no prediction. However, handling missing data also affects the training of models. Figure 5.1 demonstrates the impact that deletion schemes have on final classifier performance. Removing useful features or large proportions of training data from the cases where comparisons are complete causes underperformance.



(a) Removing informative features greatly harms classifier performance where those values are present.

(b) Removing training examples as in list-wise deletion also results in degraded classifier performance.

Fig. 5.1 ROC charts a tenfold cross-validation of a binomial regression model (a standard approach with well-understood performance) trained on the data subset for which all three features are available, under different deletion schemes.

5.1.1 Limitations of imputation

Where deletion is not appropriate, an alternative sometimes suggested is imputation of some standard value to replace the missing data [85]. However, the application of imputation for identity-resolution purposes is dubious, for a number of reasons:

1. At the level of individual profile content, imputation methods are not sufficient to produce viable text, image or complex network attributes, and are generally contra-indicated for even simple categorical variables due to the risk of violating hidden normality assumptions.

2. Assuming instead that similarity values between paired profile attributes are imputed, good imputation methods are not readily available.

    - Imputation with default scores like 0 or 1 can skew a classifier's valuation of that attribute.
    - Imputation with the mean value for the attribute can be similarly highly misleading for classifiers, as identity resolution usually suffers from highly imbalanced classes.

- Imputation with the mean value for the outcome class necessarily begs the question, leaking information from the outcome variable.
- Regression-based estimation, which predicts the missing value based on other attributes of the profile, could theoretically prove suitable, but is practically hampered by a nested variant of the original issue: the attributes being used to predict a missing value may themselves be missing in many cases. Other similarity values may not be strongly predictive of missing scores.

3. Most applications of imputation are suggested for cases where relatively small proportions of data ($<5\%$) are missing, but attributes in identity resolution commonly have data missing for $>50\%$ of cases, as is detailed in Section 4.4.

Even where possible, it should also be noted that imputation does not offer a great basis for comparison. Figure 5.2 demonstrates the significant gap in performance between a classifier exactly like those in Figure 5.1 when trained on actual values and when trained on imputed values. Different simple imputation approaches are trialled: replacing missing values with 0 or 1, and simple mean imputation (SMI), which replaces missing values with the mean of all observed values. Imputed performance is little better than chance, unlike that of the classifier trained on full data.

5.1.2 Mitigating low availability

As discussed above, traditional data deletion approaches are insufficient for identity resolution data due to a combination of imbalanced classes and content sensitivity. One approach which allows for these issues to be sidestepped is to employ a *Bayesian inference* approach to similarity scores, as opposed to a logistic regression, SVM or similar machine learning approach. These approaches are resilient to missing data by virtue of including each attribute similarity score for a profile as a distinct piece of evidence. Each piece of evidence is then used to update a prior probability, rather than as dimensions or components of an expression. Missing data items can thus be omitted,

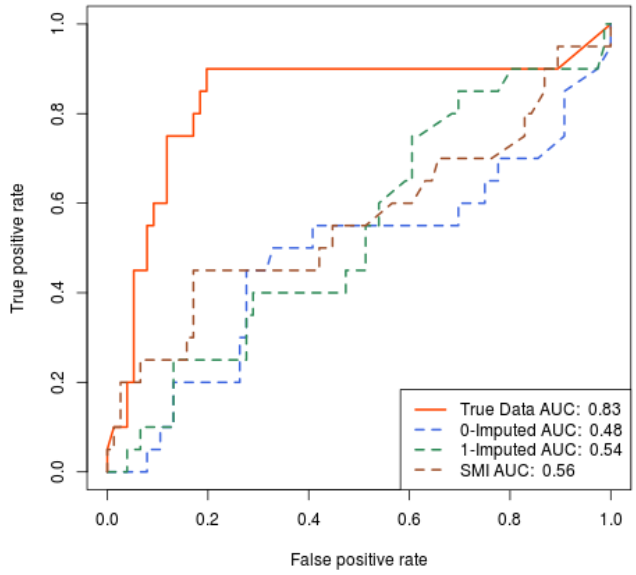
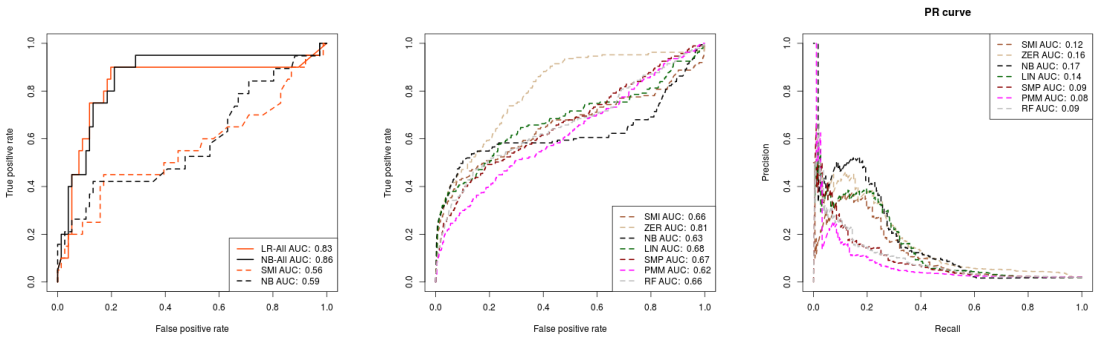


Fig. 5.2 Binomial logistic regression performance under different imputation schemes, applied to the same proportion of known data as was observed missing in the larger sample, introduced as MAR. Each of these approaches produces significantly poorer models, demonstrating the gap between imputed and real data.

necessarily reducing the evidence base for a prediction, but still producing a grounded probability.

An alternative approach is to use some form of imputation to complete missing data, excluding the outcome variable from imputation calculations to prevent data leakage (which could easily render a classifier more effective than on real complete data). There are a number of viable methods for imputing missing data. The forms of imputation trialled here are summarised in Table 5.1.

To examine how well these imputation methods perform, data was removed at random from our comparison vector where a true value was known, imputation was performed on this missing data, and then the imputed values were compared to the true values. Table 5.2 shows the mean distance between imputed values and true similarity scores for the location and image similarity data, where proportions removed reflect the real rates of loss observed in our comparison vector. The differences observed in this small dataset would suggest that linear regression of missing values is the most accurate scheme, and 0-imputation the least successful, but it is important to note that `LIN`, `PMM` and `RF` were



(a) Comparison of methods on complete data (1% of total), with proportionate MAR data introduced.

(b) Missing data methods applied to the full dataset, which contains data which should be MNAR.

(c) Missing data methods applied to the full dataset, measuring precision and recall to account for class imbalance.

Fig. 5.3 Comparison of classification performance under different missing data mitigation strategies.

Label	Method
`ZER`	Assume missing similarities are 0.
`SMI`	Impute missing data as column means from training data.
`NB`	Naive Bayesian handling of training and test data, no imputation.
`LIN`	Impute missing data as result of chained linear regression on other columns.
`SMP`	Imputation via random sampling from observed data.
`PMM`	Imputation via predictive mean matching (select real values closest to those predicted by linear regression).
`RF`	Imputation of values by a Random Forests classifier.

Table 5.1 Trialled data-robustness methods.

unable to make imputation on the extremely sparse location similarity data. Imputation with the observed mean was the best performer given the need to always produce a value.

Figure 5.3a shows the results of applying the Naive Bayes approach to complete data with missing data introduced at random. Naive Bayes performs comparably to a regression model on the complete data and matches performance with an imputation-grounded regression on the constructed missing data[1].

Figure 5.3b would appear to suggest that most imputation methods, including simple mean imputation and imputation of values suggesting the majority class (0-imputation) outperform a Naive Bayesian approach when applied to a larger MNAR dataset. However,

[1]More advanced imputation methods failed to fit to cross-validation sets on this small subset of data

Method	Image δ	Location δ
SMI	0.116	0.335
ZER	0.303	0.616
SMP	0.108	0.370
PMM	0.086	-
LIN	0.060	-
RF	0.081	-

Table 5.2 Mean absolute error rates for imputation schemes

ROC plots can be misleading about performance when classes are heavily imbalanced, as can be seen by comparison with Figure 5.3c, wherein the NB approach has a greater precision profile than any of the regression-on-imputed-data approaches.

The failure of advanced imputation strategies to outperform 0-imputation on this data is surprising, given the previous indications in Table 5.2 that this method least well reflects the true similarity values. However, this can be explained with reference to (a) the large proportions of missing data which leave advanced inference strategies with insufficient information for accurate predictions and (b) the class-imbalance inherent to the data, and the bias of available data toward positive-class examples. Imputation models fit to the available data are likely to overestimate the values of missing similarity measures because the available data is skewed towards positive classes.

These results suggest that choosing a missing-data-robust classifier such as Naive Bayes is generally more advisable than imputing missing data and using another classifier. Where data is available, classifiers are comparable, but Naive Bayesian classifiers degrade more gracefully than the imputation methods required for other classification approaches in the face of missing data. Where imputation needs to be carried out, care must be taken to avoid biasing imputed data due to MNAR effects.

For further improvements of the handling of missing data, it becomes more important to model the patterns of data availability, and the underlying causes. It is here that the measurements in Section 4.4 become instrumental.

5.2 Availability-Sensitive Feature Selection

The previous section addressed how to mitigate the issue of missing data where it appears. This section, in contrast, focuses on how estimates of availability, alone or in combination with other information, can be used to select attributes which are as available as possible, or else of such identification value that their low availability needs to be ignored. In a one-to-many identity-resolution problem, the feature selection is constrained by the binary availability of features in the source profile, but selecting the best features for comparing with candidate profiles may still help deal with missing data arising amongst candidates.

The estimates given in Section 4.4 for the availability of particular attributes are particularly useful in attribute selection for an appropriate matching schema. This can be approached in two ways: in an *a-priori* manner, using only availability information, and in an *a-posteriori* fashion, as part of a supervised feature selection process on labelled data.

5.2.1 A-Priori model

An attacker building identity-resolution solutions for general-purpose application can constrain her selection to those attributes which have high *support*. These are the attributes which are most likely to exist across different profile networks, and so the inclusion of other attributes will at best limit a method's applicability to a more specific subset of online profile networks. Focus on use of well-supported attributes makes a method generalisable.

Secondly, high-support attributes can be filtered to exclude low or medium completeness attributes, thus reducing missing data to a smaller number of cases, which may even successfully be imputed based on regression from other profile attributes. Table 5.3 takes this approach to refine Table 4.14, and lists the top attributes by these availability-only criteria, ranked by their completeness.

These figures are based on a combination of prior literature and our own measurements of a dataset, as given in Section 4.4. An attacker can adopt the same approach by

Completeness	Structure	Availability	Attribute
1.00	0.95	0.98	Bio→Username
0.80	0.72	0.76	Visual→Avatar
0.74	0.60	0.67	Degree→Subscribers
0.64	0.94	0.79	Temporal→Activity
0.64	0.63	0.64	Relationships→FollowsUser

Table 5.3 Most available attributes under an *a-priori* approach.

averaging the values given in Table 4.14 with her own domain estimates of completeness and support and selecting features through this process so that their attacks generalise well in the face of missing data.

5.2.2 A-Posteriori model

The *a-priori* approach may be most useful in domains where labelled data is not available. Where labelled data is available, it can be used to attempt to balance the demands of availability and identification value. The derivation given by Eq. 4.16 combines these concerns, weighting the expected identification value of an attribute in accordance with its expected availability.

This then provides an implicit attack model, whereby an attacker attempts to maximise their combined E or expected identification value, by selecting features which maximise Eq. 4.16 to use in an identification attack on social networking data.

Estimates of $C(f_i)$ and $U(f_i)$ for our dataset were used in combination with the $A(f_i)$ estimates in Table 4.14 to create an *a-posteriori* availability-adjusted ranking of the supported profile attributes, presented in Table 5.4.

These values, which take into account the identification value of attributes as well as their availability, paint a somewhat different picture of the desirability of various profile attributes to that given in Table 5.3.

The highest weight is given to `Contact→Web` information, which is very consistent between matched profiles and moderately available and identifiable.

Attribute	*A*	*U*	*C*	*I*
Contact→Web	0.25	1.00	0.33	6.37
Relationships→FollowsUser	0.64	0.99	0.31	4.17
Relationships→FollowedBy	0.55	0.99	0.31	3.95
Attributes→Temporal	0.68	0.94	0.86	3.26
Bio→Description	0.50	0.99	0.17	2.95
Temporal→Membership	0.75	0.98	0.20	2.84
Content→Image	0.44	1.00	0.01	2.14
Bio→Username	0.98	0.64	0.80	1.12
Content→Text	0.59	0.99	0.03	0.69
Geographical→Locations	0.45	0.72	0.77	0.30
Relationships→Interacted	0.65	0.99	0.02	0.24
Attributes→Category	0.24	0.79	1.00	0.19
Bio→Age	0.29	0.80	0.74	0.09
Temporal→Activity	0.79	0.70	0.39	0.03
Geographical→Current	0.51	0.58	0.82	-0.01
Visual→Avatar	0.76	0.58	0.46	-0.27
Attributes→Spatial	0.10	0.90	0.67	-0.59
Attributes→Impact	0.65	0.17	0.83	-0.62
Degree→Subscribers	0.67	0.23	0.74	-0.64
Degree→Contributions	0.52	0.17	1.00	-0.68
Bio→Education	0.35	0.66	0.57	-0.77
Bio→Occupation	0.32	0.65	0.58	-0.92
Bio→Gender	0.51	0.00	0.98	-1.00
Geographical→History	0.12	0.72	0.99	-1.24
Content→Links	0.22	0.99	0.02	-1.32
Degree→Subscribed	0.58	0.08	0.61	-1.38
Visual→Banner	0.29	0.65	0.44	-1.46
Bio→Relationship	0.25	0.33	0.43	-2.64

Table 5.4 Ranked availability-adjusted estimates of the identification value of profile attributes for our dataset, where supported.

Network relationships, being highly unique and moderately available and consistent, are promoted as the next most desirable, a result which is aligned with successful large-scale de-anonymisation methods from the literature [154].

Timestamp information, an often-overlooked feature, performed well across each measure. The strongest signal is given when matched content has the same timestamp, but user membership dates for social networks are often highly consistent, and the information is readily available.

In this dataset, usernames are weighted lower than may be expected from previous work due to particulars of the sampling approach used (see Chapter 3).

Similarly, despite ranking highly on availability, profile pictures are considered less useful features overall due to mediocre consistency and uniqueness.

Other good performers were matched text and images from profiles, which balance a low general consistency with high uniqueness and moderate availability. The availability and uniqueness of forms of user text may be considered the driving forces behind authorship attribution methods, which aim at improving consistency over that of the simple measure (function-word distance) used in the reported estimate.

Geographic features showed high consistency across profiles, but were not highly available and of poor uniqueness, degrading their practical value. Gender information was predictably consistent, but just as predictably of low identifying value.

Content attributes as a class were highly consistent, though it should be noted that measurement was only made between the attributes of known content matches. The usefulness of attributes such as categorical matches as they are considered on top of (for example) strong textual matches between status updates might be minor.

These rankings suggest a direction for identity-resolution methods which must balance the needs of general application (high availability) with performance (high identifiability) in terms of features and combinations of features which should be investigated.

5.2.3 Performance impact

Figure 5.4 compares the performance of a Naive Bayes classifier using the literature-based feature set used in Section 5.1.2 (labelled as `REF`) with that of other Naive Bayes classifiers using features drawn from the recommendations of the *a-priori* and *a-posteriori* models (labelled `PRI-`N and `POS-`N) respectively, where N is the number of features used counting from the top of the recommended list.

The comparison shows that the *a-priori* features match performance with those recommended from the literature (as expected, two of the features being the same), while the *a-posteriori* features, selected based on expected identification value as well as availability, perform much better, with an AUPRC difference of 0.1 for the same number of features and 0.22 for increased feature counts.

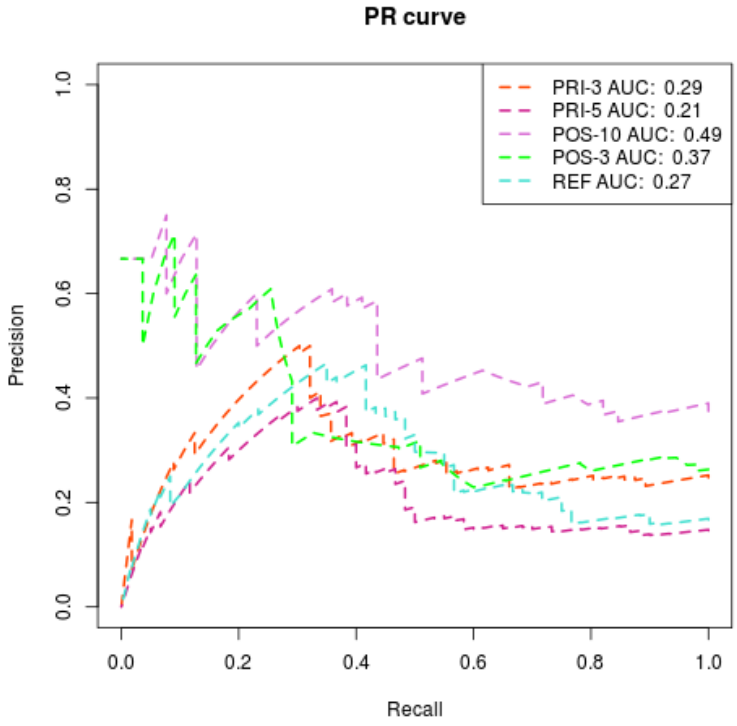


Fig. 5.4 Comparison of performance under feature selection.

Features	% data	% cases
REF	37.4	87.7
POS-3	58.2	80.5
PRI-3	8.2	12.3

Table 5.5 Rates of missing data and proportion of affected comparison cases under different feature selection schemes (3 features each).

The value of the *a-priori* feature selection is demonstrated in Table 5.5, which compares the proportions of missing data. The *a-priori* feature selection scheme results in signficantly fewer missing data items, whether examined on the basis of comparisons affected or overall missing value proportions. This means that the *a-priori* model's selection is far more workable for classifiers which are not robust to missing data.

The *a-posteriori* system manages to reduce the number of affected cases slightly in comparison to the reference scheme, but does this using an overall greater proportion of missing data values, suggesting that missing-data-robust classifiers such as Naive Bayes should still be preferred under this model.

5.3 Summary

This chapter has shown a practical application of the ACU model, and of *domain-general* estimates of measures for all attributes. The feature selection approach demonstrated here (in conjunction with existing mitigation efforts) is neither the limit of the application of the ACU model, nor a complete solution to the problem of missing data, but it does advance a means for balancing availability and identification in mitigating missing data issues. The next chapter will move on to discuss extensions of the ACU model, and how they might be applied in a real application area case study.

Chapter 6

Data Quality Measures for Applying Identity Resolution

The ACU model presented in Chapter 4 reflects data quality measurements which are central to the understanding of identity resolution systems' performance. The measurements of availability, consistency and uniqueness provide a means for understanding the value of profile attributes for identity resolution within a domain. As Chapter 5 shows, this allows for the design of more effective identity-resolution systems within said domain.

However, identity resolution is not a task to be understood as a goal in and of itself. In a police investigation, the goal is finding evidence of involvement in a crime, in a study of meme-sharing the goal is identification of memes and a user's cross-posting behaviour. What value an information item has in such an application domain depends on the investigator's end-goals, and there are many types of investigation which could benefit from linking personal profiles. Different attributes will gain greater weight in such investigations for reasons beyond their strictly identifying properties. Data quality measurements which relate to these applications are useful in identity-resolution work even if they do not contribute to understanding the core process of identity resolution.

This chapter presents and discusses three possible attribute quality measurements which are useful for understanding the value of attributes within a specific application of identity resolution: The *novelty* of information against an existing collection of attributes

about a subject, the *relevance* of information to the investigative context, and the *veracity* of profile attributes to the real attributes of the profile subject. These properties each come from disciplines closely aligned with identity resolution, and speak to different dimensions of application. In brief, these measures assume that most purposes of identity resolution reduce to the search for at least one of *new*, *useful* and/or *true* information about a profile subject. This chapter thus seeks to understand how one might judge these properties of profile attributes for a given task.

These task-specific data quality measures form a complement to the ACU model detailed in Chapter 4. Where the ACU model breaks down the important considerations of data quality for the purposes of identity resolution, the *novelty*, *relevance* and *veracity* measures described in this chapter provide a means to value information that may be obtained through, rather than *for* the purposes of, identity resolution. By better understanding which profile attributes are valuable to such a system, and why, insights can be drawn for the development of systems in the application domain.

6.1 Novelty

In a majority of identity-resolution situations, the intended purpose of resolving two identities is to make available new information about the data subject. For example, a police investigator may wish to link a suspect's social media identity to an identity they hold on an underground forum, to learn about criminal activity not recorded in one account which can be traced to a real name and image held on another. Alternatively, a people-search service may wish to learn new contact details for a subject, or an advertising agency may wish to learn more about a subject's interests and preferences, or a company may want to learn about their employees' social media behaviour in order to protect themselves from possible social engineering attacks. In each case, some profile of the user is being augmented by identity resolution.

One form of novelty measurement has already been presented in this thesis. In Chapter 3, the evaluation of the sampling method included the application of the Kullback-Leibler divergence, also known as *information gain* between distributions. By finding the

information gain between a relevant distribution on two profiles connected by identity resolution, a post-hoc value can be gathered for the accrued novelty of the information.

This definition is useful for certain applications – such as identifying when the level of novelty is decreasing over iterations when performing iterative identity resolution in a potentially unbounded database space like the internet (such as in the system proposed by Bennacer et al. [26]) – but does not allow for predictive statements to be made about the value of profile attributes. For such an understanding, novelty can be viewed as being in part a particular conditional case of availability.

For all databases in which the subject is present, one wishes to know the probability that new information is available, given current observations of the subject's records from prior databases. One means of understanding this possibility is through the *structural support* for possible attributes within a candidate database. Databases with high structural support for attributes not currently known are likely to contain novel types of information. Complementarily, the possibility of new values for already-observed attributes (such as new text posts, new location updates) should be anticipated with reference to the *completeness* of fields within a database.

In the case of unobserved attributes, the potential for finding new attributes in some possible datasets is the average *completeness* of all currently unobserved attributes from the schema, amongst the unexamined databases which have structural support for that attribute.

In the case of new values for an attribute, the above still applies, but must also be moderated by the conditional probability of new values. Specific conditional probabilities could be calculated for the given set of observations and each target database, in an environment where these probabilities can be learned, or the inverse of the attribute's *consistency* can be used to approximate this probability – being already conditional on availability. Thus novelty for an attribute f_i could be formalised as

$$N(f_i) = A(f_i)\dot{(}1 - C(f_i)) \tag{6.1}$$

6.2 Veracity

Information about the truthfulness of information on social network profiles (i.e. whether they reflect the nature of the real person making the profile) is relevant to many end-tasks. Consider the previous example of the company looking for social engineering pathways: information they obtain through linking two online accounts would only be useful to the imagined attacker if it is true information about the holder of the account or their employer. Novel location information is useful in establishing the potential whereabouts of the account holder's employer, but only so long as that location information is a reliable indicator of the account-holder's actual location.

Veracity can be expected to be both correlated and causally related to *consistency*, but it is important to understand that this is a sufficient but not necessary condition – a person's online accounts may be highly consistent even for attributes which do not relate to their offline existence. For example, friends lists online could be entirely populated with online-only contacts, with no representation from the people the account holder interacts with offline.

Coming to an estimated weight for the veracity of profile attributes is an uncertain process. The work of Rowe & Ciravegna [189, 190] proceeds by comparing the values declared in online profiles to those elicited through interview. Their discoveries include that, for example, the majority of each participant's real-world social network is represented in certain of their online social networks. This is an interesting general phenomenon, but for security purposes it may not be sufficient – it is of course possible for certain information to be provided falsely, or even copied from another person's account to imitate them. One solution might be to rely upon explicit social verification, as implemented by Bahri et al. [16] in their COIP system, whereby profile attributes can be assessed for verification as part of a light workload spread across the social network around a profile. Goga et al. [84] attempted to cover this through their term for "non-Impersonability", though they did not produce a means for estimating its value for an attribute. The insight is nonetheless useful – veracity in a security context should focus on the locus of control for an attribute.

In the case of many values, such as profile image or username, the control over the veracity lies entirely with the profile author. In situations where the author must be considered adversarial, the weight placed on general levels of veracity in these attributes must be lower. Some values are recorded by the social networking platform to reflect the profile owner's activity, such as the timestamping of status updates. While there are methods to manipulate the interpretation of these values (e.g., using an automated method to post a status update as an alibi whilst otherwise engaged), there is greater effort and less payoff from the action, as the value can only be manipulated in advance, and not at will. Finally, other values require the collaboration of others to achieve – the networks of friends and followers attached to a profile are beyond the ability of a single profile author to control, as are impact ratings on content they post, and other measures of their degree within a platform. These values could still be manipulated, but the cost of doing so is significantly higher.

On this basis, a three-level categorisation can be applied to profile attributes: **a) editable by owner**, **b) produced by owner** and **c) produced by network**. Within these categories, it may also be possible to use empirical measurements of the general population to rank attributes by their expected veracity, relying on comparisons such as that of Rowe & Carmagnola.

6.3 Relevance

A final important quality measure comes from considering identity resolution under the domain of information retrieval. *Relevance* in information retrieval settings refers to the degree to which gathered information meets a user's needs [147]. Continuing the example of the company investigating their employees' exposure: when looking at which information would be most valuable to recover, it is not sufficient to find only new (high *novelty*) and true (high *veracity*) information, if said information has no bearing on the employee's work behaviour or susceptibility to attack.

The relevance of any particular profile element is necessarily defined by the application to which an identity-resolution system is being applied, and as such an application-

independent estimation method is not possible. However, once profile attributes in a domain are assigned relevance weightings, this value can be used alongside the expectation of *novelty* to select optimal datasets to query in an iterative identity search system.

6.4 Example Application: A Social Engineering Vulnerability Detection System

Social engineering attacks pose a major risk to the security of organisations. Attackers use a range of tactics to get targeted employees to disclose sensitive information or enable compromise of business systems. Basic approaches may be as simple as a phishing email targeted at all staff members, but current research suggests that the effectiveness of such attacks is greatly increased through the use of open source intelligence (OSINT) which refines the target group and personalises the payload [96, 111].

The information to enable such customised attacks is now widely available, with individuals using social networks to release volumes of personal information to the general public. Even more worryingly, organisations themselves are engaging with social media and publishing employee rosters, activities which allow attackers to easily identify their data-harvesting targets amongst the millions of social media users.

Existing research has demonstrated the social engineering risks posed by such OSINT data [20]. However, this normally relies on labour intensive manual identity resolution [55], which is impractical to replicate in defensive manoeuvres. Other techniques rely on "active" engagement with potential targets to elicit information — through zombie profiles or misleading friend requests [194] — and hence attackers would risk detection prior to an attack being launched.

The authors of [67], including the author of this thesis, demonstrate that is it possible to use identity-resolution techniques to automatically identify the employees of an organisation amongst individuals within its online footprint. The author of this thesis developed the identity resolution and employee detection segments of that publication,

demonstrating, firstly, that it is possible to automatically resolve employee identities across multiple online social networks, for large-scale harvesting of information pertinent to launching social engineering attacks. Secondly, that such harvesting can be undertaken "passively" without resorting to invasive measures, enabling vulnerability assessments which do not rely on exercising deception during social engineering penetration tests. And finally that through automated identification of OSINT that may be used to conduct or enhance a social engineering attack against an organisation, it is possible to highlight potential risks to the target, allowing appropriate mitigation techniques to be selected.

Most relevantly for this thesis, the authors of [67] sought to identify the most valuable online profile attributes for building social engineering attacks. Interview participants from accredited penetration testing companies were solicited for research into the use of open-source intelligence in social engineering engagements[1].

Six professionally qualified penetration testing experts, with knowledge and experience in social engineering engagements, were questioned on the importance of OSINT data items for each attack vector. This included identifying the essential data needed to bootstrap an attack or payload, and the non-essential OSINT data which can still contribute to the effectiveness of an attack. The goal of the interviews was to determine the attack vectors and OSINT data that are used in the real world, and filter out the outliers that are rarely deployed effectively or are embellished in literature as to their effectiveness.

Participants were first asked to discuss their experience of social engineering attack methodologies. At this stage of the study, the aim was to determine the real-world attack vectors used by social engineers, and understand their practicality for deployment. This allowed identification of which techniques were used more often, and which would be preferred. The social engineers consistently identified the following as the attack vectors used in real-world engagements:

[1]These interviews were designed and conducted, and the initial results organised, by co-authors Robert Larson and Benjamin Green in the original paper. The following discussion details their methodology as background to the information items given in Table 6.1, but is not the work of the author of this thesis.

Table 6.1 Level of contribution of OSINT data to attack impact. B = Required to bootstrap an attack; A = accentuates an attack.

Item	**Email**	**Phone**	**Onsite**
Name	B		
Name of person with job title	B	B	B
Identity and position in company structure	B	B	B
Name linked to employment by company	B		
Name of new employees	A	A	
Format of email (e.g. ital.lastname@company.com)	B		
Specific Email address	B		
Group / generic Telephone number		B	
Direct phone number with name		B	
Cell phone number		A	
Social media posts	A	A	
Social media connections / friends	A		
Social media photos	A		A
Social media hobbies / sports / groups	A		
Email footer / communications sample	A		
Company supplier / partner information	A	A	B
Employee availability / daily routine	A		B
Absence indicators (e.g. out of office reply, Facebook)			B
Files shared on corporate website (PDF, XLS, DB etc)	B		
Identity of facilities manager			B

- **Email**: phishing / spear-phishing emails that were used to manipulate a target into visiting a malicious website, or opening a malicious file.
- **Telephone**: voice phishing or 'vishing', used to extract information directly or persuade a target into interacting with a previously delivered payload.
- **Physical**: gaining physical access to an organisation's site or systems, through use of a deceptive pretext, or delivery of physical media (e.g., drop of a USB stick).

In addition to these attack vectors, the six experts were questioned about the use of online attacks, such as strategic compromise of websites used by targets, and the use of

social network sites as an attack vector. Such attack vectors were considered by most to be out of bounds in a contract penetration test, due to reliance on services external to the customer, risk of collateral damage, and invasion of employee privacy. It was noted by the experts that such concerns were not considered by criminals.

Experts were asked to evaluate each individual attack method against criteria of:

- **Frequency of use**: rate of use in real-world engagements.
- **Effectiveness**: rate of success and detection.
- **Efficiency**: time requirement and level of automation.

Responses were largely consistent in the frequency of use and effectiveness of attack methods used, in terms of rates of success and detection of these attacks. However, it was clear from discussions that success was often interpreted as an overall objective of a penetration test, rather than an individual attack; e.g., from 100 phishing emails sent, 10 may be opened, but 1 may result in a successful compromise of the organisation.

For email-based attacks, all interviewees stated that these were frequently used (more so than any other attack vector), and all but one claimed the method to be successful in the majority of cases, with low detection rates. The level of automation and time frame varied, ranging from almost completely manual to almost fully automated, and from a few hours in one afternoon, to waiting weeks for a response. This reflects the wide range of engagements social engineering penetration testers are involved in.

For telephone-based attacks, 3 interviewees often employed this as an attack vector, 2 did around half of the time, and 1 not all. It was agreed that this was a successful method the majority of the time, with at least one set of credentials (or some other target information) gained in most engagements. Detection rates reported varied dramatically, again depending on the exact nature of the attack and the information sought. This was always done entirely manually, with each call normally lasting 10-15 minutes (except for one interviewee using much shorter phone calls of less than a minute).

Physical access attacks were part of less engagements according to our experts, but still used quite commonly (over 80% of engagements) in most cases (4/6). Success rates

ranged from 50% to above 90%, with detection rates reported as being low (except for one report of USB key drops). This was always done entirely manually, with engagements taking at least a day, and sometimes up to a week.

The main focus of the interviews was the use of OSINT data for the attacks discussed. Following discussion of each attack vector, experts were asked to detail OSINT items that facilitate it, highlighting whether they are essential to the attack process (i.e. an attack cannot occur without this OSINT item), or non-essential. For non-essential items, experts were asked to discuss the degree to which each item contributed to, or accentuated, the success of an attack. Where possible, experts were asked to rate their perceived importance of non-essential items, so as to provide a point of reference relative to the contribution of other pieces of OSINT data.

In addition to the perceived importance of each item, experts were asked to discuss the process of obtaining the OSINT data items, focusing on time required and level of automation of the process. In this manner we gained an understanding of the resources required to extract each OSINT data item. To understand the rank of importance of OSINT data items, perceived importance to the attack process was compared to the resources required to extract the data, in terms of time and level of automation. In this manner, OSINT data that is easy to obtain (i.e. fast and automated) was ranked more highly, than on requiring increased resources to extract. Furthermore, we are able to identify the level at which OSINT data contributes to individual or multiple attack vectors; flagging those items which bootstrap multiple attack vectors as more useful to an attacker.

The various OSINT items identified by the experts and their nature (bootstrap or accentuator) are shown in Table 6.1[2]. Bootstrap (B) items are shown in red, whilst Accentuators (A) are shown in yellow.

[2]Replicated from [67]

6.5 Case Study: Improving Vulnerability Detection

The above highlights the information which can be judged important for social engineering vulnerability. To move from this to an automated vulnerability scanner, the authors of [67] described two components – a classifier to identify social media accounts which belong to employees, and an identity-resolution classifier to gather more information on each target by iteratively selecting target networks – of LinkedIn, Twitter, Facebook and Google+ – in which to attempt identity resolution with the current target profile.

Our discussion in Sections 6.1, 6.2 and 6.3 provide a basis for understanding the value of information in this application. The *domain* of interest being online social networks, the schema presented in Chapter 4 can be used as a guide, and the general identification values of attributes can be understood as identical to that presented in Table 5.4. The sections below detail how application-specific measures of *Relevance*, *Veracity* and *Novelty* can be constructed and integrated into an automated vulnerability scanner.

6.5.1 Relevance

The bootstrap and accentuator information detailed in Table 6.1 provides a good basis for understanding which data can be considered relevant to the task of social engineering vulnerability detection. However, this information does not all map directly to the domain schema. In Table 6.2 the portions of the schema which appear to map to the relevant social engineering information are presented, along with a relevance score based on the number of *bootstrap* items ($\times 0.2$) and the number of *accentuator* items ($\times 0.1$) which each attribute might provide.

This mapping reveals some things which are already intuitive, such as that the name and occupation are the most relevant portions of an online profile – revealing the identity and employment status of the target – but also some less immediately clear takeaways, such as that timestamping information associated with the account can be very valuable to social engineers.

Table 6.2 Relevance ranking of profile attributes

Attribute	*Relevance*
Bio→Username	0.9
Bio→Occupation	0.6
Contact→Email	0.5
Contact→Phone	0.5
Temporal→Activity	0.5
Bio→Description	0.4
Geographical→History	0.3
Temporal→Seen	0.2
Geographical→Current	0.2
Relationships→FollowsBrand	0.2
Relationships→Contributes	0.2
Content→Text	0.2
Attributes→Temporal	0.2
Attributes→Spatial	0.2
Bio→Education	0.1
Bio→Relationship	0.1
Bio→Habits	0.1
Visual→Avatar	0.1
Visual→Photos	0.1
Opinion→Brand	0.1
Relationships→Interacted	0.1
Relationships→FollowsUser	0.1
Relationships→FollowedBy	0.1
Relationships→Grouped	0.1
Content→Image	0.1
Content→Video	0.1

The high performance of this, and `Bio→Description` is notable, as these attributes are also shown to have comparatively high general identification value under the ACU model. Attributes with high identification value and high relevance will be of particular interest to an application domain.

6.5.2 Veracity

Taking the three-level categorisation from Section 6.2, each of the profile elements considered relevant from Table 6.2 are assigned a rank weight for veracity, with the mapping (simply preserving ranks as equidistant) as follows:

- produced and editable by owner – 0.3

- produced but not editable by owner – 0.6
- produced by network – 0.9

Table 6.3 Veracity ranking of relevant profile attributes

Attribute	*Veracity*
Relationships→FollowedBy	0.9
Temporal→Activity	0.6
Geographical→History	0.6
Temporal→Seen	0.6
Attributes→Temporal	0.6
Attributes→Spatial	0.6
Relationships→Interacted	0.6
Relationships→FollowsUser	0.6
Bio→Username	0.3
Bio→Occupation	0.3
Contact→Email	0.3
Contact→Phone	0.3
Bio→Description	0.3
Geographical→Current	0.3
Relationships→FollowsBrand	0.3
Relationships→Contributes	0.3
Content→Text	0.3
Bio→Education	0.3
Bio→Relationship	0.3
Bio→Habits	0.3
Visual→Avatar	0.3
Visual→Photos	0.3
Opinion→Brand	0.3
Relationships→Grouped	0.3
Content→Image	0.3
Content→Video	0.3

In Table 6.3 the values are given for the general status of these profile attributes across the LinkedIn, Facebook, Twitter and Google+ networks which were used to inform other portions of this thesis, as well as within [67]. While this scope is somewhat limited, it has general applicability due to these being some of the largest online social networking platforms.

The least controlled element amongst those given nonzero relevance was `Relationships-→FollowedBy`, which is an artefact of the user's social network presence. It is of course possible that this attribute could be influenced by the user, by indirect methods or different forms of impersonation, but generally speaking the problem of *causing a specific*

user to follow your profile is far less tractable than superficially related problems such as increasing follower counts (which `Degree→*` information was not judged relevant).

Temporal activity is again a strong contender, being generally only modifiable in advance and thus requiring premeditation for any attempts at disguise. Geographical history also falls under this umbrella, being partially temporal in nature. Other network activity is similarly hard to edit retroactively, particularly simple public interactions which are likely to be recorded on both the originating and receiving profile.

The majority of relevant profile attributes, however, were produced and editable by the owner of the profile, and thus have low veracity. This includes many otherwise relevant attributes such as *Bio*Description and *Visual*Avatar, all of which should be trusted little under this approach.

6.5.3 Novelty

As discussed in Section 6.1, novelty estimates will change based upon the information available at a particular stage in an iterative identity resolution operation. However, for a general a-priori estimate of the likely novelty of an attribute, a linear combination of *availability* and the inverse of *consistency* can be used. Table 6.4 shows these (normalised) values for the relevant attributes.

These novelty estimates reflect the general possibility of finding new values for these attributes. The most plausibly novel attribute is `Relationships→Interacted`, indicating that there is a high probability of finding new associates of a target profile. Two `Content→*` items are the next-highest in novelty, suggesting new text and image updates from a person are likely possible through identity resolution, followed by self-descriptive text, other relationship and images, other names and then temporal indicators. Note the high availability and consistency of time updates means that *new* observations of time activity are probably correlated with existing observations, so there is only a mediocre value to collecting more of them.

Table 6.4 Novelty ranking of profile attributes

Attribute	*Availability*	*1 - Consistency*	*Novelty*
Relationships→Interacted	0.65	0.98	0.82
Content→Text	0.59	0.97	0.78
Content→Image	0.44	0.99	0.72
Bio→Description	0.50	0.83	0.67
Relationships→FollowsUser	0.62	0.69	0.66
Visual→Avatar	0.75	0.54	0.65
Relationships→FollowedBy	0.54	0.69	0.62
Bio→Username	0.98	0.20	0.59
Temporal→Activity	0.79	0.30	0.55
Temporal→Seen	0.55	-	0.55
Attributes→Temporal	0.68	0.14	0.41
Bio→Relationship	0.25	0.57	0.41
Relationships→FollowsBrand	0.40	-	0.40
Bio→Education	0.35	0.43	0.39
Bio→Occupation	0.32	0.42	0.37
Relationships→Grouped	0.36	-	0.36
Geographical→Current	0.51	0.18	0.35
Contact→Email	0.28	-	0.28
Visual→Photos	0.25	-	0.25
Relationships→Contributes	0.24	-	0.24
Bio→Habits	0.23	-	0.23
Attributes→Spatial	0.10	0.33	0.22
Opinion→Brand	0.22	-	0.22
Content→Video	0.22	-	0.22
Contact→Phone	0.10	-	0.10
Geographical→History	0.12	0.01	0.07

6.5.4 Summary

Each of the above tables demonstrates one measure of the value of online profile attributes to this specific task – detecting social engineering vulnerability. Table 6.5 combines all three measures to rank profile attributes such that the top-ranked are attributes most likely to produce relevant, true and novel information for this goal.

Three different combinations of these metrics are given in the table, the mean, the product, and the geometric mean. These Importance measurements are ones which gives equal weight to all concerns, which may not be appropriate for all situations – it is likely that a higher weight should be assigned to the defined relevance score, for example, rather than considering new and true information to be as important.

Table 6.5 Combined task-importance ranking of profile attributes

Attribute	*Relevance*	*Veracity*	*Novelty*	μ	$\prod$	$\sqrt[3]{\prod}$
Bio→Username	0.90	0.30	0.59	0.60	0.16	0.54
Temporal→Activity	0.50	0.60	0.55	0.55	0.17	0.55
Relationships→FollowedBy	0.10	0.90	0.62	0.54	0.06	0.38
Relationships→Interacted	0.10	0.60	0.82	0.51	0.05	0.37
Bio→Description	0.40	0.30	0.67	0.46	0.08	0.43
Temporal→Seen	0.20	0.60	0.55	0.45	0.07	0.40
Relationships→FollowsUser	0.10	0.60	0.66	0.45	0.04	0.34
Content→Text	0.20	0.30	0.78	0.43	0.05	0.36
Bio→Occupation	0.60	0.30	0.37	0.42	0.07	0.41
Attributes→Temporal	0.20	0.60	0.41	0.40	0.05	0.37
Content→Image	0.10	0.30	0.72	0.37	0.02	0.28
Contact→Email	0.50	0.30	0.28	0.36	0.04	0.35
Visual→Avatar	0.10	0.30	0.65	0.35	0.02	0.27
Attributes→Spatial	0.20	0.60	0.22	0.34	0.03	0.30
Geographical→History	0.30	0.60	0.07	0.32	0.01	0.23
Contact→Phone	0.50	0.30	0.10	0.30	0.02	0.25
Relationships→FollowsBrand	0.20	0.30	0.40	0.30	0.02	0.29
Geographical→Current	0.20	0.30	0.35	0.28	0.02	0.28
Bio→Relationship	0.10	0.30	0.41	0.27	0.01	0.23
Bio→Education	0.10	0.30	0.39	0.26	0.01	0.23
Relationships→Contributes	0.20	0.30	0.24	0.25	0.01	0.24
Relationships→Grouped	0.10	0.30	0.36	0.25	0.01	0.22
Visual→Photos	0.10	0.30	0.25	0.22	0.01	0.19
Bio→Habits	0.10	0.30	0.23	0.21	0.01	0.19
Opinion→Brand	0.10	0.30	0.22	0.21	0.01	0.19
Content→Video	0.10	0.30	0.22	0.21	0.01	0.19

Nonetheless, these baselines provides some insight into the social engineering vulnerability task. Generally speaking, the ranking by product or geometric mean is less forgiving of low values than the arithmetic mean (by which the table is ordered). This is particularly critical for attributes given low *relevance*, such as relationship information, which typically drop several places.

Taking the top results, it can be seen that `Bio→Username` is suggested as a critical value, reflecting its heavy relevance from expert judgement and a respectable novelty score due to its wide availability across sites making it possible for other names to be found.

`Temporal→Activity` is the second highest-scoring attribute under the arithmetic mean, and the highest under the other two measures. Activity logs can be useful for social

engineers in identifying business activity and staff absences and are not retroactively editable. With reasonable novelty due to different revelation patterns across different social networks, this attribute would appear to be generally of high value for collection. The `Temporal→Seen`, as a singular data point of a similar nature, also ranks respectably.

The `Relationships→FollowedBy` and `Relationships→Interacted` are only accentuator values for social engineering attacks, allowing attackers to refer to plausible context when delivering ploys, but are hard for a target to disguise and can present a range of new values in an iterative identity search system. They rank lower than other attributes under the measurements which more heavily penalise their low relevance in expert opinion, but still perform moderately, and are notable for high veracity.

These task-importance measures form an important complement to the ACU model for improving the identity-resolution system described in [67]. The approach taken by that paper is broad. A wide variety of attributes are included to capture possible identification value – these might now be more judiciously selected based on attribute availability, weighted by the consistency and uniqueness of methods operating on those attributes. This would reduce development effort and increase performance. Validation of this approach through user trials and/or expert review would be a useful direction for further work.

For the purposes of linking together online profiles, the ACU model describes which online profile attributes will be most useful. For the purposes of deciding which OSNs to investigate, however, the combination of relevance, novelty and veracity is most instructive – by first identifying the most important attributes to the purpose of social engineering vulnerability detection, OSNs providing these attributes can be included in identity resolution attempts. The results for this case are not necessarily limiting – expert social engineering advice most values an employee's `Bio→Username`, which is broadly available on most social networks. However, the importance of `Temporal→Activity` does highlight that platforms which encourage regular updates are preferred targets, whereas the low importance of `Content→Video` suggests video-sharing sites should

be among the last sources searched. These measures demonstrate value by imparting guidance for the development of identity-resolution applications.

Chapter 7

Conclusion & Future Work

This thesis has defined the ACU data quality measurement framework for the process and application of identity resolution. Throughout, the focus has been on methodological contributions to identity resolution as applied in online social networks.

By tying the central theory developed in Chapter 4 to clear methodological issues faced in building identity resolution classifiers, this thesis has aimed to ensure that data quality measurements can be connected with practical benefits for the development and application of identity resolution in online contexts.

7.1 Thesis Objectives Revisited

A solution to the ground-truth problem in identity resolution

Data collection and scientific replication of identity resolution results has been hampered by ethical and legal prohibitions against sharing personally identifiable information. Chapter 3 outlines and evaluates a method which would allow researchers to sidestep these issues, from the perspective of sampling theory. A method for drawing comparable datasets from online social networks is described, and realised in a publicly-available implementation. The results show that identity resolution performed on these samples performs similarly enough for the replication and validation of methods.

A model enabling comparative assessment of the identification value of all common profile attributes

The previous understanding of which online information items hold value for identity resolution was patchy and incomplete. Chapter 4 presents the ACU model for understanding the three dimensions of information value in identity resolution: *availability*, *consistency* and *uniqueness*, along with a well-grounded schema of online profile attributes and estimates of values based on prior literature and original observations. The insights drawn from these estimates are suggestive of areas of exploration for future identity resolution methods.

The purposes of an identity-resolution system are also paramount to evaluating its performance. Positing the use of online identity resolution as a form of search system, Chapter 6 outlines data quality measures which can be applied to understand the value of revealed attributes to a particular task, with a case study in a real application domain of social engineering vulnerability detection.

Improving the reliability of identity resolution methods on real datasets

The issue of missing data is often overlooked in the extant literature on identity resolution in online social networks, despite the widespread nature of this issue. Chapter 5 details the particular impact of missing data values and naive coping strategies, before applying the estimates for ACU model components to improve the situation through feature selection.

7.2 Implications for Preservation of Privacy

This thesis has so far focused on methodological issues for identity-resolution researchers and investigators seeking to apply such methods. However, these results also hold meaning for individuals seeking to protect their privacy on and between social networks and similar profile networks.

The attribute rankings provided in the tables of Section 5.2 isolate those profile attributes which are most open to identification attacks. These attributes are ones which should be targeted for defensive measures, and the ranking of identification value suggests a priority ranking for any defensive approaches. Our model also suggests three dimensions by which privacy protection may be approached on any attribute:

Availability The work in Section 4.4 produces general availability estimates for a range of profile features, while the combined model in Section 4.3 also reveals valuations weighted by estimates of the consistency and uniqueness of certain attributes. By not filling in these fields, or otherwise protecting access to profile fields wherever this is possible, users can effectively hamper the application of identity-resolution techniques to their own profiles, as demonstrated by the performance penalties of classification in Section 5.1. Moreover, users can intelligently focus their efforts to reduce availability on the more consistent and unique attributes, depriving attackers of their most valuable features.

However, social networks can hinder users in these efforts by making fields mandatory (or practically mandated for ordinary operation of the service) or else by recording data without the user being directly aware of or able to control its collection. Users have long been advised not to reveal personal information publicly online, and while this may still prove effective with critically identifying information, the consistent uptake of social media indicates that there is a real desire to share certain incidental information in public. As such, alternative defences should be explored and employed alongside availability-reduction, as is discussed next.

Consistency Where it is not possible or desirable to omit features, a user can alternately attack the consistency of their profile information across different networks by creating contradictory values for the same attribute in each profile. With dissimilar values in the fields most expected to be consistent, such as username, gender or location, most identity-resolution approaches will fail. Consistency is hard for social networks or other

interested parties to affect, as it regards a property that exists between what are by definition unlinked data records.

The most troublesome areas for consistency-based defences are inferential attacks based on data over which a user has at best limited control, such as their writing style, or the times at which they interact with social networking services. The ability of the user to make these properties inconsistent is limited. This might be mitigated through the use of technological aids which filter user output to disguise identifying traits.

There is a potential usability obstacle to consistency-based privacy preservation. Users who are maintaining separate accounts but not entire separate identities may find the requirement to use names, locations or genders other than their own in their online account is socially or otherwise disadvantageous. This fundamental tendency towards veracity could present a limit to the applicability of the principle for any users who are not strictly adversarial in their intentions.

Uniqueness Alternatively, a user can choose to avoid outlandishly inconsistent profile attributes by "blending in with the crowd" and adopting highly non-unique values for usually-unique attributes. This can be a more difficult approach for some attributes (such as network connections), but for more biographical details, selecting from the modal choices can effectively reduce the identifiability of a user.

The primary challenge for implementing such defences is practical difficulties for an individual attempting to learn a property of the entire network. Systems that could automate a sampling process to gather up-to-date attribute value modes for a particular network could form a technical solution to this hurdle. A more social solution could require the collusion of the social network in ploys such as replacing any 'private' with the platform's own measure of the least-informative value, or a randomised value. Introducing this unreliability instead of more identifiable missing data would contribute to active defences against identity-resolution attacks on a particular platform.

7.3 Future Work

All the developments reported in this thesis reveal areas where further research could be directed.

7.3.1 Reproducibility of identity-resolution results

There are two key strands of work which could form significant benefit to the reproducibility of identity-resolution results.

Firstly, the methodology tested in Chapter 3 was realised in a particular implementation centred on the Google+ social network. This network is becoming dated as a relevant source of profile connections. Finding an alternative source for realising this data collection method, particularly a more high-volume or current source, could provide immense practical benefit to the field. One avenue here is the use of Twitter URL fields, which are not explicitly marked as identifiers, but often practically used as such. Other sources, such as blogging platforms, might also prove useful. More broadly, future work in this area might focus on improving this approach by finding other practical search methods than name-based searches to use for random sampling, such as content or network-based search.

Moving beyond data collection, the establishment of an appropriate means of sharing standardised data transformations – such as a library of competitive similarity functions that operate over the range of data types available in OSNs – would help standardise the field. A common evaluation framework based on current practices in machine learning could also greatly improve the comparability of state-of-the-art methods, and one such as that provided by Köpcke et al. [124] could be used for the organisation of competitive events to spur development and comparison of new methods

7.3.2 Validating the ACU model

With regard to the ACU model developed in Chapter 4, whilst the profile schema was grounded in a broad survey of online profile information, the original measurements

leading to attribute-level estimates of *availability*, *consistency* and *uniqueness* were drawn from data sampled according to the method from Chapter 3, and as such these estimates relate mostly to a few of the more dominant social networks at the time of publication. Drawing comparable measurements from identity-resolution data collected by other means would provide valuable perspective on the generalisability of these estimates.

More broadly, constructing a profile schema and associated *availability*, *consistency* and *uniqueness* measurements for an identity-resolution domain outside of online social networking (such as bibliometrics) would be a valuable test of the generalisability of the ACU model.

7.3.3 Application to adversarial cases

The ACU model could in general be applicable to adversarial scenarios, but the data sampling methodology in Chapter 3 does not speak to adversarial profiles, and so the estimates given in Chapter 4 do not apply to adversarial profiles. It may also be the case that the ACU model lacks elements which are important for adversarial usage, and that e.g., *veracity* information should be integrated. Missing data issues such as those explored in Chapter 5 are also affected, as information revelation patterns are likely distinct in adversarial datasets.

The primary hurdle for any research into adversarial identity resolution will be constructing a suitable database of intentionally-disconnected profiles at a volume and under sampling criteria suitable for machine learning. Once this is overcome, modelling the information value of attributes under ACID, ACU or an adjusted construction is enabled, and information revelation patterns can also be studied.

7.3.4 Missing data in identity resolution

With regard to the challenge of missing data, ongoing research might focus on the result that classifiers which are robust in the face of missing data are often more effective than statistical imputation in these types of high-dimensional datasets. Replication of this

result in classification problems across different domains should establish how generally this may apply.

It is possible that classification problems under missing data conditions could benefit from a dual-classifier approach – using a robust method such as Naive Bayes to handle cases where data is missing, and using a more fragile but more powerful classification system for complete cases. Exploration of this angle could build upon the results in Chapter 5.

7.3.5 Validating application-specific measures

Finally, regarding application-area quality measurements, trials could be made of the predictive power of the *novelty* metric as conditional on an accumulating body of attributes on a subject. This would require an intensive iterative search process against a ground-truth dataset across many social networks, building on the designs of Bennacer [26]. Keeping such a result up to date with the social networking landscape is also a challenging problem for the area.

The *veracity* estimation method described in Chapter 6 could be tuned to include the suggested adjustments based on observed values rather than relying solely on a judgement of the locus of control for an attribute. Surveying the truthfulness of attribute reports as Rowe et al. did is probably the most reliable method for this, but does not permit great scale or particularly diverse sampling, so an alternative methodology may have to be developed. This result would likely be of interest to sociologists and psychologists using OSN data, outside of its relevance to identity resolution.

7.4 Concluding Remarks

The internet provides investigators with a glut of data which may or may not be of aid to their work. In many situations, this overabundance has a paralysing effect – some law enforcement departments are unable to properly handle the volume of information, and are forced to ignore it all for fear of being overwhelmed. But as traces of crime

increasingly appear – perhaps solely – online, this approach cannot be maintained. To enable investigations, technologies must be developed which can retrieve from the data the information which is *useful* to a case, and present it appropriately to investigators.

For this to be possible, a theoretical understanding of what *useful* means for investigators is required. This thesis has sought to address this question within the narrow domain of identity resolution, which seems to be a problem of much relevance for online investigation. Data quality measures for the identifiability of online profile data have been developed, and some benefits from them demonstrated. The results are no doubt incomplete – this is the scientific process – but, hopefully, will contribute meaningfully to the cohesion and reliability of ongoing research. .

References

[1] Abbasi, A. (2007). Affect intensity analysis of dark web forums. In *Proceedings of the 2007 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 282–288. IEEE.

[2] Abbasi, A. and Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems*, 20(5):67–75.

[3] Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7:1–7:29.

[4] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12:1–12:34.

[5] Abel, F., Henze, N., Herder, E., and Krause, D. (2010). Interweaving public user profiles on the web. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP)*, pages 16–27. Springer.

[6] Abel, F., Herder, E., Houben, G.-J., Henze, N., and Krause, D. (2013). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209.

[7] Aggarwal, S., Bali, J., Duan, Z., and Kermes, L. (2007). The design and development of an undercover multipurpose anti-spoofing kit (UnMask). In *Proceedings of the 23rd Annual Computer Security Applications Conference (ACSAC)*, pages 141–150. IEEE.

[8] Airoldi, E. and Malin, B. (2004). Scamslam: An architecture for learning the criminal relations behind scam spam. Technical Report CMU-ISRI-04-121.

[9] Al-Zaidy, R., Fung, B., Youssef, A. M., and Fortin, F. (2012). Mining criminal networks from unstructured text documents. *Digital Investigation*, 8(3):147–160.

[10] Appavu alias Balamurugan, S. and Rajaram, R. (2008). Learning to classify threaten e-mail. In *Proceedings of the 2nd Asia International Conferenced on Modeling & Simulation (AICMS)*, pages 522–527. IEEE.

[11] Appavu alias Balamurugan, S., Rajaram, R., Athiappan, G., and Muthupandian, M. (2007). Data mining techniques for suspicious email detection: a comparative study. In *Proceedings of the European Conference on Data Mining (ECDM)*, pages 213–217. IADIS.

[12] Appavu alias Balamurugan, S., Rajaram, R., Muthupandian, M., Athiappan, G., and Kashmeera, K. (2009). Data mining based intelligent analysis of threatening e-mail. *Knowledge-Based Systems*, 22(5):392–393.

[13] Atig, M. F., Cassel, S., Kaati, L., and Shrestha, A. (2014). Activity profiles in online social media. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 850–855. IEEE.

[14] Atkinson, M., Belayeva, J., Zavarella, V., Piskorski, J., Huttunen, S., Vihavainen, A., and Yangarber, R. (2010). News mining for border security intelligence. In *Proceedings of the 2010 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 173–173. IEEE.

[15] Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy. Technical report, Department of Computer Engineering, Chalmers University of Technology.

[16] Bahri, L., Carminati, B., and Ferrari, E. (2016). COIP—continuous, operable, impartial, and privacy-aware identity validity estimation for OSN profiles. *ACM Transactions on the Web (TWEB)*, 10(4):23.

[17] Baili, N., D'Souza, D., Mohamed, A., and Yampolskiy, R. (2011). Avatar face recognition using wavelet transform and hierarchical multi-scale LBP. In *Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, pages 194–199. IEEE.

[18] Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., and Kruegel, C. (2010). Abusing social networks for automated user profiling. In *Proceedings of the 13th International Symposium on Recent Advances in Intrusion Detection (RAID)*, pages 422–441. Springer.

[19] Bali, J. S. (2007). Automation of email analysis using a database. Master's thesis, Florida State University.

[20] Ball, L. D., Ewan, G., and Coull, N. J. (2012). Undermining: social engineering using open source intelligence gathering. In *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, pages 275–280. SciTePress.

[21] Banday, M. T., Qadri, J. A., Jan, T., Shah, N., et al. (2011). Detecting threat e-mails using bayesian approach. *International Journal of Secure Digital Information Age*, 1(2):10.

[22] Barbian, G. (2011). Detecting hidden friendship in online social network. In *Proceedings of the 2011 European Intelligence and Security Informatics Conference (EISIC)*, pages 269–272. IEEE.

[23] Bartunov, S., Korshunov, A., Park, S. T., Ryu, W., and Lee, H. (2012). Joint link-attribute user identity resolution in online social networks. In *Proceedings of the 6th Workshop on Social Network Mining and Analysis, at the 18th International Conference on Knowledge Discovery and Data Mining (SNA-KDD)*. ACM.

[24] Bekkerman, R. and McCallum, A. (2005). Disambiguating web appearances of people in a social network. In *Proceedings of the 14th International Conference on World Wide Web (WWW)*, pages 463–470. ACM.

[25] Benjamin, V. and Chen, H. (2012). Securing cyberspace: Identifying key actors in hacker communities. In *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 24–29. IEEE.

[26] Bennacer, N., Jipmo, C. N., Penta, A., and Quercini, G. (2014). Matching user profiles across social networks. In *Proceedings of the 26th International Conference on Advanced Information Systems Engineering (CAiSE)*, pages 424–438. Springer.

[27] Bernard, K., Cassidy, A., Clark, M., Liu, K., Lobaton, K., McNeill, D., and Brown, D. (2011). Identifying and tracking online financial services through web mining and latent semantic indexing. In *Systems and Information Engineering Design Symposium (SIEDS), 2011 IEEE*, pages 158–163. IEEE.

[28] Black, S., Creese, S., Guest, R., Pike, B., Saxby, S. J., Stanton Fraser, D., Stevenage, S. V., and Whittty, M. (2012). Superidentity: Fusion of identity across real and cyber domains. In *Proceedings of ID360: Global Issues in Identity*. CSID.

[29] Bogdanova, D., Rosso, P., and Solorio, T. (2012a). Modelling fixated discourse in chats with cyberpedophiles. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 86–90. Association for Computational Linguistics.

[30] Bogdanova, D., Rosso, P., and Solorio, T. (2012b). On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 110–118. Association for Computational Linguistics.

[31] Bouquet, P. and Bortoli, S. (2010). Entity-centric social profile integration. In *Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web (LUPAS)*, pages 52–57. CEUR-WS.org.

[32] Broadway, J., Turnbull, B., and Slay, J. (2008). Improving the analysis of lawfully intercepted network packet data captured for forensic analysis. In *Proceedings of the 3rd International Conference on Availability, Reliability and Security (ARES)*, pages 1361–1368. IEEE.

[33] Bruce, J., Scholtz, J., Hodges, D., Emanuel, L., Stanton Fraser, D., Creese, S., and Love, O. J. (2014). Pathways to identity: using visualization to aid law enforcement in identification tasks. *Security Informatics*, 3(1):12.

[34] Buccafurri, F., Lax, G., Nocera, A., and Ursino, D. (2012a). Crawling social internetworking systems. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 506–510. IEEE.

[35] Buccafurri, F., Lax, G., Nocera, A., and Ursino, D. (2012b). Discovering links among social networks. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 467–482. Springer.

[36] Budgen, D., Turner, M., Brereton, P., and Kitchenham, B. (2008). Using mapping studies in software engineering. In *Proceedings of the 20th Annual Meeting of the Pschology of Programming Interest Group (PPIG)*, volume 8, pages 195–204. Lancaster University.

[37] Carmagnola, F., Osborne, F., Torre, I., and White, B. (2009). Cross-systems identification of users in the social web. In *Proceedings of the IADIS International Conference WWW/Internet*, pages 129–134. IADIS.

[38] Chandler, D. (1998). Personal home pages and the construction of identities on the web. http://www.aber.ac.uk/~dgc/Webident.html, Accessed: June 2014.

[39] Chang, W., Ku, Y., Wu, S., and Chiu, C. (2012). CybercrimeIR–a technological perspective to fight cybercrime. In Chau, M., Wang, G., Yue, W., and Chen, H., editors, *Intelligence and Security Informatics*, volume 7299 of *Lecture Notes in Computer Science*, pages 36–44. Springer.

[40] Chau, M. and Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1):57 – 70.

[41] Chaurasia, N., Dhakar, M., Chharia, A., Tiwari, A., and Gupta, R. (2012). Exploring the current trends and future prospects in terrorist network mining. In *Proceedings of The 2nd International Conference on Computer Science, Engineering and Applications (CCSEA 2012)*, volume 2, pages 379–385. CS&CP.

[42] Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., and Weimann, G. (2008a). Uncovering the dark web: A case study of jihad on the web. *Journal of the American Society for Information Science and Technology*, 59(8):1347–1359.

[43] Chen, H., Qin, J., Reid, E., and Zhou, Y. (2008b). Studying global extremist organizations' internet presence using the darkweb attribute system. *Terrorism Informatics*, 18:237–266.

[44] Chen, T., Chaabane, A., Tournoux, P. U., Kaafar, M. A., and Boreli, R. (2013). How much is too much? Leveraging ads audience estimation to evaluate public profile uniqueness. In *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS)*, pages 225–244. Springer.

[45] Chen, T., Kaafar, M. A., Friedman, A., and Boreli, R. (2012a). Is more always merrier?: A deep dive into online social footprints. In *Proceedings of the 2012 ACM Workshop on Online Social Networks (WOSN)*, pages 67–72. ACM.

[46] Chen, X., Hao, P., Chandramouli, R., and Subbalakshmi, K. (2011). Authorship similarity detection from email messages. In Perner, P., editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 6871 of *Lecture Notes in Computer Science*, pages 375–386. Springer.

[47] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012b). Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and the 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE.

[48] Cheng, H., Liang, Y.-L., Xing, X., Liu, X., Han, R., Lv, Q., and Mishra, S. (2012). Efficient misbehaving user detection in online video chat services. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 23–32. ACM.

[49] Cheong, M. and Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, 13(1):45–59.

[50] Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media. ISBN: 9783642430015 3642430015.

[51] Chung, W. (2012). Categorizing temporal events: A case study of domestic terrorism. In *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 159–161. IEEE.

[52] Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *Proceedings of the 2003 KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, volume 3, pages 73–78. ACM.

[53] Corney, M. W. (2003). *Analysing e-mail text authorship for forensic purposes*. PhD thesis, Queensland University of Technology.

[54] Cortis, K., Scerri, S., Rivera, I., and Handschuh, S. (2012). Discovering semantic equivalence of people behind online profiles. In *Proceedings of the 5th International Resource Discovery Workshop (RED), co-hosted with the 9th Extended Semantic Web Conference*, pages 1–22. Springer.

[55] Creese, S., Goldsmith, M., Nurse, J. R. C., and Phillips, E. (2012). A data-reachability model for elucidating privacy and security risks related to the use of online social networks. In *Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, (TrustCom)*, pages 1124–1131. IEEE.

[56] Dardick, G. S., La Roche, C. R., and Flanigan, M. A. (2007). Blogs: Anti-forensics and counter anti-forensics. In *Proceedings of the 5th Australian Digital Forensics Conference*. School of Computer and Information Science, Edith Cowan University, Perth, Western Australia.

[57] DARPA (2003). Lifelog proposer information pamphlet, solicitation number BAA03-30DARPA IPTO. https://web.archive.org/web/20040226044110/http://www.darpa.mil/ipto/solicitations/closed/03-30%5FPIP.htm Accessed: 2016-02-23.

[58] Dazeley, R., Yearwood, J. L., Kang, B. H., and Kelarev, A. V. (2010). Consensus clustering and supervised classification for profiling phishing emails in internet commerce security. In *Proceedings of the Pacific Rim Knowledge Acquisition Workshop (PKAW)*, pages 235–246. Springer.

[59] De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Multi-topic e-mail authorship attribution forensics. In *Proceedings of the Workshop on Data Mining for Security Applications at the ACM Conference on Computer Security (CCS)*. ACM.

[60] Ding, L., Zhou, L., Finin, T., and Joshi, A. (2005). How the semantic web is being used: An analysis of FOAF documents. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS)*, pages 113c–113c. IEEE.

[61] Ding, X., Zhang, L., Wan, Z., and Gu, M. (2010). A brief survey on de-anonymization attacks in online social networks. In *Proceedings of the 2010 International Conference on Computational Aspects of Social Networks (CASoN)*, pages 611–615. IEEE.

[62] Do, T., Chang, K., and Hui, S. (2004). Web mining for cyber monitoring and filtering. In *Proceedings of the 2004 IEEE International Conference on Cybernetics and Intelligent Systems*, pages 399–404. IEEE.

[63] Dreier, D. J. (2009). Blog fingerprinting identifying anonymous posts written by an author of interest using word and character frequency analysis. Master's thesis, Monterey, California; Naval Postgraduate School.

[64] Dudas, P. M. (2013). Cooperative, dynamic twitter parsing and visualization for dark network analysis. In *Proceedings of the 2nd IEEE International Network Science Workshop (NSW)*, pages 172–176. IEEE.

[65] Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nation's Health*, 36(12):1412–1416.

[66] Dwyer, C., Hiltz, S., and Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proceedings of the 13th Americas' Conference on Information Systems (AMCIS)*, pages 339–351. AIS.

[67] Edwards, M., Larson, R., Green, B., Rashid, A., and Baron, A. (2017). Panning for gold: Automatically analysing online social engineering attack surfaces. *Computers & Security*, 69:18–34.

[68] Edwards, M., Rashid, A., and Rayson, P. (2014). A service-independent model for linking online user profile information. In *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 280–283. IEEE.

[69] Edwards, M., Rashid, A., and Rayson, P. (2015). A systematic survey of online data mining technology intended for law enforcement. *ACM Computing Surveys*, 48(1):15:1–15:54.

[70] Edwards, M., Wattam, S., Rayson, P., and Rashid, A. (2016). Sampling labelled profile data for identity resolution. In *Proceedings of the IEEE International Conference on Big Data (BigData)*, pages 540–547. IEEE.

[71] Elovici, Y., Kandel, A., Last, M., Shapira, B., and Zaafrany, O. (2004). Using data mining techniques for detecting terror-related activities on the web. *Journal of Information Warfare*, 3(1):17–29.

[72] Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., and Kandel, A. (2010). Detection of access to terror-related web sites using an advanced terror detection system (ATDS). *Journal of the American Society for Information Science and Technology*, 61(2):405–418.

[73] Endy, E., Lim, C., Eng, K., and Nugroho, A. (2010). Implementation of intelligent searching using self-organizing map for webmining used in document containing information in relation to cyber terrorism. In *Proceedings of the 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT)*, pages 195–197. IEEE.

[74] Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

[75] Fong, S., Roussinov, D., and Skillicorn, D. B. (2008). Detecting word substitutions in text. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1067–1076.

[76] Fournier, R., Cholez, T., Latapy, M., Magnien, C., Chrisment, I., Daniloff, I., and Festor, O. (2014). Comparing paedophile activity in different P2P systems. *Social Sciences*, 3(3):314–325.

[77] Frank, R., Westlake, B., and Bouchard, M. (2010). The structure and content of online child exploitation networks. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD)*, pages 3:1–3:9. ACM.

[78] Friedland, G., Maier, G., Sommer, R., and Weaver, N. (2011). Sherlock Holmes' evil twin: on the impact of global inference for online privacy. In *Proceedings of the 2011 Workshop on New Security Paradigms*, pages 105–114. ACM.

[79] Gawron, J. M., Gupta, D., Stephens, K., Tsou, M.-H., Spitzberg, B., and An, L. (2012). Using group membership markers for group identification in web logs. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 467–470. Association for the Advancement of Artificial Intelligence.

[80] Ge, A., Mao, W., and Zeng, D. (2010). Story extraction from the web: A case study in security informatics. In *Proceedings of the IEEE International Conference on Service Operations and Logistics and Informatics (SOLI)*, pages 306–310. IEEE.

[81] Gerstenfeld, P. B., Grant, D. R., and Chiang, C.-P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy*, 3(1):29–44.

[82] Giacobe, N., Kim, H.-W., and Faraz, A. (2010). Mining social media in extreme events : Lessons learned from the DARPA network challenge. In *Proceedings of the IEEE International Conference on Technologies for Homeland Security (HST)*, pages 165–171. IEEE.

[83] Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., and Teixeira, R. (2013a). Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 447–458. IW3C2.

[84] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., and Gummadi, K. P. (2015). On the reliability of profile matching across large online social networks. In *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining (SIGKIDD)*, pages 1799–1808. ACM.

[85] Goga, O., Perito, D., Lei, H., Teixeira, R., and Sommer, R. (2013b). Large-scale correlation of accounts across social networks. Technical Report TR-13-002, University of California at Berkeley.

[86] Golbeck, J. and Rothstein, M. (2008). Linking social networks on the web with FOAF: A semantic web case study. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, volume 8, pages 1138–1143. AAAI.

[87] Gonzalez, R., Cuevas, R., Motamedi, R., Rejaie, R., and Cuevas, A. (2013). Google+ or Google-?: dissecting the evolution of the new OSN in its first year. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 483–494. IW3C2.

[88] Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576.

[89] Gray, A., Sallis, P., and MacDonell, S. (1997). Software forensics: Extending authorship analysis techniques to computer programs. In *Information Science Discussion Papers Series*. University of Otago.

[90] Gray, G. L. and Debreceny, R. (2007). Data mining of emails to support periodic and continuous assurance. Technical report, College of Business and Economics, California State University at Northridge.

[91] Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., and Benredjem, D. (2009). Towards an integrated e-mail forensic analysis framework. *Digital Investigation*, 5(3):124–137.

[92] Haggerty, J., Llewellyn-Jones, D., and Taylor, M. (2008). Forweb: file fingerprinting for automated network forensics investigations. In *Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia (e-Forensics)*, pages 29:1–29:6. ICST.

[93] Hidalgo, J. M. G. and Díaz, A. A. C. (2012). Combining predation heuristics and chat-like features in sexual predator identification. In *CLEF (Online Working Notes/Labs/Workshop)*.

[94] Hosseinkhani, J., Chaprut, S., and Taherdoost, H. (2012). Criminal network mining by web structure and content mining. In *Proceedings of the 11th WSEAS International Conference on Information Security and Privacy*, pages 24–26. WSEAS.

[95] Hu, W., Wu, O., Chen, Z., Fu, Z., and Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1019–1034.

[96] Huber, M., Kowalski, S., Nohlberg, M., and Tjoa, S. (2009). Towards automating social engineering using social networking sites. In *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE)*, volume 3, pages 117–124. IEEE.

[97] Hui, W., Yin, H., and Lin, C. (2009). Design and deployment of a digital forensics service platform for online videos. In *Proceedings of the First ACM workshop on Multimedia in forensics*, MiFor '09, pages 31–36, New York, NY, USA. ACM.

[98] Hui, W., Zhao, H., Lin, C., and Yang, Y. (2012). Videcloud: Efficient support for large-scale video copy detection. *Journal of Computational Information Systems*, 8(3):1055–1062.

[99] Ibrahim, A. A. (2009). Detecting and preventing the electronic transmission of illicit images. Master's thesis, University of Ontario Institute of Technology.

[100] Inches, G. and Crestani, F. (2011). Online conversation mining for author characterization and topic identification. In *Proceedings of the 4th Workshop for Ph.D. Students in Information & Knowledge Management (PKIM)*, pages 19–26. ACM.

[101] Inches, G. and Crestani, F. (2012). Overview of the international sexual predator identification competition at PAN-2012. In *CLEF (Online Working Notes/Labs/Workshop)*.

[102] Inyaem, U., Meesad, P., Haruechaiyasak, C., and Tran, D. (2009). Ontology-based terrorism event extraction. In *Proceedings of the 1st International Conference on Information Science and Engineering (ICISE)*, pages 912–915. IEEE.

[103] Iofciu, T., Fankhauser, P., Abel, F., and Bischoff, K. (2011). Identifying users across social tagging systems. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*, pages 522–525. AAAI.

[104] Iqbal, F. (2011). *Messaging Forensic Framework for Cybercrime Investigation*. PhD thesis, Concordia University.

[105] Iqbal, F., Binsalleeh, H., Fung, B., and Debbabi, M. (2010a). Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1):56–64.

[106] Iqbal, F., Binsalleeh, H., Fung, B. C., and Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112.

[107] Iqbal, F., Hadjidj, R., Fung, B., and Debbabi, M. (2008). A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:S42–S51.

[108] Iqbal, F., Khan, L. A., Fung, B. C. M., and Debbabi, M. (2010b). e-mail authorship verification for forensic investigation. In *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC)*, pages 1591–1598. ACM.

[109] Irani, D., Webb, S., Li, K., and Pu, C. (2009). Large online social footprints–an emerging threat. In *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE)*, volume 3, pages 271–276. IEEE.

[110] Islam, M., Watters, P. A., and Yearwood, J. (2011). Real-time detection of children's skin on social networking sites using markov random field modelling. *Information Security Technical Report*, 16(2):51 – 58.

[111] Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10):94–100.

[112] Jain, P. (2015). Automated methods for identity resolution across heterogeneous social platforms. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT)*, pages 307–310. ACM.

[113] Jain, P., Kumaraguru, P., and Joshi, A. (2013). @i seek 'fb.me': Identifying users across multiple online social networks. In *Proceedings of the 22nd International Conference on World Wide Web (WWW) Companion*, pages 1259–1268. IW3C2.

[114] Jayanthi, S. and Sasikala, M. S. (2011). XGraphticsCLUS: Web mining hyperlinks and content of terrorism websites for homeland security. *International Journal of Advanced Networking and Applications*, 2(6):941–949.

[115] Johnson, J., Miller, A., Khan, L., and Thuraisingham, B. (2012). Measuring relatedness and augmentation of information of interest within free text law enforcement documents. In *Proceedings of the 2012 European Intelligence and Security Informatics Conference (EISIC)*, pages 148–155. IEEE.

[116] Kaafar, M. A. and Manils, P. (2010). Why spammers should thank Google? In *Proceedings of the 3rd Workshop on Social Network Systems (SNS)*, pages 4:1–4:6. ACM.

[117] Karran, A. J. and Llewellyn-Jones, D. (2009). A digital forensics analytical process model for the investigation, analysis and visualisation of social networks derived from e-mail. Master's thesis, Liverpool John Moores University.

[118] Khan, A. M. R. (2012). A simple but powerful e-mail authorship attribution system. In *Proceedings of the 4th International Conference on Machine Learning and Computing*, pages 151–155. IACSIT.

[119] Klimt, B. and Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning (ECML)*, pages 217–226. Springer.

[120] Klinger, E. and Starkweather, D. (2016). pHash – the open source perceptual hash library. http://www.phash.org/apps/ Accessed 2016-05-19.

[121] Kontaxis, G., Polakis, I., Ioannidis, S., and Markatos, E. P. (2011). Detecting social network profile cloning. In *Proceedings of the 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 295–300. IEEE.

[122] Kontostathis, A., Edwards, L., Bayzick, J., Leatherman, A., and Moore, K. (2009). Comparison of rule-based to human analysis of chat logs. *Communication Theory*, 8:2–13.

[123] Kontostathis, A., Edwards, L., and Leatherman, A. (2010). Text mining and cybercrime. In *Text Mining: Applications and Theory*, pages 149–164. John Wiley & Sons, Ltd.

[124] Köpcke, H., Thor, A., and Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493.

[125] Koppel, M. and Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72. Morgan Kaufmann.

[126] Krishnamurthy, B. and Wills, C. E. (2009). On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 7–12. ACM.

[127] Ku, C. H., Iriberri, A., and Leroy, G. (2008). Natural language processing and e-government: crime information extraction from heterogeneous data sources. In *Proceedings of the 2008 International Conference on Digital Government Research (DG.O)*, pages 162–170. Digital Government Society of North America.

[128] Labitzke, S., Taranu, I., and Hartenstein, H. (2011). What your friends tell others about you: Low cost linkability of social network profiles. In *Proceedings of the 5th International ACM Workshop on Social Network Mining and Analysis*, pages 1065–1070. ACM.

[129] Lauw, H., Lim, E.-P., Pang, H., and Tan, T.-T. (2005). Social network discovery by mining spatio-temporal events. *Computational & Mathematical Organization Theory*, 11:97–118.

[130] Layton, R., Watters, P., and Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In *Proceedings of the 2nd Cybercrime and Trustworthy Computing Workshop (CTC)*, pages 1–8. IEEE.

[131] Lenselink, L. (2011). Radicalization online. patterns of social interaction on the al-faloja and as-ansar forums. Master's thesis, Utrecht University.

[132] Lim, M.-H., Negnevitsky, M., and Hartnett, J. (2007). Detecting abnormal changes in e-mail traffic using hierarchical fuzzy systems. In *Proceedings of the 2007 IEEE International Fuzzy Systems Conference (FUZZ-IEEE)*, pages 1–6. IEEE.

[133] Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R. (2014). Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM International Conference on Management of Data (SIGMOD)*, pages 51–62. ACM.

[134] Liu, Z., Yang, Z., Liu, S., and Shi, Y. (2012). Semi-random subspace method for writeprint identification. *Neurocomputing*, 108:93–102.

[135] Ma, J., Li, Y., Teng, G., Wang, F., and Zhao, Y. (2008). Sequential pattern mining for chinese e-mail authorship identification. In *Proceedings of the 3rd International Conference on Innovative Computing Information and Control*, pages 73–73. IEEE.

[136] Ma, J., Teng, G., Chang, S., Zhang, X., and Xiao, K. (2011). Social network analysis based on authorship identification for cybercrime investigation. *Intelligence and Security Informatics*, pages 27–35.

[137] Ma, J., Teng, G., Zhang, Y., Li, Y., and Li, Y. (2009a). A cybercrime forensic method for chinese web information authorship analysis. *Intelligence and Security Informatics*, pages 14–24.

[138] Ma, L., Yearwood, J., and Watters, P. (2009b). Establishing phishing provenance using orthographic features. In *Proceedings of the 2009 eCrime Researchers Summit (eCRIME)*, pages 1–10. IEEE.

[139] Malhotra, A., Totti, L., Meira Jr., W., Kumaraguru, P., and Almeida, V. (2012). Studying user footprints in different online social networks. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1065–1070. IEEE.

[140] Malin, B. (2005). Unsupervised name disambiguation via social network similarity. In *Proceedings of the 4th Workshop on Link Analysis, Counterterrorism, and Security at the 2005 SIAM International Conference on Data Mining*, pages 93–102. SIAM.

[141] Marcus, S. (1998). Dynamic data mining for information exploitation. In *Proceedings of the Information Technology Conference*, pages 79–82. IEEE.

[142] Marjuni, S., Mahmod, R., Ghani, A., Bin Mohd Zain, A., and Mustapha, A. (2009). Lexical criminal identification for chatting corpus. In *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology, (ICCSIT)*, pages 360–364. IEEE.

[143] Markham, A. and Buchanan, E. (2012). Ethical decision-making and internet research. *Recommendations from the AoIR Ethics Working Committee (Version 2.0)*.

[144] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., and Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122.

[145] Meissel, K. (2010). A practical guide to using Cliff's delta as a measure of effect size where parametric equivalents are inappropriate. In *Proceedings of the 2nd ACSPRI Social Science Methodology Conference*.

[146] Michalopoulos, D. and Mavridis, I. (2011). Utilizing document classification for grooming attack recognition. In *Computers and Communications (ISCC), 2011 IEEE Symposium on*, pages 864–869. IEEE.

[147] Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society of Information Science*, 48(9):810–832.

[148] Modupe, A., Olugbara, O., and Ojo, S. (2011). Exploring support vector machines and random forests to detect advanced fee fraud activities on internet. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 331–335. IEEE.

[149] Mohamed, A. A. and Yampolskiy, R. V. (2012a). Using discrete wavelet transform and eigenfaces for recognizing avatars' faces. In *Proceedings of the 17th International Conference on Computer Games (CGAMES)*, pages 143–147. IEEE.

[150] Mohamed, A. A. and Yampolskiy, R. V. (2012b). Wavelet based statistical adapted local binary patterns for recognizing avatar faces. In *Proceedings of the 1st International Conference on Advanced Machine Learning Technologies and Applications (AMLTA)*, pages 92–101. Springer.

[151] Morris, C. (2013). Identifying online sexual predators by SVM classification with lexical and behavioral features. Master's thesis, Department of Computer Science, University of Toronto.

[152] Morris, C. and Hirst, G. (2012). Identifying sexual predators by SVM classification with lexical and behavioral features. In *CLEF (Online Working Notes/Labs/Workshop)*.

[153] Motoyama, M. and Varghese, G. (2009). I seek you: searching and matching individuals in social networks. In *Proceedings of the 11th International Workshop on Web Information and Data Management (WIDM)*, pages 67–75. ACM.

[154] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., and Song, D. (2012). On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (S&P)*, pages 300–314. IEEE.

[155] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P)*, pages 111–125. IEEE.

[156] Narayanan, A. and Shmatikov, V. (2009). De-anonymizing social networks. In *Proceedings of the 2009 IEEE Symposium on Security and Privacy (S&P)*, pages 173–187. IEEE.

[157] Negnevitsky, M., Lim, M.-H., Hartnett, J., and Reznik, L. (2005). Email communications analysis: how to use computational intelligence methods and tools? In *Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety (CIHSPS)*, pages 16–23. IEEE.

[158] Newcombe, H. B. (1967). Record linking: the design of efficient systems for linking records into individual and family histories. *American Journal of Human Genetics*, 19(3.1):335–359.

[159] Newcombe, H. B., Kennedy, J. M., Axford, S., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130(3381):954–959.

[160] Nirkhi, S., Dharaskar, R., and Thakre, V. (2012). Analysis of online messages for identity tracing in cybercrime investigation. In *Proceedings of the 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, pages 300–305. IEEE.

[161] Nizamani, S., Memon, N., Wiil, U. K., and Karampelas, P. (2013). Modeling suspicious email detection using enhanced feature selection. *International Journal of Modeling and Optimization*, 2(4):371–377.

[162] Nosko, A., Wood, E., and Molema, S. (2010). All about me: Disclosure in online social networking profiles: The case of Facebook. *Computers in Human Behavior*, 26(3):406–418.

[163] Okoli, C. and Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Sprouts: Working Papers on Information Systems*, 10(26):49–99.

[164] Orebaugh, A. (2006). An instant messaging intrusion detection system framework: Using character frequency analysis for authorship identification and validation. In *Proceedings of the 40th Annual IEEE International Carnahan Conference on Security Technology*, pages 160–172. IEEE.

[165] Orebaugh, A. and Allnutt, D. J. (2010). Data mining instant messaging communications to perform author identification for cybercrime investigations. *Digital Forensics and Cyber Crime*, pages 99–110.

[166] Orebaugh, A. and Allnutt, J. (2009). Classification of instant messaging communications for forensics analysis. *The International Journal of Forensics Computer Science*, pages 22–28.

[167] Paice, C. D. et al. (1990). Another stemmer. In *ACM SIGIR Forum*, volume 24, pages 56–61. ACM.

[168] Panchenko, A., Beaufort, R., and Fairon, C. (2012). Detection of child sexual abuse media on P2P networks: Normalization and classification of associated filenames. In *Proceedings of the LREC Workshop on Language Resources for Public Security Applications*, pages 27–31. LREC.

[169] Panchenko, A., Beaufort, R., Naets, H., and Fairon, C. (2013). Towards detection of child sexual abuse media: categorization of the associated filenames. In *Advances in Information Retrieval*, pages 776–779. Springer.

[170] Pandit, S., Chau, D. H., Wang, S., and Faloutsos, C. (2007). Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 201–210. ACM.

[171] Patil, G., Manwade, K., and Landge, P. (2012). A novel approach for social network analysis & web mining for counter terrorism. *International Journal on Computer Science and Engineering*, 4(11):1816.

[172] Pearl, L. and Steyvers, M. (2012). Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and Linguistic Computing*, 27(2):183–196.

[173] Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 37–44. ACM.

[174] Peersman, C., Vaassen, F., Van Asch, V., and Daelemans, W. (2012). Conversation level constraints on pedophile detection in chat rooms. In *CLEF (Online Working Notes/Labs/Workshop)*.

[175] Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *Proceedings of the 2007 International Conference on Semantic Computing (ICSC)*, pages 235–241. IEEE.

[176] Peng, Y.-T. and Wang, J.-H. (2008). Link analysis based on webpage co-occurrence mining - a case study on a notorious gang leader in taiwan. In *Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 31–34. IEEE.

[177] Penna, L., Clark, A., and Mohay, G. (2010). A framework for improved adolescent and child safety in MMOs. In *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 33–40. IEEE.

[178] Perito, D., Castelluccia, C., Kaafar, M. A., and Manils, P. (2011). How unique and traceable are usernames? In *Proceedings of the 11th International Symposium on Privacy Enhancing Technologies (PETS)*, pages 1–17. Springer.

[179] Pham, D. D., Tran, G. B., and Pham, S. B. (2009). Author profiling for vietnamese blogs. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 190–194. IEEE.

[180] Plaxo (2016). Building an open social graph. http://www.plaxo.com/info/opensocialgraph Accessed: 2016-01-30.

[181] Prichard, J., Watters, P. A., and Spiranovic, C. (2011). Internet subcultures and pathways to the use of child pornography. *Computer Law & Security Review*, 27(6):585–600.

[182] Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K., and Momouchi, Y. (2010). In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research*, 1(3):135–154.

[183] Qin, J., Zhou, Y., Reid, E., Lai, G., and Chen, H. (2007). Analyzing terror campaigns on the internet: Technical sophistication, content richness, and web interactivity. *International Journal of Human-Computer Studies*, 65(1):71–84.

[184] Raad, E., Chbeir, R., and Dipanda, A. (2010). User profile matching in social networks. In *Proceedings of the 13th International Conference on Network-Based Information Systems (NBiS)*, pages 297–304. IEEE.

[185] Ríos, S. A. and Muñoz, R. (2012). Dark web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD)*, pages 2:1–2:7. ACM.

[186] Rohn, E. and Erez, G. (2012). Fighting agro-terrorism in cyberspace: A framework for intention detection using overt electronic data sources. In *Proceedings of the 9th International ISCRAM Conference*, pages 1–5. ISCRAM.

[187] Romaniuk, S. (2000). Using intelligent agents to identify missing and exploited children. In *IEEE Intelligent Systems and their Applications*, pages 27–30. IEEE.

[188] Rowe, M. (2009). Applying semantic social graphs to disambiguate identity references. In *Proceedings of the 6th European Semantic Web Conference (ESWC)*, pages 461–475. Springer.

[189] Rowe, M. (2010). The credibility of digital identity information on the social web: a user study. In *Proceedings of the 4th Workshop on Information Credibility (WICOW) at the 19th International Conference on World Wide Web*, pages 35–42. ACM.

[190] Rowe, M. and Ciravegna, F. (2010). Harnessing the social web: The science of identity disambiguation. In *Proceedings of the 2nd International Web Science Conference (WebSci)*.

[191] Sahito, F., Latif, A., and Slany, W. (2011). Weaving twitter stream into linked data: a proof of concept framework. In *Proceedings of the 7th International Conference on Emerging Technologies (ICET)*, pages 1–6. IEEE.

[192] Salem, A., Reid, E., and Chen, H. (2006). Content analysis of jihadi extremist groups' videos. *Intelligence and Security Informatics*, pages 615–620.

[193] Salem, A., Reid, E., and Chen, H. (2008). Multimedia content coding and analysis: Unraveling the content of jihadi extremist groups' videos. *Studies in Conflict & Terrorism*, 31(7):605–626.

[194] Scheelen, Y., Wagenaar, D., Smeets, M., and Kuczynski, M. (2012). The devil is in the details: Social engineering by means of social media. Technical report, System & Network Engineering, Universiteit van Amsterdam.

[195] Schmid, M. (2012). Computer-aided writeprint modelling for cybercrime investigations. Master's thesis, Concordia University.

[196] Shekar, D. C. and Imambi, S. S. (2008). Classifying and identifying of threats in e-mails–using data mining techniques. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, page 5.

[197] Shen, Q. and Boongoen, T. (2012). Fuzzy orders-of-magnitude-based link analysis for qualitative alias detection. In *Knowledge and Data Engineering, IEEE Transactions on*, pages 649–664. IEEE.

[198] Shupo, A., Martin, M. V., Rueda, L., Bulkan, A., Chen, Y., and Hung, P. C. (2006). Toward efficient detection of child pornography in the network infrastructure. *IADIS International Journal on Computer Science and Information Systems*, 1(2):15–31.

[199] Skillicorn, D. (2004). Detecting related message traffic. In *Workshop on Link Analysis, Security and Counterterrorism, SIAM Data Mining Conference*, pages 39–48.

[200] Skillicorn, D. (2010). Applying interestingness measures to ansar forum texts. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD)*, pages 7:1–7:9. ACM.

[201] Skillicorn, D. and Vats, N. (2007). Novel information discovery for intelligence and counterterrorism. *Decision Support Systems*, 43(4):1375 – 1382.

[202] Sobkowicz, P. and Sobkowicz, A. (2010). Dynamics of hate based internet user networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 73(4):633–643.

[203] Spencer, J. F. (2008). Using XML to map relationships in hacker forums. In *Proceedings of the 46th Annual Southeast Regional Conference on XX (ACM-SE)*, pages 487–489. ACM.

[204] Stallings, T., Wardman, B., Warner, G., and Thapaliya, S. (2012). "whois" selling all the pills. *International Journal of Forensic Computer Science*, 2:46–63.

[205] Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05):823–838.

[206] Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799.

[207] Statistic Brain Research Institute (2017). Google plus demographics & statistics. https://www.statisticbrain.com/google-plus-demographics-statistics/ Accessed: 2018-03-10.

[208] Steel, C. (2009). Child pornography in peer-to-peer networks. *Child Abuse & Neglect*, 33(8):560–568.

[209] Stolfo, S. J., Creamer, G., and Hershkop, S. (2006a). A temporal based forensic analysis of electronic communication. In *Proceedings of the 2006 International Conference on Digital Government Research (DG.O)*, pages 23–24. Digital Government Society of North America.

[210] Stolfo, S. J. and Hershkop, S. (2005). Email mining toolkit supporting law enforcement forensic analyses. In *Proceedings of the 2005 National Conference on Digital Government Research (DG.O)*, pages 221–222. Digital Government Society of North America.

[211] Stolfo, S. J., Hershkop, S., Hu, C.-W., Li, W.-J., Nimeskern, O., and Wang, K. (2006b). Behavior-based modeling and its application to email analysis. *ACM Transactions on Internet Technology*, 6(2):187–221.

[212] Sun, B. and Ng, V. (2011). Lifespan and popularity measurement of online content on social networks. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 379–383. IEEE.

[213] Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. (2010). Mining youtube to discover extremist videos, users and hidden communities. In *Asia Information Retrieval Symposium*, pages 13–24. Springer.

[214] Szomszor, M. N., Cantador, I., and Alani, H. (2008). Correlating user profiles from multiple folksonomies. In *Proceedings of the 19th ACM conference on Hypertext and Hypermedia*, pages 33–42. ACM.

[215] Tam, J. K. (2009). Detecting age in online chat. Master's thesis, Monterey, California Naval Postgraduate School.

[216] Teng, G.-F., Lai, M.-S., Ma, J.-B., and Li, Y. (2004). E-mail authorship mining based on svm for computer forensic. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, pages 1204–1207 vol.2. IEEE.

[217] Tian, X.-P., Geng, G.-G., and Li, H.-T. (2010). A framework for multi-features based web harmful information identification. In *Proceedings of the 2010 International Conference on Computer Application and System Modeling (ICCASM)*, pages 11:614–11:618. IEEE.

[218] Tinguriya, D. and Kumar, B. (2010). Detecting terror-related activities on the web using neural network. *Oriental Journal of Computer Science and Technology*, 3(2):331–336.

[219] Tseng, Y.-H., Ho, Z.-P., Yang, K.-S., and Chen, C.-C. (2012). Mining term networks from text collections for crime investigation. *Expert Systems with Applications*, 39(11):10082 – 10090.

[220] Uke, N. J. and Thool, R. C. (2012). Detecting pornography on web to prevent child abuse–a computer vision approach. *International Journal of Scientific and Engineering Research*, 3(4):1–3.

[221] Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., y Gómez, M. M., and Villasenor-Pineda, L. (2012). A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*.

[222] Vosoughi, S., Zhou, H., and Roy, D. (2015). Digital stylometry: Linking profiles across social networks. In *Proceedings of the 7th International Conference on Social Informatics (SocInfo)*, pages 164–177. Springer.

[223] Walgampaya, C., Kantardzic, M., and Yampolskiy, R. (2010). Real time click fraud prevention using multi-level data fusion. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1, pages 20–22. Citeseer.

[224] Wang, H., Cai, C., Philpot, A., Latonero, M., Hovy, E. H., and Metzler, D. (2012a). Data integration from open internet sources to combat sex trafficking of minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research (DG.O)*, pages 246–252. Digital Government Society of North America.

[225] Wang, K.-J., Han, X.-Z., Sun, X.-S., Chang, S.-H., and Qi, H.-F. (2006). Research on forecasting the dangerous level to illegal email based on integrated immune evolution algorithm. In *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, pages 2112–2116. IEEE.

[226] Wang, N., Jiang, K., Meier, R., and Zeng, H. (2012b). Information filtering against information pollution and crime. In *Proceedings of the International Conference on Computing, Measurement, Control and Sensor Network (CMCSN)*, pages 45–47. IEEE.

[227] Wang, X., Brown, D., and Gerber, M. (2012c). Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 36–41. IEEE.

[228] Wang, X., Gerber, M., and Brown, D. (2012d). Automatic crime prediction using events extracted from twitter posts. *Social Computing, Behavioral-Cultural Modeling and Prediction*, 7227:231–238.

[229] Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media (LSM)*, pages 19–26. Association for Computational Linguistics.

[230] Watters, P. A., Layton, R., and Dazeley, R. (2011). How much material on BitTorrent is infringing content? a case study. *Information Security Technical Report*, 16(2):79 – 87.

[231] Wei, C., Sprague, A., Warner, G., and Skjellum, A. (2008). Mining spam email to identify common origins for forensic application. In *Proceedings of the 2008 ACM Symposium on Applied computing (SAC)*, pages 1433–1437. ACM.

[232] Wenhua, L. and Na, L. (2010). Application of unstructured data processing and analyzing base on chinese in digital data evidence collecting. In *Proceedings of the 2nd International Conference on Computer Engineering and Technology (ICCET)*, pages 7:780–7:783. IEEE.

[233] Wilder, N., Smith, J. M., and Mockus, A. (2016). Exploring a framework for identity and attribute linking across heterogeneous data systems. In *Proceedings of the 2nd International Workshop on BIG Data Software Engineering (BIGDSE)*, pages 19–25. ACM.

[234] Winkler, W. E. (2015). Probabilistic linkage. In *Methodological Developments in Data Linkage*, pages 8–35. Wiley Online Library.

[235] Wondracek, G., Holz, T., Kirda, E., and Kruegel, C. (2010). A practical attack to de-anonymize social network users. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy (S&P)*, pages 223–238. IEEE.

[236] Xing, X., Liang, Y.-L., Cheng, H., Dang, J., Huang, S., Han, R., Liu, X., Lv, Q., and Mishra, S. (2011). SafeVchat: detecting obscene content and misbehaving users in online video chat services. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 685–694. ACM.

[237] Xu, J. and Chau, M. (2006). Mining communities of bloggers: A case study on cyber-hate. In *Proceedings of the 27th International Conference on Information Systems (ICIS)*, pages 135–144.

[238] Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012a). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 656–666. Association for Computational Linguistics.

[239] Xu, J.-M., Zhu, X., and Bellmore, A. (2012b). Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, pages 10:1–10:6. ACM.

[240] Yampolskiy, R. V., Klare, B., and Jain, A. K. (2012). Face recognition in the virtual world: recognizing avatar faces. In *Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA)*, pages 40–45. IEEE.

[241] Yang, C. and Ng, T. (2008). Analyzing content development and visualizing social interactions in web forum. In *Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 25–30. IEEE.

[242] Yang, C. and Ng, T. (2009). Web opinions analysis with scalable distance-based clustering. In *Proceedings of the 2009 International Conference on Intelligence and Security Informatics (ISI)*, pages 65–70. IEEE.

[243] Yang, L., Liu, F., Kizza, J., and Ege, R. (2009). Discovering topics from dark websites. In *Proceedings of the IEEE Symposium on Computational Intelligence in Cyber Security*, pages 175–179. IEEE.

[244] Yang, M., Kiang, M., Chen, H., and Li, Y. (2012). Artificial immune system for illicit content identification in social media. *Journal of the American Society for Information Science and Technology*, 63(2):256–269.

[245] Yang, M., Kiang, M., Ku, Y., Chiu, C., and Li, Y. (2011). Social media analytics for radical opinion mining in hate group web forums. *Journal of Homeland Security and Emergency Management*, 8(1).

[246] Yearwood, J., Mammadov, M., and Banerjee, A. (2010). Profiling phishing emails based on hyperlink information. In *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 120–127. IEEE.

[247] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009a). Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB 2.0 (CAW2. 0) Workshop at WWW2009*, pages 20–24. ACM.

[248] Yin, H., Hui, W., Miao, Q., Li, Z., and Lin, C. (2009b). Ivforensic: a digital forensics service platform for internet videos. In *Proceedings of the 17th ACM International Conference on Multimedia (MM)*, pages 1015–1016. ACM.

[249] Zafarani, R. and Liu, H. (2009). Connecting corresponding identities across communities. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM)*, pages 354–357. AAAI.

[250] Zhang, C., Chen, W.-B., Chen, X., and Warner, G. (2009). Revealing common sources of image spam by unsupervised clustering with visual features. In *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*, pages 891–892. ACM.

[251] Zheng, R., Li, J., Chen, H., and Huang, Z. (2005). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.

[252] Zheng, R., Qin, Y., Huang, Z., and Chen, H. (2003). Authorship analysis in cybercrime investigation. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics (ISI)*, pages 59–73. Springer-Verlag.

[253] Zhou, Y., Qin, J., Lai, G., and Chen, H. (2007). Collection of u.s. extremist online forums: A web mining approach. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS)*, pages 70–70. IEEE.

[254] Zhou, Y., Qin, J., Reid, E., Lai, G., and Chen, H. (2005a). Studying the presence of terrorism on the web: an knowledge portal approach. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 402–402. ACM.

[255] Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. (2005b). US domestic extremist groups on the web: link and content analysis. *Intelligent Systems*, 20(5):44–51.

[256] Zhu, Z. (2007). Deconstruction and analysis of email messages. Master's thesis, Florida State University.

[257] Zhuge, J., Holz, T., Song, C., Guo, J., Han, X., and Zou, W. (2009). Studying malicious websites and the underground economy on the Chinese web. In *Managing Information Risk and the Economics of Security*, pages 225–244. Springer.

[258] Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325.