Finding You on the Internet
An Approach for Finding On-line Presences of People for Fraud Risk Analysis

Henry Been[1] and Maurice van Keulen[2]
[1] *Snelstart Software, Harkebuurt 3, 1794HM Enschede, The Netherlands*
[2] *Faculty of EEMCS, University of Twente, POBox 217, 7500AE Enschede, The Netherlands*

Keywords: Fraud Detection, Web Crawling, Twitter, Entity Resolution.

Abstract: Fraud risk analysis on data from formal information sources, being a 'paper reality', suffers from blindness to false information. Moreover, the very act of providing false information is a strong indicator for fraud. The technology presented in this paper provides one step towards the vision of harnessing real-world data from social media and internet for fraud risk analysis. We introduce a novel iterative search, monitor, and match approach for finding on-line presences of people. A real-world experiment showed that Twitter accounts can be effectively found given only limited name and address data. We also present an analysis of the ethical considerations surrounding the application of such technology for fraud risk analysis.

1 INTRODUCTION

In the Netherlands, the governmental organization *Inspectie Sociale Zaken en Werkgelegenheid (ISZW)* (Inspectorate for the Ministry of Social Affairs and Employment) is responsible for "detection of fraud, exploitation and organised crime within the chain of work and income (labour exploitation, human trafficking and large scale fraud in the area of social security)".[1] The ISZW is facing both budget cuts and demand for more fraud detection[2] To accommodate this, they have been changing their work process to be risk-driven, which is believed to increase efficiency. This means that analysts of the ISZW first classify subjects into categories of low through high risk of fraud. Resources are then distributed accordingly.

Although in theory a sound approach, results in practice are not as good as the ISZW believes they can be. One major cause is that the ISZW uses only formal information sources available within the Ministry and other ministries. Especially in fraudulent cases, people are inclined to provide false information, hence the formal information sources can be considered to reflect a 'paper reality'. A risk analysis performed based on this data, assesses risk in the paper reality, not in the real world. Hence it will miss concealed fraudulent cases.

To overcome problems like these, the ISZW is looking for other sources for predictive characteristics of subjects that it can use to discover fraud. One successful experiment involved water usage (Inspectie SZW, 2012). It is known that water usage at a certain address correlates strongly with the number of persons living at that address. By gathering water usage data, discrepancies can be detected between the number of people registered to be living at an address and the predicted number of people based on the water usage. The discrepancy is an indicator for, for example, someone falsifying his/her address to meet the requirements for receiving welfare support.

Unfortunately, water usage can only be used to predict certain types of fraud, like welfare or unemployment benefits, which involve household characteristics. Other types of fraud, like undeclared capital or income cannot be related to falsified addresses. Nevertheless, the experiment showed that involving 'data from the real world' allows for indicators for falsification and attempts at concealment, and that such indicators are strong predictors for high risk of fraud.

One direction the ISZW is exploring, is the use of social media. People voluntarily share a lot of information about themselves online. Furthermore, research has shown that, contrary to popular belief, online profiles do not depict who we want to be, but who we are (Back et al., 2010). For example, some-

[1] http://www.inspectieszw.nl/english

[2] http://www.rijksoverheid.nl/documenten-en-publicaties/persberichten/2011/03/14/kleinere-en-efficientere-overheid-bij-szw-uwv-en-svb-bespaart-410-miljoen.html.

Been H. and van Keulen M..
Finding You on the Internet - An Approach for Finding On-line Presences of People for Fraud Risk Analysis.
DOI: 10.5220/0004985006970706
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ISS-2014), pages 697-706
ISBN: 978-989-758-028-4

one who Twitters about bikes he repairs, buys or sells might have undeclared income. Or someone who is posting Facebook updates from all over Europe, is apparently in possession of sufficient funds for traveling, hence likely of undeclared capital or income.

Note that possibly incriminating information is not only disclosed because of carelessness, but also because many on-line activities unavoidably leave traces. For example, if someone raises a substantial income by buying and selling items on an online auction website, the website unavoidably contains all data necessary to calculate this income, hence for checking whether or not that income has been declared and whether or not that income violates a requirement for receiving welfare support.

If data from social media and other websites could be harnessed, many real-world fraud predictors might be derived. However, before the ISZW is able to harness information from social media, they need to be able to identify which online presences belong to a certain person. With *online presence* we mean any account on some social media or other website, such as Twitter, Facebook, eBay or some blog. The ISZW does not have much information in its information sources usable for this identification: typically only name, address, phone number(s), and possibly an email address.

This paper focusses on this particular problem: Given only limited details about a person, can one automatically determine his/her online presences with sufficient certainty? In essence an entity resolution problem (Brizan and Tansel, 2006) on internet-scale. In our experiments, we use a Twitter account as a representative examples of an online presence. There are two important characteristics of working with internet data that make such a task difficult: First, the vast amount of Twitter accounts (around 900 million at the time of the experiment) makes a side-by-side comparison of details of a person and a Twitter account infeasible. Secondly, the influence of *noise*, i.e., nicknames, duplicates, false and incomplete accounts, inflicts a significant performance penalty.

Contributions. This paper presents (a) a novel iterative search, monitor, and match approach for finding on-line presences of people given only limited name, address data; (b) a real-world experiment with two subject groups of 22 (voluntary sign-up) and 85 subjects (from ISZW), respectively, where our IMatcher prototype gathered candidate accounts for extended periods of time. Although our initial attempt at pinpointing the correct account for each subject proved ineffective, we showed that in almost all cases, the correct account was among the candidates. And (c) an analysis of the ethical considerations surrounding the application of such technology for fraud risk analysis. In this way, the paper provides one step towards the vision of harnessing real-world data from social media and internet for fraud risk analysis.

Outlook. The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the approach on a conceptual level. Section 4 presents some technical aspects of realizing the approach in practice. Section 5 describes the performed experiments, which are discussed in Section 6. Section 7 discusses some of the ethical concerns surrounding both the experiments and the application of this approach for fraud detection. Finally Section 8 concludes the paper.

2 RELATED WORK

The task of finding a person's Twitter account can be seen as an entity resolution problem (Brizan and Tansel, 2006). Many approaches from this field cannot be applied, however, because one of the sources is not directly accessible: the database behind twitter.com. The task can also be described as an entity extraction and disambiguation problem. Often, approaches in this field use a knowledge base, such as Aida (Yosef et al., 2011) which uses Yago (Suchanek et al., 2007). These are also not applicable here, because the sought entities are not famous, hence do not appear in any knowledge base. An open world approach such as (Habib and van Keulen, 2013) comes closer, but as opposed to this work, in the end we are not interested in any homepage for the entity, but in a specific one: the person's Twitter account. Another difference of this work with many others, is that our approach is incremental: new candidates and/or attributes are discovered and evaluated continuously.

Linking social media profiles is a related problem. Veldman (Veldman, 2009) investigated the value of the connection network for linking the Dutch social network site Hyves with LinkedIn. Narayanan and Shmatikov (Narayanan and Shmatikov, 2009) used a known, labeled network surrounding a user to deanonymize an anonymous network surrounding the same user. Both confirm that the network alone is not discriminative enough, but including the network in an attribute-based matching improves performance. Nemo (Jain and Kumaraguru, 2012) uses three dimensions: profile details, user generated content, and connection network. Starting with a social media profile provides more information for search than in our application. Therefore, one clue they found most effective cannot be exploited: self-mentioning, where someone tweets a reference to his/her Facebook page.

Other approaches towards using online profiles to gather data about individuals have surfaced. An example is RIOT[3]. This system finds people online, tracks their past activities and makes predictions about future movements. RIOT is semi-automatic in its disambiguation.

Other forms of input have been investigated. (Perito et al., 2011) use usernames from Google and EBay services to find accounts belonging to the same user. Face recognition can also be used as input (Minder and Bernstein, 2011). Although in itself it does not perform better than text-based approaches, they do show that combining both dimensions increases accuracy.

Note that many approaches mentioned here work with pre-fetched databases, not directly on internet-scale data as we did.

3 CONCEPTUAL APPROACH

When looking for something, humans often take a longlist/shortlist approach: using a set of heuristic queries they quickly gather a set of possible answers to their search: the *longlist*. Each possible answer undergoes a brief examination to select the few answers that (s)he deems most likely to be correct: the *shortlist*. The answers on the shortlist undergo a more thorough examination for determining the final answer.

To automatically find online presences of a person, we take a similar, but automated approach. First, a set of possibly matching online presences are gathered. Together they form the *longlist*. The challenge is to ensure that the longlist includes the correct ones while at the same time keeping the length of the list within practical bounds. Every entry of the longlist is examined to gather some characteristics about it. Comparing these characteristics with the subject of the initial search yields a similarity score. Selecting only the entries with the highest similarity scores yields the *shortlist*, a list of likely correct results.

We deviate from the human longlist/shortlist approach in two ways:

- Besides a shortlist of likely correct results, also a list of highly likely incorrect results is determined (the *exclusion list*).
- The process is iterative: the longlist, shortlist and exclusion list are updated and improved at each iteration with the aim that the shortlist converges to the correct result.

3.1 Formalization of Approach

We assume that a set of persons P is given. For each person $p \in P$, data on a number of attributes $a \in A$ is also given, denoted as $p.a$. In our experiments, $A = \{\mathsf{firstname}, \mathsf{lastname}, \mathsf{address}, \mathsf{email}, \mathsf{telephone}\}$.

Let T be the set of all Twitter accounts. The goal of our approach is to identify for all $p \in P$, the Twitter account of this person, denoted as $\tilde{t}_p$.

Note that it is infeasible to iterate over all accounts $t \in T$ as there were around 900 million at the time of the experiment. Therefore, we follow a three-step strategy. Executing the three steps once is called an *iteration* denoted with I. A set of consecutive iterations is call a *run* denoted with $r = \{I_1, \ldots, I_N\}$. Each subsequent iteration I_n produces, for each $p \in P$, an improved *candidate set* or *longlist* C_p^n of accounts possibly belonging to p and an *exclusion set* E_p^n of accounts that have been considered but found to be not belonging to p with sufficient certainty.

Step 1. The aim of the first step is to construct a longlist of candidates while ensuring that it includes the correct one if there is one. The first step starts with a search for possibly matching Twitter accounts using a number of queries Q. Each query $q \in Q$ results in a set of Twitter accounts $q(p) = T'$.

Queries we used in our experiments were, for example, Google queries like '$p.\mathsf{firstname}$ $p.\mathsf{lastname}$ `twitter`' from which we extracted Twitter accounts from the top-8 (later top-20) results, or spatial queries on Twitter producing all tweets sent within a 200 meter radius around the coordinates of $p.\mathsf{address}$ from which we extracted the accounts of the users who sent them.

The union of all query results is the candidate set of that iteration: $c_p^n = \bigcup_{q \in Q} q(p)$. The found candidates are accumulated. Let $C_p^n = (C_p^{n-1} \cup c_p^n) \backslash E_p^{n-1}$ be the set of candidates after iteration I_n. Obviously, the start situation is $C_p^0 = \emptyset$ and $E_p^0 = \emptyset$.

The set of queries are constructed in such a way that they achieve a high likelihood that the correct result is in the candidate set, i.e., $0 \ll P(\tilde{t}_p \in c_p^n) \leq 1$ even though the queries can only produce a very limited result set $|q(p)| \ll |T|$. Sections 4 and 5.1 provide more details on the exact construction of the queries.

Step 2. The goal of the second step is to gather more details about the candidates on the longlist, including by assumption, about the person who owns the account. Then, for each candidate of a person, we determine how similar it is to the data we have on that person.

[3] http://www.guardian.co.uk/world/2013/feb/10/software-tracks-social-media-defence

We use a set of similarity functions S. Each $s \in S$ determines a similarity score $s(p,t) \in [0,1]$ based on certain criteria. The overall similarity for all $p \in P$ and for all $t \in C_p^n$ is computed as $os(p,t) = \frac{1}{W}\sum_{s \in S} w_s s(p,t)$ where w_s is the weight of similarity function s and $W = \sum_{s \in S} w_s$.

Step 3. Along with (hopefully) the correct result $\tilde{t}_p$, there are obviously many incorrect results in C_p^n called *false positives* $F_p^n = C_p^n \setminus \{\tilde{t}_p\}$.

The goal of the third step, is to mark those accounts as false positives for which sufficient certainty has been obtained that they do not belong to p, i.e., which accounts to add to the exclusion set. Three thresholds are in play here:

- Let β be the number of iterations during which no accounts are excluded.
- Let α_1 be the similarity score below which accounts are excluded.
- Let α_2 be the similarity score above which accounts are considered plausible.

Therefore, $E_p^n = \begin{cases} \emptyset & \text{if } n \leq \beta \\ \{t \in C_p^n | os(p,t) \leq \alpha_1\} & \text{otherwise} \end{cases}$

From the remaining candidates, we determine a set of *most plausible candidates* or *shortlist* $\tilde{C}_p^n = \{t \in C_p^n | os(p,t) \geq \alpha_2\}$

Convergence. The expectation is that with a growing number of iterations, the steps converge to the following two desirable situations: $\tilde{C}_p^n = \{\tilde{t}_p\}$ in case p has a Twitter account in reality, or $\tilde{C}_p^n = \emptyset$ in case (s)he hasn't. There are two important reasons that strengthen this expectation.

- Query results may sometimes include $\tilde{t}_p$, but also many times not. Repeating the queries over time, increases the likelihood that $\tilde{t}_p$ is encountered eventually. Just once does a person need to write a tweet sufficiently close to the known address or sufficiently popular to appear in the top-20 Google result, for the system to pick it up.
- By accumulating the data gathered in step 2, the similarity functions have more data to go one, hence can be expected to gain in accuracy.

For these two reasons, it is expected that the correct Twitter account will ‘surface’ eventually.

Note that for our application, it is not absolutely necessary that the desirable situation is actually achieved. A shortlist with several possible candidates is usable as input to the risk analysis process. The latter is statistical in nature, hence it is straightforward to deal with more than one candidate as long as each has an adequate probability estimate (van Keulen, 2012).

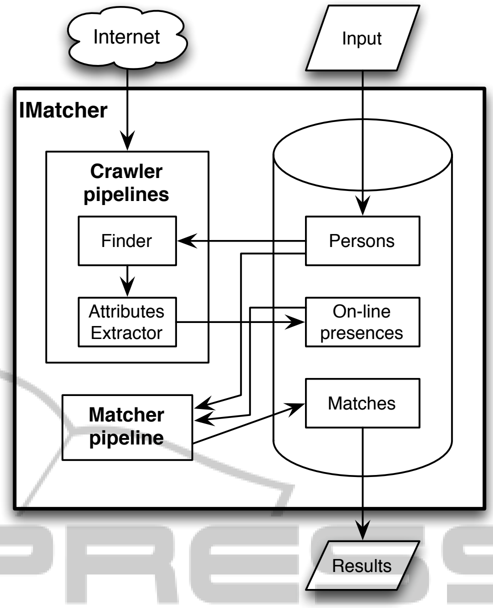


Figure 1: Overview of the IMatcher system.

4 TECHNICAL ASPECTS

Overview Architecture. The system implementing the conceptual approach of Section 3 is called *IMatcher*. Figure 1 presents the architectural overview. The core of IMatcher is an XML database storing all data: input data on the persons to be found as well as data on the candidate Twitter accounts including user profile data, photo, collected tweets, and any email addresses and phone numbers found in the tweets.

The two other main components are the crawler and the matcher. The crawler is responsible for all Internet access: executing the queries, retrieving resulting web pages, and extracting attributes and accounts from those pages, as well as Twitter access: for all found accounts, retrieving profile data, collect all tweets, and extract information from the tweet texts. The matcher is responsible for a multi-criteria comparison between the input person data with the data collected on the found accounts. Both components use a pipeline architecture (also known as pipe-and-filter) (Hofmann et al., 1996) for easy scheduling and monitoring.

Crawler Pipelines. The pipelines for the crawler are presented in Figure 2.

The inflow of the first pipeline are the persons. The first few sinks execute the queries in an attempt to find Twitter accounts. Thus found accounts pro-

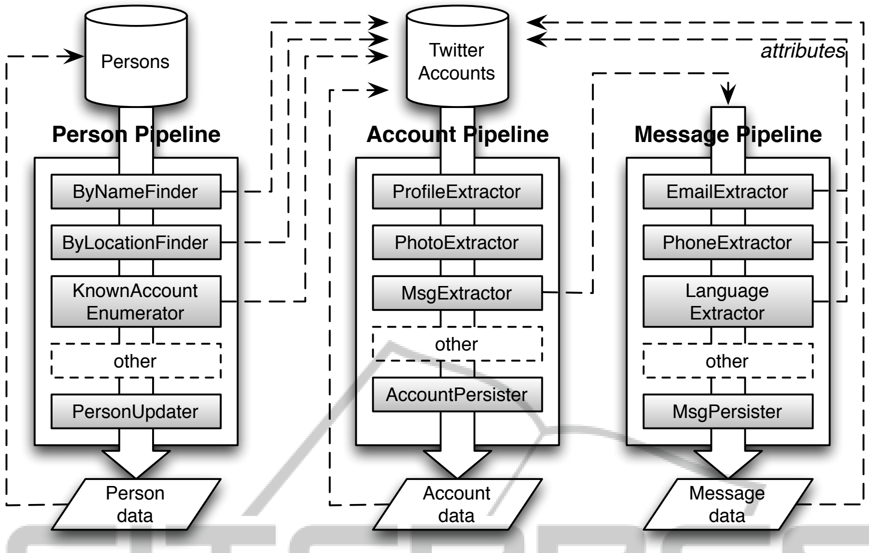


Figure 2: The IMatcher crawler pipelines.

vide the inflow for the second pipeline. To complete the inflow, the 'KnownTwitterEnumerator' also sends the candidate accounts from previous iterations. Any newly found accounts are stores with the person by the 'PersonUpdater'.

The second pipeline is responsible for extracting information about the accounts. Profile information and a photo are extracted at each iteration to ensure that changes are caught. The pipeline also contains a 'TwitterAccountPruner' (not shown) that filters accounts in the exclusion list thereby avoiding that they get processed. Furthermore, from the moment an account becomes a candidate for some person, its tweets are collected.

Those tweets are the inflow for the third pipeline responsible for extracting information from the tweets, such as the language, email addresses, phone numbers, etc. This information is added to the stored account data as it provides valuable clues for matching.

Note that although only Twitter-related sinks are shown, the architecture is extensible with more sinks and pipelines exploring other data sources, such as Facebook or e-bay.

Matcher Pipeline. The matcher pipeline is rather straightforward. Its inflow consists of all combinations of a person and one of its candidate accounts. Each sink computes a different similarity score based on some specific criterion. A final sink aggregates the individual scores to one overall score. All scores are stored.

Input Data. The IMatcher requires that for each person an unique number, e.g., the social security number is provided. This number is not used in searching or matching and only serves to uniquely identify the person. Furthermore, the following can be optionally provided for better search and matching:

- Firstname(s)
- Lastname
- 'Tussenvoegsel' (if any)[4]
- Any addresses known for the person, such as home and work address. If unknown or not known precisely, the information can be provided on street, city, or country level.
- Any phone numbers known for the person.
- Any email addresses known for the person.
- Any aliases known for the person, such as nicknames, often used online pseudonyms or avatar names in online games.

Request Limits and Proxies. Many on-line services impose strict limits on the usage of their APIs. IMatcher uses Google search and the Twitter API. Twitter limits the number of requests per hour. Google search not only has a limit on the number of requests, but also requires that requests are not sent in

[4] A 'tussenvoegsels' is a typical Dutch phenomenon. It is a prefix in the lastname. For example, in "Jan van der Sloot", "Jan" is the first name, "van der Sloot" the lastname and "van der" the 'tussenvoegsel'. The name would be alphabetically ordered under 'S' of "Sloot". In other countries, the 'tussenvoegsels' are often contracted, e.g., "Jan vanderSloot".

quick succession in an attempt to only allow search by humans.

In the envisioned production setting, IMatcher is expected to run within the Dutch iRN infrastructure[5] already available for the Dutch national police to research and investigation of the internet in a safe and forensically secure way. Since iRN already provides mechanisms for dealing with request limits, we did chose to use a simple intermediary solution based on proxies for carrying out the experiments. A proxy is a service that forwards requests, hence mimicking a different origin of the request with its own limit. In the implementation, we re-used code from the "Neogeo" project[6], which is a research prototype from the University of Twente that allows for (almost) automatic crawling. Its automated crawling and information extraction functionality is not used; only its lower-level functionality for handling proxies and requesting, parsing and storing web pages.

For more details on IMatcher, we refer to (Been, 2013).

5 EXPERIMENTS

We identify four goals in our experiments investigating how certain factors would influence the resulting longlist. After these experiments, the gathered data was analyzed to see how well it could be used to pinpoint the correct account or at least identify a shortlist of most likely matches.

The first goal is to establish a relation between the number of iterations n and the average setsize $L_n = \frac{1}{|P|}\sum_{p \in P}|C_p^n|$. This investigates the feasibility of the chosen approach. The average setsize was expected to increase with each iteration, but also that it would stabilize and, over time, even reduce with a growing exclusion list E_p^n.

The second goal is to investigate the influence of the input attributes A on the inclusion of correct accounts $I_n = \left|\{t \in C_p^n \mid \exists p \in P : t = \tilde{t}_p\}\right|$. It was expected that not using all the input attributes available, but leaving out, for example, the lastname, would negatively influence inclusion. Reversing this result would show how inclusion would benefit from adding a specific input parameter.

The third goal is to investigate the relation between the number of iterations n and inclusion I_n. It was expected that inclusion, i.e., the number of discovered correct accounts, would rise during the first few iterations, and then stabilize at some number with only an incidental increment.

Finally, the fourth goal is to study the influence of broader search criteria on the setsize L_n. Broader search queries produce longer longlists, but are also expected to influence inclusion favorably.

5.1 Experimental Setup

Subject Selection. Two different datasets were available for experimentation. The first dataset, called 'sign-up', consisted of 22 subjects, 16 male and 6 female. All subjects were self-selected and signed up under full informed consent. It was explained that the goal of the experiment was to discover their Twitter account using an undirected exploratory approach and that it was difficult to predict the results and what kind of form they would take. Of these subjects, 12 had a Twitter account and provided the name of that account. Therefore, for this dataset the ground truth is known, hence inclusion could be measured.

The second dataset, called 'ISZW', was provided by the ISZW and consisted of 85 subjects that were picked from a real risk-analysis project. It was claimed by the ISZW that the subjects were a representative sample of the group currently under investigation by the ISZW. These subjects did not sign up voluntarily and were (and are) not informed that they were part of an experiment. A contract was signed that defined the authors as responsible for carrying out a pilot for the ISZW and specified the conditions they had to uphold. This made it both legal as well as moral to make this data available to the authors.

Execution. A number of runs were done with the IMatcher to gather all data needed. Due to the dependence on external services and data-intensive tasks, the runs could take a lot of time (48 hours for the longest). Therefore, great care was taken to select a minimum number of runs to facilitate all goals. The first four runs were done on the subjects of the sign-up dataset. One run, including all information available about the subjects, had 15 iterations. The other three runs, where some input attributes were left out, were meant to run five times. Only the last iteration of run 4 failed, due to an unfortunate reaching of a request limit. A fith run was performed on the subjects from the ISZW dataset.

All runs were repeated with an altered version of the IMatcher that had a broader search for building the longlist (see 'Queries' below). Due to result sizes and data volumes involved, each run consisted of only one iteration. Again, there is a missing iteration for the final run, run 9, with the sign-up dataset.

[5] http://columbo.nl

[6] https://github.com/utwente-db/neogeo

Table 1: Runs performed with the IMatcher.

Run	**data set**	n	**Input attributes** A
1	sign-up	15	all
2	sign-up	5	all except lastname
3	sign-up	5	all except address
4	sign-up	5	all except e-mail & telephone
5	ISZW	3	all
6	sign-up	1	all
7	sign-up	1	all except lastname
8	sign-up	1	all except address
9	sign-up	1	all except e-mail & telephone
10	ISZW	1	all

Table 1 describes all runs, the dataset they used, the number of iterations, and which input attributes were used.

The reasons for choosing these 10 runs are as follows. The results of the first four runs were used for the first three goals. Comparing the average setsize and inclusion of the different runs, the influence of input attributes can be established. Also, each run provides data related to the relation between the number of iterations and setsize and inclusion. These measurements are meant to be interpreted in relation to each other, to see the value of each input attribute. Run 1, which had the most input available and was thus expected to perform best, was continued longer than the other runs to establish a better absolute projection of the results. Runs 6 to 10 were run for the fourth goal to discover the influence of search queries on inclusion.

To make the results of certain iterations as comparable as possible, they should run against an Internet in more-or-less the same state. Therefore, all iterations i of a run were started as close to each other in time as possible. As a consequence, there was a gap of about 16–25 hours between iterations of a run, depending on the moment of completion of other runs. After each iteration a snapshot was made of the database, so metrics could be calculated afterwards.

Queries. For all input attributes except address, keyword sets were constructed: each attribute individually as well as all combinations of firstname(s) and lastname(s). With each such keyword set, Google search queries were constructed in two ways.

- **direct**: the query is equal to the keyword set.
- **guided**: two queries are constructed as "Twitter *keyword set*" and "*keyword set* site:twitter.com".

Twitter accounts were extracted from a top-k of Google results: either directly if a result refers to a Twitter profile, or from the page the result refers to by means of the regular expressions `@[a-zA-Z_]*` and `(http(s)?://)?(www.)?twitter.com/[a-zA-Z_]`. For address, a geographical search was performed with a radius of r meters using the now retired Twitter API V1[7].

For runs 1–5, IMatcher was configured with search criteria **direct**, $k = 8$, and $r = 200$. For runs 6–10, search criteria were **guided**, $k = 20$, $r = 200$.

5.2 Results

Experimental results for average setsize L_n and inclusion I_n are shown in Figures 3 and 4, respectively. Each subfigure shows a pair of runs that use the same input parameters.

It can be seen that specific input attributes do influence the results. Leaving out lastname shrinks the average setsize by roughly 50%, but also reduces inclusion 7 → 2 and 11 → 3, respectively. Leaving out address reduces the average setsize by roughly 33%, but has no influence on inclusion. Leaving out phone numbers and e-mail adresses does not seem to influence average setsize nor inclusion.

The convergence expectations are not confirmed by our experiments. The average setsize did not stabilize and inclusion did not increase with more iterations.

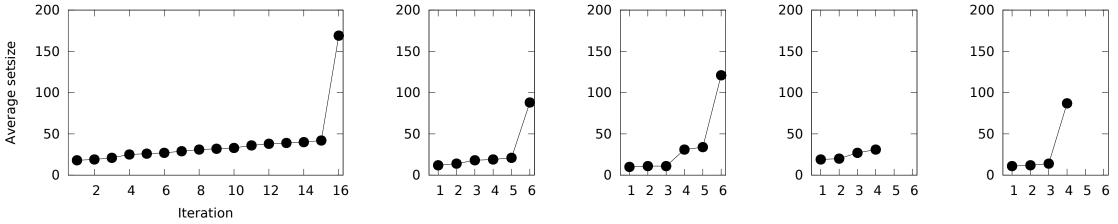
Pinpointing $\tilde{t}_p$ among C^n_p. Close to 37k rows of raw feature data was extracted from the gathered data. It consists of the person, the twitter candidate, a number of similarity scores, the run, and the iteration number. An attempt was made to predict whether the person has a Twitter account, as well as whether a candidate is the correct one. This relates to step 3 of our approach. Unfortunately, this attempt proved too ineffective for it to be useful to present here.

Note that this analysis could only be performed for the sign-up subject group, since no ground truth was known for the ISZW dataset.

6 DISCUSSION

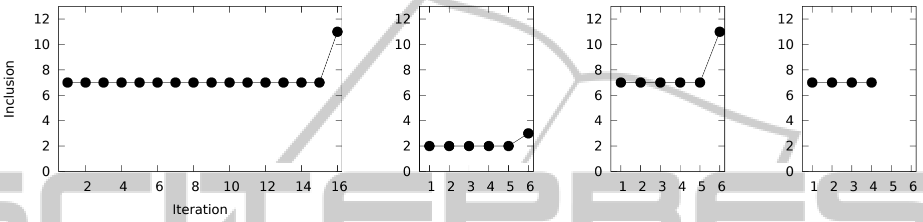
From the experimental results, we can conclude a number of things. First of all, the convergence expectations were not confirmed. Therefore, the amount of data to be managed is larger and will grow longer than the lengths of the runs we experimented with (15 being the longest). than expected. However, we still expect that, after even more iterations, the pruner will prune enough accounts to keep the candidate set size constant or even decrease it over time.

[7] https://blog.twitter.com/2013/api-v1-retirement-final-dates



(a) Run 1 & 6 (b) Run 2 & 7 (c) Run 3 & 8 (d) Run 4 & 9 (e) Run 5 & 10

Figure 3: Average setsize L_n after n iterations. Runs 6–10 are depicted in the graphs as the last iteration ($n = 16$, 6, 6, 6, and 4, respectively). Last iteration of run 4 and 9 failed.



(a) Runs 1 & 6 (b) Runs 2 & 7 (c) Runs 3 & 8 (d) Runs 4 & 9

Figure 4: Inclusion I_n, i.e., the number of correct Twitter accounts included (out of 12) after n iterations. Since there is no ground truth known for the ISZW dataset, there is no graph for runs 5&10. Again, Runs 6–9 are depicted in the graphs as the last iteration ($n = 16$, 6, 6, and 6, respectively). Last iteration of run 4 and 9 failed.

Secondly, the approach is sensitive to certain input attributes. The full name needs to be available, since leaving it out will decrease inclusion dramatically. E-mail addresses and phone numbers do not seem to influence effectiveness.

Thirdly, there is no rise in inclusion after the first iteration. This is unexpected and most likely due to the small scale of our experiments and the open nature of our volunteers. Fraudulent people are expected to be less open, hence their account is not expected to be found with a first search. We still believe that increasing the number of subjects and the number of iterations will show that inclusion will slowly converge to the number of subjects with a Twitter account.

Finally, broadening the search queries Q did have an impact on both setsize and inclusion. Research is needed to investigate this relation in more detail. In practice, an iterative deepening approach might prove powerful in addressing the fact that candidates accounts can be ranked at any position in query results. Increasing the top-k of query results that is explored with each iteration if no definite candidate is found yet, might prove a good strategy.

Future Work. As stated above, we were not able to perform step 3 of our approach satisfactorily. We suspect that the problem lies in the chosen features we extracted from the accounts and our similarity functions. As we have shown the viability of a search, monitor and match approach on internet-scale, we aim to improve the second and third step in future work. We intend to focus on retrieving more characteristics of the online presences we find and comparing them to known characteristics. We expect that including more characteristics like usernames, photographs, language usage and locations mentioned in text will improve our matching algorithms and thus allow us to pinpoint the one correct account or at least a group of 2 or 3 most likely candidates.

We are also interested in extending our research beyond Twitter to other online presences, like Facebook, E-Bay and even homepages in general. Gathering more online presences will allow for better characteristic extraction and thus for better fraud prediction. It will also enable the exploitation of self-mentioning.

7 ETHICAL CONSIDERATIONS

When undertaking research such as ours, it is important to consider its morality. We co-operated with an ethicist to see if our work is ethically justifiable. We were confident that our experiment was, since we followed university guidelines for subject selection and experiment design. However, the context necessitates considering its (possible) uses to assess the desirability of pursuing the research.

Aime van Wynsberghe (Wynsberghe et al., 2013) used our work to develop a set of generic guidelines for working with data from social network sites. Together we published a paper describing the guidelines and applying them to our research.

Using these guidelines and elements from value sensitive design (Friedman et al., 2006), we concluded that our research was ethically justifiable as a value trade-off taking into account the interests of people investigated, people who's account is included in a candidate set as false positive, the ISZW and all Dutch citizens. One important factor was also that, although requesting welfare support is not really by choice, the receiver is not obliged to do so. By requesting welfare support, someone voluntarily gives up some privacy to allow the government to investigate if he rightfully does so. This aspect also shows that using our design to investigate other groups has to be considered for its own merits. For more details, we refer to (Been, 2013).

8 CONCLUSIONS

Fraud risk analysis on data from formal information sources, being a paper reality, suffers from blindness to false information. Moreover, the very act of providing false information is a strong indicator for fraud. As a step towards the vision of harnessing real-world data from social media and internet for fraud risk analysis, we present a novel iterative search, monitor, and match approach for finding on-line presences of people. The approach needs only limited name/address input data available to governmental organizations responsible for fraud detection. A real-world experiment showed that Twitter accounts can be effectively found: from a voluntary sign-up subject group of 22 subjects, the correct account was almost always captured. Our initial attempt at pinpointing the correct account for each subject proved ineffective, but we expect this to be a matter of choosing other features and classification techniques, since the correct account is included and rich data is gathered. We also experimented with a larger subject group of 85 subjects from the ISZW. Finally, an analysis is given of the ethics surrounding the application of such technology for fraud risk analysis. We aim to extend IMatcher to search for more kinds of on-line presences such as other social networks, extract and monitor more characteristics, and improve the person vs. on-line presence matching.

ACKNOWLEDGMENTS

This publication was supported by the Dutch national program COMMIT/.We thank Victor de Graaff for neogeo support and ISZW for providing the data.

REFERENCES

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., and Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3):372–374.

Been, H. (2013). Finding you on the internet: Entity resolution on twitter accounts and real world people. Master's thesis, Univ. Twente, Netherlands.

Brizan, D. and Tansel, A. (2006). A survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 6(3):41–50.

Friedman, B., Kahn Jr, P. H., and Borning, A. (2006). Value sensitive design and information systems. *Human-computer interaction in management information systems: Foundations*, 5:348–372.

Habib, M. B. and van Keulen, M. (2013). A generic open world named entity disambiguation approach for tweets. In *KDIR 2013, Vilamoura, Portugal*. SciTePress.

Hofmann, C., Horn, E., Keller, W., Renzel, K., and Schmidt, M. (1996). The field of software architecture. Technical Report TUM-I9641, TU Munich, Germany.

Inspectie SZW (2012). Bestandskoppeling bij fraudebestrijding. Technical Report Nvb-Info 12/062, Ministerie van Sociale Zaken en Werkgelegenheid.

Jain, P. and Kumaraguru, P. (2012). Finding nemo: Searching and resolving identities of users across online social networks. Technical Report arXiv:1212.6147, arXiv.

Minder, P. and Bernstein, A. (2011). Social network aggregation using face-recognition. In *SDoW 2011, Bonn, Germany*, volume 830. CEUR. ISSN 1613-0073.

Narayanan, A. and Shmatikov, V. (2009). De-anonymizing social networks. In *30th IEEE Symposium on Security and Privacy*, pages 173–187.

Perito, D., Castelluccia, C., Kaafar, M. A., and Manils, P. (2011). How unique and traceable are usernames? In *Privacy Enhancing Technologies*, pages 1–17. Springer.

Suchanek, F., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *WWW 2007, Banff, Canada*, pages 697–706.

van Keulen, M. (2012). Managing uncertainty: The road towards better data interoperability. *IT - Information Technology*, 54(3):138–146.

Veldman, I. (2009). Matching profiles from social network sites. Master's thesis, Univ. Twente, Netherlands.

Wynsberghe, A., Been, H., and Keulen, M. (2013). To use or not to use: guidelines for researchers using data from online social networking sites. *Responsible Research and Innovation Observatory*.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). AIDA: an online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4(12):1450–1453.