OCEAN
Open source Collation of eGovernment data and Networks

Srishti Gupta*, Mayank Gupta*, Ponnurangam Kumaraguru*
*Indraprastha Institute of Information Technology, Delhi
+Delhi College of Engineering
{sishti, pk}@iiitd.ac.in {mayank.gupta@dce.edu}

1. INTRODUCTION
In the world of computerization, database records consisting of personal attributes related to an individual are maintained and updated online by the government of Delhi. The Delhi government runs several information portals (Voter ID, Driving License, MTNL directory) which provide access to personal information of an individual. The attributes like name, address, phone number, age, Date of Birth (DOB), Father’s/Mother’s name etc. which govern the offline identity of a person are stored in such databases. A user is bound to share his/her details with the government and hence has no control on its security. Even though different government departments maintain individual databases, this project aims to showcase that the data can be aggregated together and thus potentially exploited [1]. An offline footprint of the user can be created collating all the available information which leads to a privacy breach. For e.g., all such information can be used to create fake documents, open fake bank accounts. Also, users today exhibit an online identity by using various social networking sites. Here the user shares his/her information according to his/her choice. These networking sites thus present a rich source of information. Since user himself gives all the information, it cannot be considered as a privacy loss but if all the sources are combined together, it may give out much more information which is undesirable by the user, as they may have expected the information to remain in the boundaries of that particular site. So, an online footprint of the user can be created which gives additional information about him/her. This project caters to the former problem of aggregating data from various open government sites. The aim is to identify such information and analyse possible privacy related concerns that may arise out of easy accessibility and public availability of the data. Data from OSN’s are collected to show the availability of the data like name, location, profile image and link to profile page from these networking sites.

2. RESEARCH METHODOLOGY
The project was divided into three phases. The first phase involved identifying all the databases that hold public information readily available on the Internet. We identified Driving licence, Voter ID, Twitter, Facebook, Foursquare, GooglePlus and LinkedIn as possible databases from which we could extract personal data. Since the MTNL website changed its policies by not allowing reproducing their content, we could not get the data from their site. The next phase was creating a possible notion of extracting information programmatically. Mechanisms like CAPTCHA, IP-address blocking often make such databases non-accessible for mining. For e.g. the website containing PAN card details contains the CAPTCHA thus making it difficult for extraction. The data was collected from all the sources and stored in a local data store, containing records for a large number of individuals. Our considerable success in collecting Personally Identifiable Information (PII) for a large number of individuals, pointed out many privacy issues which may be exploited by a malicious entity. The third phase was the aggregation of data across the government data sources.

As a part of this research, we developed OCEAN: Open source Collation of eGovernment data and Networks [http://precog.iiitd.edu.in/research/ocean], a system where the user can enter little information about a person and get all possible data about him/her. Basic interface of the system is shown in Figure 1. The system architecture is described in Figure 2.

3. IMPLEMENTATION
Below we describe the various stages of the system.

- When the user first visits the site, he is required to input the name of the person he wants to search and his location (optional). The database is queried at the back-end. In order to ensure the security of the system from automated attacks,
CAPTCHA is being used. If the user tries to query the system multiple times, he is asked to re-enter the CAPTCHA. After giving the inputs, the user is directed to the next page.

- Here the user has an option to select one of the four databases (Voter-ID, Driving license, Social profiles, Results across all government databases). For further discussion, let us assume that the user chose the Voter-ID database source to get the details. The system gives a list of all the people existing in the database with similar names. It includes details like voter ID, address, age, gender, father’s name. When the user clicks on a specific person, he can view his/her family tree as shown in Figure 3. Information about user’s parents, siblings and spouse could be identified with such a tree. It helps in visualizing the family structure of the user. The aggregated data is shown under the Results across all government databases tab.

- Figure 4 shows the social footprint of the user aggregated from three OSN’s viz. Twitter, Foursquare and Facebook. The data from social networks is obtained through API calls. Each social network has its own public API which can be used to get the results. Aggregation of data from social networks help in getting the information which can not be obtained from a single OSN. This is because some attributes are made private in particular OSN and hence cannot be retrieved while it may be possible to get from some other OSN.

4. **DATASET**

The data that we have collected from various data sources until 17 January is summarized in Table 1.

![Image of Family Tree](image)

**Figure 3: Family Tree:** Clicking any link from list of all the names, the user can view the family tree of the person which shows information about his/her parents, siblings and spouse.

![Image of Social footprint](image)

**Figure 4: Social footprint on OSN:** It shows the social footprint of the user from the results aggregated from three OSN’s (Twitter, Foursquare and Facebook).

5. **DISCUSSION**

The system was made LIVE on January 20, 2013. As on January 21, 2013, the number of unique visitors recorded was 103 and the number of unique searches was 390. The feedback given by the users shows that they liked the system. Some of the comments collected are “This is awesome..we’ve got PeekYou - Free People Search and other searches but this one is specialized for us :).” “Truly awesome!!.” Using the system built by us, we demonstrated that we can find out a lot of private information about a user. We could extract the name, address, father’s name, age, Voter-ID, gender, driving license. Social network gave us information like name, profile image, profile url. This is a huge source of data which can be potentially exploited. The fetched details can be used to create fake documents, fake bank accounts.

6. **LIMITATION AND FUTURE WORK**

Datasets available are limited to the residents of Delhi. We have only considered databases which are maintained by the Delhi government, election commission. The data collection is not complete as the scraping missed out some pages. Since the API calls to the social networks is rate limited, the process of collection is slow. As part of the future work, national level databases can be incorporated. We can focus on a specific user collecting his information from all possible sources. It can be used to calculate the reputation score of the person in the society. We plan to aggregate the results from social media to associate large number of attributes to a single entity and hence depicting privacy loss.

7. **REFERENCES**