Privacy Risks from Public Data Sources

Zacharias Tzermias, Vassilis Prevelakis, Sotiris Ioannidis

Privacy Risks from Public Data Sources

Zacharias Tzermias[1], Vassilis Prevelakis[2], and Sotiris Ioannidis[1]

[1] Institute of Computer Science,
Foundation for Research and Technology - Hellas (FORTH)
{tzermias,sotiris}@ics.forth.gr,
[2] Institute of Computer and Network Engineering,
Technical University Carolo-Wilhelmina zu Braunschweig
prevelakis@ida.ing.tu-bs.de

Abstract. In the fight against tax evaders and other cheats, governments seek to gather more information about their citizens. In this paper we claim that this increased transparency, combined with ineptitude, or corruption, can lead to widespread violations of privacy, ultimately harming law-abiding individuals while helping those engaged in criminal activities such as stalking, identity theft and so on.
In this paper we survey a number of data sources administrerd by the Greek state, offered as web services, to investigate whether they can lead to leakage of sensitive information. Our study shows that we were able to download significant portions of the data stored in some of these data sources (scraping). Moreover, for those datasources that were not ammenable to scraping we looked at ways of extracting information for specific individuals that we had identified by looking at other data sources. The vulnerabilities we have discovered enable the collection of personal data and, thus, open the way for a variety of impersonation attacks, identity theft, confidence trickster attacks and so on. We believe that the lack of a big picture which was caused by the piecemeal development of these datasources hides the true extent of the threat. Hence, by looking at all these data sources together, we outline a number of mitigation strategies that can alleviate some of the most obvious attack strategies. Finally, we look at measures that can be taken in the longer term to safeguard the privacy of the citizens.

1 Introduction

The increasing computerization of government departments and state organizations often places in jeopardy the confidentiality of the private information of the citizens. Moreover, there exists an additional incentive in collecting as much data, financial or otherwise, into central databases and improving the interconnection of existing databases, namely that the information from all these sources can be combined to discover illegal activities. Examples of this holistic approach abound: the US Drug Enforcement Agency checks electricity bills because high electricity consumption without an obvious reason, may indicate an underground marijuana plantation [19]. The Greek government wants to gain

access to water utility bills to discover unregistered swimming pools (in Greece, swimming pools above a certain size are levied a special wealth tax, as well as increasing the minimum taxable income that the owner must declare).

Nevertheless, these actions run counter to basic privacy principles (which, by the way, the private sector is expected to follow), and may lead to widespread leakage of information that can then be used for fraud or other illegal activities. For example, the widespread use in the US of the social security number as a unique identification number has been implicated in numerous cases of fraud, such as identity theft [3]

Another example would be the case where the state requires that citizens submit inventories of valuable articles in their possession (e.g. paintings, works of art, electronic equipment etc.) together with the physical address, name and telephone number of the owner. If such lists are ever leaked, they will constitute a guide to potential thieves or other scammers that may use this information to defraud the owners.

In this paper we look at the information stored in various repositories and look at how the unintended use of the information, or services offered by these repositories can affect the privacy of the citizens. We will use the Greek state as an example of the case where the lack of commonly applied data protection rules and the casual disregard for such rules when they exist, allows infringement of privacy in a massive scale. Miscreants, can extort valuable private information from the aforementioned repositories, that can be used in targeted attacks.

The rest of the paper is organized as follows. In Section 2 we survey a wide number of data sources operated by the Greek state describing their purpose and what data is available from them. Section 3 presents our crawling methods and infrastructure that we used towards the aforementioned data sources. We present the evaluation of our approach in Section 4. In Section 5 we propose countermeasures than can be incorporated to repulse future attacks. Related work is presented in Section 6. Finally we conclude in Section 7.

2 Public Datasources

The Greek government has offered a variety of data to the public, either to accommodate taxpayers' need to access their own personal information, or in terms of transparency. The most prominent example is Diavgeia [6], a government repository where every public entity is obliged to upload its decisions. This section depicts every public datasource that contains sensitive information about individuals as well as enumerates a variety of identification numbers used by the Greek government for various transactions with citizens.

2.1 The Greek Tax Registration Number

Greek Tax Registration Number (in Greek,"ΑΦΜ"), is a unique number provided by the Greek Ministry of Finance to every person or legal entity exhibiting

[3] `http://www.consumer.ftc.gov/articles/0271-signs-identity-theft`

financial activity in Greece. Every transaction with public sector services, like tax offices, require this number among others. The ubiquitous nature of Greek Tax Registration Number (TRN), made other institutions, such as banks or insurance services to incorporate it in their transactions as an additional means of identification. The Greek TRN is a non-sequential number consisting of 9 digits and generated by an algorithm. This algorithm has been published and used in forms of state websites to detect mistyped TRNs.

Greek Tax Registration Numbers are separated in two main categories. The former includes numbers issued to wage earners and pensioners, while the latter includes freelancers, legal entities, organizations, institutions and businesses. In an effort to aid transactions among entities belonging in the second category, the General Secretariat of Information Systems (GSIS) [4] responsible of computerization of the public sector, introduced a web service that given an TRN that belongs to the second category, provides an abbreviated version of the record of the entity which matches that TRN. The information provided includes name and commercial title of business, location of headquarters, telephone number as well as the type of business activity. [11]. If the submitted number does not belong in the second category, the system returns an error which indicates the reason for the refusal, such as unregistered Tax Registration Number, number belongs to a wage earner or pensioner (and hence information about that TRN cannot be released), and so on. This information, as we shall see later, can be used to determine which TRNs to look for in other on-line datasources

Generally self-employed people and single-person businesses (e.g. farmers, plumbers, electricians and many other categories) tend to submit their residential addresses and phone numbers as their contact information, during the TRN registration procedure. Some even submit mobile telephone numbers. Consequently, the TRN search web service can inadvertently expose potentially private data to the public.

2.2 Greek Identification Card Number

Along with the Tax Registration Number, every Greek citizen is issued with an Identification (ID) card. It holds information about its owner, namely first name, last name, father's name, date of birth, as well as a unique Identification Card Number. Each Identification Card Number (ID) consists of 1 or 2 Greek capital letters and a 6-digit sequence. The ID card number is the primary form of person identification, and is widely used by whenever a government ID is required. Hence, its ubiquity can be considered similar with the Social Security Number (SSN) in the US. Unlike the social security number and the Tax Registration Number, the Identification Card Number identifies the ID card, not the person, hence changes each time a new ID card is issued. This means that there may well be multiple ID card numbers corresponding to a given person. In many cases organisations such as banks, request the TRN along with the ID card number, implying that there numerous databases in both the private and the

[4] `http://www.gsis.gr`

public sector that may be used to match TRNs to ID card numbers and vice versa.

2.3 Greek Social Security Number

In 2009, the Greek state, introduced the Social Security Registration Number (AMKA) aiming to unify transactions among insurance and healthcare institutions in Greece. Every Greek national is required to have an AMKA number. AMKA is a structured number consisting of 11 digits. The first six encode the person's date of birth while the following 4 is an incremental sequence number. The last digit usually indicates the sex of the owner. During registration for an AMKA, the Identification Card Number is required. Optionally, one may provide his TRN, as well [1]. Individuals can find their AMKA, by supplying their first name, last name, father's and mother's name as well as their full date of birth to a government operated website [4]. If the supplied information does not match exactly the corresponding information in the AMKA database, the site returns a message to the effect that the person does not exist. The only exception is the date of birth. If only the year of birth is supplied, the site may ask for additional information (such as person's TRN or Identification Card Number) towards validation. If all information is correct, the AMKA of the person is returned.

2.4 Phone Directory

As in many countries, the use of telephones in Greece (both fixed and cellular) is widely established. OTE, the main telecommunications provider in Greece, offers a phone directory service from its site [3] for landline phones. A newly updated version of the directory website offers features such as reverse search (using the phone number itself instead of the name), and location search using specific addresses, postal codes or even cities. Additionally, the directory includes cellular numbers, from OTE's subsidiary mobile telecommunication company, Cosmote. A cellular entry also exhibits the address that the cellphone owner had declared during number registration. This address usually points to owners residence.

2.5 Greek Voter Registration

The Ministry of the Interior, deployed a new web service to help voters find their electoral center on election day [8]. Voters, enter their first name, last name, father's name, mother's name and their year of birth and get back a screen with information from the Registrar General specifying their assigned electoral center. While this service is only useful during elections, it is available continuously.

2.6 Tax card for recording transactions

The Greek state, in an effort to monitor retail transactions and force shop owners to issue receipts, introduced a "tax receipt card" that the customer presents to

the shop owner as part of the payment. This allows the customer to declare the transaction to the tax authority on-line, thus earning some tax discount at the end of the year, while the tax authorities get instant notification of the transaction.

The disadvantage of this method is that the shops can use the number in the "receipt card" to track customers, thus dispensing of the need for loyalty cards. Moreover, if the use of these cards becomes obligatory (through, for example the imposition of financial penalties if an insufficient number of receipts is collected) then the customers will lose their right to refuse to identify themselves when making certain purchases.

2.7 Governmental Documents

From October 1, 2010, any organization receiving funding from the Greek government, is obliged to upload decisions and other documents to a publicly available repository, called Diavgeia [6] (in Greek, transparency). According to Diavgeia statistics, [5] so far, more than 10 million documents have been published by approximately 3,500 public organizations. Documents include hiring or purchasing decisions, detailed payroll lists, balance sheets and so forth. Hence it is an obvious source of private data such as names, Identification Card Number, Tax Registration Number, etc.

2.8 Other public resources

Without effective guidance, or coordination, the various systems are developed in an ad hoc manner with a shocking disregard of the preservation of private and confidential data on Greek nationals. Hence sensitive information can be easily collected from many sources. The following serve as examples of this trend:

Municipalities and other institutions upload lists of persons who have been hired. Like Diavgeia, these lists contain a wealth of personal information.

Moreover, lists of persons that are selected to work programmes and are published by ministries or other state organizations, are rarely anonymized [2,5] from any identification numbers. Worse, they include information that can be misused. In particular, such lists contain selection criteria like martial status, time of unemployment or disability degree. Each criterion is assigned a discrete score. Thus, it is easy to infer one's disability degree using the assigned score on that criterion. As email addresses can be collected using social networks [24], a spammer could send targeted spam to that individual. As of the writing of this paper, these lists still remain available to public.

Data security in such sites is also a problem. Even in the flagship IT system of the Greek state, the Taxis system operated by GSIS (which stores tax data for every company or person that has income or property in Greece), recently, had an extensive breach of security. This resulted in the leak of large parts of the database with tax returns and real estate ownership. In an unprecedented move,

[5] `http://et.diavgeia.gov.gr/f/all/stats`

the Hellenic Data Protection Authority (HDPA) fined GSIS for the security breach. The HDPA report [6] indicates that large chunks of the database (about 70GB) had been copied and made available to private companies who processed the data and stored it in a MySQL database.

3 Data Crawling infrastructure

As we have noted in the previous section, a lot of private data is held in publicly accessible repositories protected with weak authentication mechanisms. The common assumption is that only the owner of this information can access it, because the systems rely on knowledge of personal information that only the owner can know. This assumption can be broken by, on one hand, searching for the required information in other sites, and by guesswork (trying out repeated guesses until we hit the correct value). In this section we describe various techniques used and methodology followed towards collecting data from a variety of sources.

Our first target was the Tax Registration Number web service, described in the previous section. Our objective was to submit all possible TRNs and thus get all the records that were accessible to public. As stated in section 2, the algorithm for generating TRNs is used in form validation of state websites. Therefore, we generated every valid Tax Registration Number, using the TRN generation algorithm. This resulted in approximately 90 million valid Tax Registration Numbers. For each one, we queried the Tax Registration Number web service and recorded the returned information. To make our collection mechanism more efficient, we segregated the set of valid Tax Registration Numbers into 10 subsets using the first digit as a filter. Each subset contained approximately 9 million elements. As Greece's population does not exceed 11 million residents, it is obvious that the search space would be sparse. From empirical observations, we concluded that registered TRNs are those starting from 0, 1, 2 and 99. Hence, we requested numbers from these ranges first and then search for the remaining space for completeness. To avoid triggering any detection mechanism we implemented the collection mechanism in a distributed fashion; A searcher component performs requests to the service using a set of TRNs, while results are stored in a centralized machine. Searchers were deployed on multiple PlanetLab nodes located in Greece. To avoid possible lock-outs from the web-service, we requested random TRNs from each subset using an interval of 100ms between consecutive requests.

The Diavgeia document repository was our second target. Although, it employs indexing mechanisms to document data, it does not index the documents' contents. Luckily a third-party service called yperdiavgeia [10], indexes both data and metadata of documents posted on Diavgeia. Moreover, it incorporates faster search mechanisms as well as applies OCR techniques to scanned documents, broadening the scope of potential results. In order to automate the process of locating and extracting Identification Card Numbers that appear in documents

[6] HDPA, decision 98/2013.

we used the following heuristic; Many instances of Identification Card Numbers are preceded by the string AΔT (the initials of Identification Card Number in Greek). Moreover, we observed that the person's name, surname usually follow the Identification Card Number in the text. We, therefore, searched the yperdiageia documents for instances of the string "AΔT" (a possible reference to identification numbers) and download them locally. Using pattern matching we identified each Identification Card Number location within the document, and tried to find person's name that may have been located near the ID reference via common syntax patterns. A valid reference to a name along with an Identification Card Number would be "John Papadopoulos, son of George, with AΔT AB-123456". This heuristic produced very good yields, because most official documents follow set patterns for stating the name, Identification Card Number, and address of individuals.

In the case of the telephone directory (white pages) crawling, we used the location search feature offered by the OTE website. One could search for phones using street names, postal codes or even names of cities (e.g Athens). In cases where the name and telephone number fields were left blank, the engine returned every phone number from the specified area or region. In such cases the results were limited to 1,000 records. To extract the telephone numbers and their owners from the website, we relied to scraping mechanisms along with headless browsers provided by Selenium. Queries were constructed using a combination of known addresses, postal codes and cities. To extract the information we wanted (telephone numbers and owners) from the HTML text, we used the crawling and screen-scraping framework, Scrapy in conjunction with the headless browser framework, Selenium. We implemented a Scrapy spider that issued synthesized queries to the telephone number directory website, and extracted our data from the results. Queries were constructed using a combination of known addresses, postal codes, and cities provided by the Greek Postal Service. To avoid being blocked for potential rate-limiting mechanisms, queries were performed periodically using a fixed 90-second interval.

Despite the fact that crawling utilizes multiple machines to throttle data collection, a malicious user without adequate collection resources, could still considered as a threat, by targeting specific individuals of his interest and collecting only the relevant data for them. Therefore, the collection effort is minimized.

4 Data Analysis

In this section we perform a coarse-grained analysis of collected data in order to prove our hypothesis that the combination of discrete datasources results in greater information leakage for individuals. We also present some proof-of concept data mining attacks that may be implemented against these repositories.

4.1 Data Completeness

The first question, concerns the quantity of data we were able to acquire using our design. In particular we seek to answer, what is the size of each data source,

Datasource	Results
Active Business Tax Registration Numbers	1,900,035
Phones (cellphones and landlines)	3,326,658
Identification numbers (total)	74,760
Voters	720,255
AMKA numbers	108,867

Table 1. Data collected from public sources

hence answering what is the portion of data we have collected. Data collection results are depicted on Table 1. According to a report issued by the General Secretariat for Information Systems [7], 2,082,396 Tax Registration Numbers are assigned to businesses and persons working free lance. Our crawling mechanism gathered 1,900,035 business Tax Registration Numbers covering the 91.2 % of the aforementioned set. Regarding non-business related TRNs (i.e. those assigned to salaried persons and pensioners), even if no actual record was returned, the GSIS web service gave a big hint. In particular, it is possible to discriminate between a Tax Registration Number belonging to an individual and an unassigned TRN, using the returned error code. Despite providing no additional information, the indication that the TRN is valid allows us to reduce the set of TRNs that would be used in a future brute force attack. In total we have discovered 7,397,876 Tax Registration Numbers belonging to individuals.

Regarding the acquisition of information on telephone numbers, we have gathered 1,676,195 PSTN and 1,650,377 cellphone numbers. There were 4,927,000 PSTN and 15,254,000 subscribers in Greece at the end of 2012 [8] [9] . Hence we located the 34.2 % of PSTN numbers and the 10.81 % of cellphone numbers.

Harvesting Identification Card Numbers from documents in the Diavgeia repository, resulted in 9,370 unique numbers. We were able to identify the person associated with the specific Identification Card Number in half of our set. If we add Identification Card Numbers leaked by documents published from municipalities, the total number rises to 74,760 records. As every Greek citizen is obliged to have an Identification Card, we can assume that the set of Identification Card Numbers is at least the size of the population which is 10,815,197 according to the Greek demographics [9]. Thus, our collected data concerning Identification Card Numbers cover the 0.69 % of the aforementioned set.

The Greek Ministry of the Interior [7], reported that in the recent elections held in June 2012, there were 9,947,876 registered voters. Our crawling methodology resulted in acquiring information for 720,255 voters, thus covering 7.2 % of the set.

[7] `http://www.gsis.gr/gsis/export/sites/default/gsis_site/PublicIssue/documents_Statistics/statdeltio2011.pdf`

[8] `http://www.3comma14.gr/pi/?survey=15701`

[9] `http://www.3comma14.gr/pi/?survey=13821`

4.2 A Taxonomy of Tax Registration Numbers

Each record of the Tax Registration Number database includes information about business activity that the owner of the TRN is engaged in. Thus we can identify persons according to their line of work. For example, we can identify journalists or newspaper publishers based on the business activity identifier in their tax record. Such information could be misused for malevolent purposes. As the home address of these journalists becomes public knowledge, a terrorist or a crook, could use it for blackmailing or terrorist attacks.

4.3 A closer look at Diavgeia Documents

The Diavgeia document repository contains miscellaneous documents mostly about the public sector. Despite, the obvious benefits of Diavgeia in promoting transparency in government decisions, we believe that specific and serious privacy concerns arise from the publication of such documents on the Internet.

Diavgeia includes contracts of individuals with specific organizations. The contract usually specifies first and last name as well as the monthly wage. Therefore, we can acquire information about one's income. However, not all contracts are been published in Diavgeia.

Furthermore, many municipalities publish construction permits. Each permit contains information about its owner, namely full name as well as Tax Registration Number. As every individual can acquire a construction permit, the referred Tax Registration Number, could potentially point to a wage earner or pensioner, mappings between non-commercial TRNs and names. The fact that most of these documents are scanned (hence bitmaps) is only a minor obstruction, as OCR software can easily convert them to searchable form.

Moreover, documents may contain Identification Card Numbers. The prevalence of Identification Card Number as a primary means of identification (such as the SSN in the US), could lead to impersonation attacks. One could leverage Identification Card Number and owner names, performing a social engineering attack to phone customer services.

5 Mitigation

The problem of balancing transparency against the protection of privacy is very hard indeed, and to a large extent philosophical, rather than technical. Nevertheless, there are numerous techniques that, if deployed, would diminish the extent of leaks by limiting or eliminating the effectiveness of our methods. In this section we discuss countermeasures that could be adopted by government entities towards prevention of similar information leaks in the future.

Rate-limiting Rate-limiting techniques are widely used to throttle the number of requests originating from a specific user or host. Despite its primary use in thwarting Denial-of-Service (DoS) attacks, rate-limiting can be also used to prevent rapid-fire requests of the type that we described earlier in this paper.

For example, the TRN web service can introduce a daily limit on the number of requests that can be issued from a given IP address. The granularity of this limit can be adjusted to accommodate legitimate uses of the facility. From our crawling experience, we faced lock-outs from the TRN web service after a long period of crawling and only to the extent that specific IP addresses we used for our crawler were blocked. We are not aware as to whether this block was manually enforced, or automatically, triggered by a rate limiting mechanism. Since we were not the only ones downloading the contents of that particular service (we are aware of at least one company which apparently did something similar) we were not surprised when the service was eventually discontinued temporarily, with no specific justification.

CAPTCHA: Such methods can be applied to web services to prevent brute-force attacks. After a number of successive requests originating from the same IP address, a CAPTCHA must be solved. This would spurn most automated brute force attempts. During our crawling, we did not encounter any website incorporating CAPTCHA techniques. For some web services, such as the TRN service, the root of the problem was lack of any user authentication. By authenticating users and asking them to be bound by some guidelines on the use of the service, there would be little scope for the kind of massive data downloads we have carried out. In many cases, the mere fact that the client has been identified will be sufficient to deter abuse.

Data Sanitization: As we have observed, much sensitive information was hidden inside Government documents found either on the Diavgeia repository or on municipality websites. With the introduction of the yperdiavgeia search engine [10], this unstructured information can be indexed. Thus an attacker may search for specific names or TRNs of his interest, performing a more targeted attack. As a countermeasure, Diavgeia document repository can sanitize references to names or surnames prior to document publication. Instead of displaying the full name of an individual, only a portion may be visible. For example, a reference to "John Papadopoulos" would be sanitized to "J. Papad.". Moreover, governments can enforce a stricter policy, for making Personally Identifiable Information (PII) available from sources outside Diavgeia. As we have shown, major privacy leaks were effected though municipal or other institutions linked to the public sector. Decisions containing sensitive information must be sanitized or anonymized and sent to Diavgeia.

Coordination: The plethora of data sources and the disparities in their design and operation significantly contributed to the creation of the vulnerabilities we exploited during our crawling.

Since 2011, the UK government has been trying, with some success, to bring state-run websites into the fold of gov.uk. In this way redundant websites can be axed while the rest can be made to comply with a common set of rules[10].

Greece has to do something similar to prevent each new website from the pitfalls experienced by other, older, web sites. Already the GSIS site is providing single log-on services to a very small number of websites (e.g. companies wishing

[10] `http://www.bbc.co.uk/news/uk-politics-25950004`

to register with the appropriate Chamber of Commerce can authenticate via GSIS). Soon every Greek citizen will have log-on credentials with the GSIS website which means that GSIS will be able to function as an authentication service for other state-run websites. However, this assumes that at least the GSIS site is itself secure. However, we have not seen any announcements to the effect that the recommendations offered by the HDPA after the security breach we discussed earlier have been implemented. Moreover, all sites authenticating via GSIS will have to meet some common privacy and security guidelines and undergo security audits at regular intervals.

Accountability: By identifying civil servants who are responsible of PII leaks, we envisage that proper vigilance will be observed on the part of the authorities who publish documents on public websites. To that effect, analysis of document metadata (e.g. Word documents storing the name of the author or the modification date) may produce valuable information leading to the source of disclosures of PII to the public [12,18]. Finally, Government can also use decoy documents [14], with "bait" information like TRNs or AMKAs, as a method to identify leaks.

6 Related Work

The concept on privacy exposure through public available data is not new. United States Social Security Numbers (SSN) is susceptible to conduct fraudulent actions through social engineering. Their prevalence as a means of identification made them a prime target for someone attempting identity thefts. Studies [13, 16] have indicated that the use of SSNs to identify individuals should be discouraged.

The first research, concerning privacy leakage from a Greek State datasource was conducted by Gessiou et al. [17]. The authors investigated whether personal identifiable information are publicly available on Greek web sites and documents, and if they are sufficient to extract a person's AMKA number for the AMKA web service [4]. Using these past results as a starting point, the work presented in this paper carries it a step forward. Our study is also related with the article [20], where various governmental open databases are discussed in terms of preserving citizen privacy. A similar study has been conducted by Simpson [25] for the UK government repository, `data.gov.uk`, showing how public data can be misused in terms of a privacy breach. Whang and Garcia-Molina [26] showed that adversaries can collect various private data from diverse sources and combine them resulting in a more precise piece of information for individuals. They proposed a model that can quantify the amount of a person's information leakage from a collection of data sources.

Personally Identifiable Information (PII) are present not only in the text of a document, but also to its embedded metadata. The first work investigating the problem was conducted by Byers [15] where Microsoft Word documents were crawled on the Web and searched for hidden words or deleted SSNs within document metadata. Gessiou et al. [18] collected over 10 million Microsoft Office doc-

uments from Google using synthesized queries. Using information present only on document metadata (such as documents contributors), constructed the relation graph of document contributors and searched Twitter to investigate whether such relationships are retained on social networks. Aura et al. [12] implemented a tool capable of identifying PII in documents that can potentially be used for tracing document authors or organizations. The tool automatically harvests sensitive information using heuristics from the user's computer, and searches for their presence in a collection of documents. Our work focuses on the document's context instead. In particular, we focus on documents like hiring lists or public documents uploaded on the Diavgeia repository to extract sensitive information like Identification Card Number and associate it with their respective owners. Narayanan and Shmatikov [23] denote that typical de-identification techniques are not sufficient for privacy, as information that distinguish one person from another (commercial transactions, browsing and search histories) can be used to re-identify anonymous data. They discuss that privacy cannot be guaranteed solely by anonymizing the data, but rather by enforcing policies concerning their usage.

A rich source of Personally Identifiable Information are social networks. Mao et al. [22] indicated that users can inadvertently release sensitive information to the public, such as vacation plans or disclosure of one's medical condition. Authors showed how a miscreant count leverage such information to perform automated attack on specific victims. Krisnamurthy et al. [21] showed that social networks can leak PII of their users to third-parties, like ad services. Wondracek et al. [27] introduced an attack to deanonymize social network users. They indicated that memberships to groups of a user can act as a fingerprint and can be exploited using history stealing attacks.

7 Conclusion

In this paper we discussed ways that information from multiple government sites or lists available on-line may be combined to create profiles of Greek citizens. As this sensitive information is publicly available, a miscreant may exploit it for malevolent purposes. Furthermore, ethical questions arise from the publication of this type of sensitive information. We looked at specific examples related to terrorism, identity theft, stalking, and spam, but these are only the obvious cases of the unauthorised use of the information provided by state institutions. As more private data are released on the Internet, there will be many more novel abuses of this information.

Acknowledgments

This work was supported in part by the project ForToo, funded by the Directorate-General for Home Affairs under Grant Agreement No. HOME/2010/ISEC/AG/INT-002 and by FP7 project SysSec, funded by the European Commission under Grant Agreement No. 257007.

References

1. http://www.amka.gr/odigos4.html.
2. http://www.asep.gr/asep/site/home/Tabs/autepistasia/autepistasia-sub1.csp.
3. 11888.gr (Greek Phone Catalogue). http://11888.ote.gr/web/guest/home.
4. AMKA Web Service. https://www.amka.gr/AMKAGR/.
5. Charitable Work Programme. http://www.epanad.gov.gr/.
6. Diavgeia Document Repository. http://diavgeia.gov.gr.
7. Greek Elections 2012 - Ministry of Interior. http://ekloges.ypes.gr/v2012b/public/.
8. Greek Electorate Web Service. http://www.ypes.gr/services/eea/eea.htm.
9. Hellenic Statistical Authority. http://www.statistics.gr.
10. UltraCl@rity - Search in the depths of the Cl@rity program. http://www.yperdiavgeia.gr.
11. VAT Registration Numbers Web Service. http://www.gsis.gr/wsnp/wsnp.html.
12. T. Aura, T. A. Kuhn, and M. Roe. Scanning Electronic Documents for Personally Identifiable Information. In *Proceedings of the 5th Annual ACM Workshop on Privacy in the Electronic Society*. ACM, 2006.
13. H. Berghel. Identity Theft, Social Security Numbers, and the Web. *Communications of the ACM*, 43(2):17–21, 2000.
14. B. M. Bowen, S. Hershkop, A. D. Keromytis, and S. J. Stolfo. Baiting Inside Attackers Using Decoy Documents. In *Proceedings of the 5th International ICST Conference on Security and Privacy in Communication Networks*, 2009.
15. S. Byers. Information Leakage Caused by Hidden Data in Published Documents. *Security & Privacy*, 2(2):23–27, 2004.
16. S. Garfinkel. Risks of Social Security Numbers. *Communications of the ACM*, 38(10):146, 1995.
17. E. Gessiou, A. Labrinidis, and S. Ioannidis. A Greek (privacy) Tragedy: The Introduction of Social Security Numbers in Greece. In *Proceedings of the 8th Annual ACM Workshop on Privacy in the Electronic Society*. ACM, 2009.
18. E. Gessiou, S. Volanis, E. Athanasopoulos, E. P. Markatos, and S. Ioannidis. Digging up Social Structures from Documents on the Web. In *Proceedings of the Global Communications Conference (GLOBECOM)*. IEEE, 2012.
19. S. Glenn. Marijuana bust shines light on utilities, Jan. 29, 2012. http://www.postandcourier.com/article/20120129/PC1602/301299979.
20. T. Keenan. Are They Making Our Privates Public?–Emerging Risks of Governmental Open Data Initiatives. *Privacy and Identity Management for Life*, pages 1–13, 2012.
21. B. Krishnamurthy and C. E. Wills. On the Leakage of Personally Identifiable Information via Online Social Networks. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*. ACM, 2009.
22. H. Mao, X. Shuai, and A. Kapadia. Loose Tweets: An Analysis of Privacy leaks on Twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*. ACM, 2011.
23. A. Narayanan and V. Shmatikov. Myths and Fallacies of Personally Identifiable Information. *Communications of the ACM*, 53(6):24–26, 2010.
24. I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas, and E. P. Markatos. Using Social Networks to Harvest Email Addresses. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*. ACM, 2010.

25. A. Simpson. On Privacy and Public Data: A study of data.gov.uk. *Journal of Privacy and Confidentiality*, 3(1):4, 2011.
26. S. Whang and H. Garcia-Molina. A model for Quantifying Information Leakage. *Secure Data Management*, pages 25–44, 2012.
27. G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A Practical Attack to De-Anonymize Social Network Users. In *Proceedings of 2010 IEEE Symposium on Security and Privacy*, 2010.