Leaked Databases and User Risk

Alex Bainbridge[1] Kenny Cohen[2] Frank Liao[3]
Carnegie Mellon University

***Abstract*— Database leaks are becoming more common as the world moves their data online. Companies need to consider database security and user privacy issues now more than ever as they have become targets for stealing identities. User information is no longer confined to the domain where it was originally generated, and now captures a larger picture of the user. From this, data from different domains can be combined to create a wholistic "User Profile" which describes aspects about a users many different personas found throughout the various domains on the internet. This "User Profile" can also be used as a starting point for inference attacks, greatly increasing the amount of risk users undertake by using online services. We explored the practical dangers of leaked datasets with our user profile definition and were able to make inferences about various users we encountered. Due to this, we feel that there should be more future work in obfuscating databases in order to help protect users in the event of a database leak.**

I. Introduction

Our research explored publicly leaked datasets and how the information contained in these datasets can be used to narrow the anonymity set of an online persona - thus allowing us to tie personas back to the origin user. Our work focuses on two aspects of analysis, the potential harm a dataset may have and the combined harm many datasets may have. To accomplish this, we gathered numerous datasets leaked over the years and developed a code book and a methodology for joining datasets together. Our codebook was used to quantify harm that could come to the user for a particular field being leaked - we utilized 4 main categories: no risk, small risk, large risk, detrimental risk. Further, we defined a "profile" for an online persona that contains what we believe to be the most relevant factors of someone's identity. Using this, we attempted to fill in as many fields as we could by cross referencing the datasets we collected. Together, our work attempts to quantify not only theoretical risk of leaked datasets, but practical risk, and attempts to tackle the issues surrounding large anonymity sets.

II. Problem Statement

Anonymity is a large issue that effects many online users. Information collected by sites is expected to be confidential and not used to bring any harm to the users providing the information. However, there are issues when the service is attacked and information is leaked to the general public. Due to this, it is important to measure how much harm a particular dataset may have for the users present. Further, this information may be used by itself, or in combination with other information from other sources. Unless the risks are appropriately articulated, many databases may not have the proper safeguards and access controls to protect their users from harm.

III. Literature Review

Due to recent events, there has been an increased pressure on companies to secure their databases and protect user data. In fact, in the first half of 2017 alone there were 1.9 Billion records of data leaked and stolen from company servers. [1] This is very worrisome as with the internet of things providing more avenues for data collection, attackers may be getting their hands on more information than ever before. To further complicate matters, social media has become an ever evolving data source that has allowed miners to gain meaningful insight about people, activities, and events.

A significant advancement in the use of social media for data inferencing was done by Minkus et al., 2015. [2] Their efforts focused on using purchased voter registration data, to search social media for other data points. These were then combined to create verbose data profiles on both adults and children. From their starting dataset of about 70,000 people, they were able to match 10,000 to a facebook profile with a "precision" value just above 90

To acquire enough related information on a given user, we will be relying on the insight from Perito et al., 2011. [5] Their work in demonstrating that usernames can be tied to user online identities and can be used as rough identifiers. From their data, they found that users tend to reuse usernames or make similar usernames between sites. (They also acknowledged that various usernames were also "non-identifying", and couldn't be tied to an identity with a high confidence.) The basis that usernames can be used to identify users is an assumption that we take on in our work, in order to link datasets together.

Finally, our work involves coding several datasets and our target online profile. The inspiration for these codes from Delugach and Hinke, 1994. [6] Delugach and Hinke discuss that there are 7 classes of inference targets, and that they range from entities, activities, and various relationships built up from these bases. In particular, they define the concept of inference path - "the sequence of steps used in making an inference". These definitions are important to our work because they give a baseline for how inferences are made, and how lower levels of classified data can be combined in

[1]Alex Bainbridge is a student in Information Systems at Carnegie Mellon University abainbri at andrew.cmu.edu
[2]Kenny Cohen is a student in Information Systems at Carnegie Mellon University kicohen at andrew.cmu.edu
[3]Frank Liao is a student in Information Systems at Carnegie Mellon University fliao at andrew.cmu.edu

an algorithmic way to infer information at higher levels of classification. Our code book extends upon this and labels the levels of classification in the data sets, and our online profile. Our analysis then works on building some of the various rules of interaction between the classification based on the data we collected.

Our contribution to the field relies on quantifying the harm leaked datasets may have for users and their online personas. We explore harm in both the theoretical bound and the practical bound, and also tackle issues surrounding large anonymity sets and propose methods to reduce their size. Further, we provide a prognosis for dangers emerging with IOT devices becoming more prominent, and how social media and people search services can also bring users into additional harm when combined with leaked datasets.

IV. Methodology

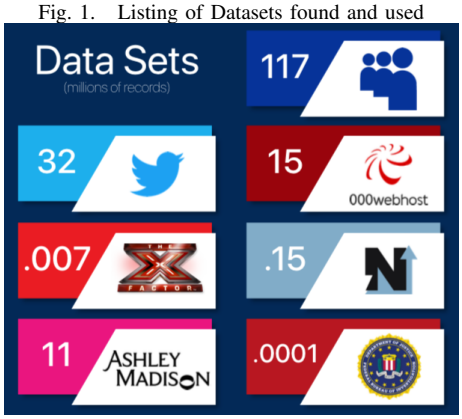
A. Gathering Datasets

To explore the types of data that is leaked containing user information, we utilized publicly available datasets that we found online. These datasets covered a wide array of services, were from different years, and contained differing fields. We performed two different analyses on these datasets, the first was through a code book and the other through cross referencing the datasets.

We found datasets by looking at the news for recent database breaches, and searching The Pirate Bay for leaked databases. Additionally, we identified a prominent ethical hacker named Cthulhu who posted many public datasets that we were able to access using the Wayback Machine service. The datasets we collected are: Ashley Madison Database, Myspace Database, Twitter Database, X-Factor Contestants, Nulled_IO, and Webhost. These datasets covered a wide range of user fields and covered a large timeframe. Some datasets had 100's of millions of records, and others had only hundreds. Following this, we decided to clean the datasets and put them into the same format. Our cleaning process involved using the regex:

```
"^[A-Za-z0-9\.\+_-]+@[A-Za-z0-9\._-]+
  \.[a-zA-Z]*$"
```

Fig. 1. Listing of Datasets found and used



B. User Profile

First, to go over a few definitions we will be using. A "User" is defined as the actual person who is using the accounts online. In our work we simplified this so that multiple people using an account will only reflect a single "User". A "Persona" is a single online profile that a user has. A "User Representation" is the culmination of a user's personas into a condensed format. Our goal is to make our "User Profile" match closely with the user representation for a particular user. Our definition and creation of User Profiles follows.

We define "User Profile" for use in our research as facets of a user's online presence. These facets were selected from an example list of general Personally Identifiable Information or PII data. The example attributes were defined as data which support the ability to distinguish or trace an individual, such as name and social security number, and information which is linked or linkable to an individual, such as medical and financial records.[7] Linkable information is less variable, and therefore less distinguishable when analyzing specific individuals; however, linkable information is especially prominent online because it is made readily available by social platforms, which will aggregate data such as religion, geographical information, activities, etc. For this work, PII attributes were chosen based on their online prominence and presence. Certain data, such as driver's license number and complex biometric data (retina scans, x-rays, etc), have almost no online presence and were rejected.

C. Code Book

For our code book, we used two different coding schemes. The first being a risk analysis one, and the other being a inference target analysis - our inference targets come from Delugach's work[6].

The risk categories were no risk, small risk, large risk, and detrimental risk. Our definitions for each were: No Risk: The field offers no risk to the user if leaked OR no effect on the anonymity set. Small Risk: The field offers minimal risk to the user, minimal in that something is learned about the user, but has no lasting impact OR slight changes to anonymity set. Large Risk: The field offers large risk to the user, the user may be severely impacted by this AND decreases the anonymity set significantly. Detrimental Risk: The field offers large risk to the user, the user may be wholly compromised with this information AND small anonymity set.

We performed 4 rounds of coding in order to achieve an inter-rater agreement of 100% through Cohen's kappa. We found that on average, most of the information leaked is small risk to no risk. We conjecture that this is a symptom of the datasets we have gathered, as the ones most likely to be leaked are the ones with less security measures because they are guarding less sensitive material. However, in our datasets we found that the most likely to be leaked information was legal_name (found in 35 datasets) and email (found in 56 datasets). This can have a major impact on the user as these fields can be used to join other datasets for inference making.

The inference targets were Class 1, Class 2, Class 3, Class 4, Class 5, Class 6, Class 7. Which were defined as, Class 1: The materialization of an entity. Class 2: The materialization of an activity. Class 3: The materialization of a sensitive relationship between materialized entities. Class 4: The materialization of a sensitive relationship between materialized activities. Class 5: The materialization of a sensitive relationship between one or more materialized activities or entities. Class 6: The materialization of a sensitive relationship between sensitive relationships. Class 7: The materialization of a sensitive rule from existing classes.

We performed one round of inter-rater coding, and achieved an agreement of 100%. Only four of our fields were considered to be of class 2 - IP address, Zip Code, Location, and Timestamp , all others were class 1. We believe other classes would develop in future analysis where we could use expert systems to inference un-hashed data together.

D. Dataset Cross Referencing

For our dataset cross referencing, we first hashed all PII and secondary form PII in compliance with the IRB's suggestion. We hashed the data using the SHA256 algorithm with the built-in python hashing library. This process took a considerable amount of time because of the sheer size of the datasets, using several computers was key for us. In order to work with the data however, we still had to chunk the file into parts that could be read into memory easily.

After hashing the data, we cleaned the data so that it was all in a similar format, and removed bad records identified via malformed emails. Many had issues with whitespacing, some were just raw SQL dumps, some were separated data values with commas and others with colons, and many datasets had different ways of representing null values. So to make this process feasible, we decided upon a format to convert all of the files to and wrote many different conversion algorithms.

Once the data files were hashed and cleaned, we came up with three different algorithms for cross referencing the datasets. We consider two datasets A and B where each set has m and n records respectively. The first options was to simply provide no further analysis and iterate over each record in file A and iterate through each record in file B to see if there were any matches.

The second option was to sort both files and run binary search for matching. However on trial of writing this algorithm, we discovered that we did not have an easily accessible sort function that didn't read the target into memory. Due to this we retired this option, however we suggest it for future works.

The third option was to create a hash table in the form of a dictionary and then keep track of the number of records that hash to each place in the table. (We re-hashed the data in order to be consistent with the paradigm that we had access to the raw data and were attempting to perform this analysis in the wild.) We ultimately used the third option. The only caveat being that the MySpace file had to be broken down into smaller files because even the hashed identifiers from that dataset was too large to read into memory on their own.

V. Findings

A. Code Book

Our codebooks are presented below, the first being the Risk Levels coded for each field found in our datasets (duplicates removed). The second is the coding for inference targets. The third is a list of possible inferences made from the data, and their respective Risk Level and Inference Target Class.

TABLE I
CODE BOOK - RISK LEVELS

Variable	Value Label
Username	No Risk
Email	Small Risk
IP Address	Small Risk
Password	No Risk
SHA(Password Truncated)	No Risk
SHA(Salted Password)	No Risk
First Name + Last Name	Detrimental Risk
First Name	Small Risk
Last Name	Small Risk
Location	Detrimental Risk
Date of Birth	Small Risk
Phone	Small Risk
Gender	Small Risk
Zip Code	Small Risk
Timestamp	No Risk

TABLE II
CODE BOOK - INFERENCE TARGETS

Variable	Value Label
Username	Class 1
Email	Class 1
IP Address	Class 2
Password	Class 1
SHA(Password Truncated)	Class 1
SHA(Salted Password)	Class 1
First Name + Last Name	Class 1
First Name	Class 1
Last Name	Class 1
Location	Class 2
Date of Birth	Class 1
Phone	Class 1
Gender	Class 1
Zip Code	Class 2
Timestamp	Class 2

TABLE III

CODE BOOK - INFERENCES FROM DATA

Combined Fields	Resultant	Risk Level	Inference Target
Username + Password	Persona Account Access	Large Risk	Class 3
Email + Password	Persona Account Access	Large Risk	Class 3
Zip Code + Birthdate + Gender[8]	Persona Identification	Large Risk	Class 5
First_name + Last_name	Legal_name	Detrimental	Class 3
Phone + Email	User Contact Information	Large Risk	Class 3
IP Address + Password	Login Access	Large Risk	Class 5
Username + IP address	Persona Tagging	Small Risk	Class 5
Location + Timestamp	User Location at Time	Detrimental Risk	Class 4
User Location at Time + User Tagging	User Tied to Internet Activity	Detrimental Risk	Class 6

From table 1, we can see that most of the risks for individual fields are actually no risk or small risk. This is concerning because it can lead to this information being less protected or in some cases given out freely to whoever wants it. We defined these risks to be in accordance with two factors, learning about a user, and narrowing down the anonymity set. As an adversary is able to learn more about a user, they can gain more information to be used for further inferences or attacks, and as they narrow the anonymity set a user's identity may be revealed along with their personas. From this, we conjecture that an adversary could develop and fill out a user representation based off of all the knowledge gained from numerous no risk - small risk database leaks - and further that this user representation will actually hold significant information that could be used against the user.

For the second table, we found that most inference targets released were of class 1 or class 2. This provided the groundwork for our inferences as we had well defined rules for how to combine the various classes. This means that a simple rule based system combined with an expert system would be able to make inferences about most data that comes from leaked databases - thus significantly lowering the barrier for entry for inference attacks. Further, an interesting aspect of these databases was that we saw mainly entities and few activities, we conjecture that this is a symptom of the online service databases we captured, as many of the activity logic may be at the application level.

For the final table, we explored possible inferences that could be made, their Risk Levels, and the Inference Target classes. We were able to develop inferences from Classes 3-6, but not 7. The inferences ranged from Small Risk to Detrimental Risk with a varying amount of information needed. We developed these inferences from either work in our literature review, or from the meta data we believe these fields provide. Some of these inferences deal with gathering more information, attacks on privacy, or attacks on the user directly. As seen in the second to last row of the table, these inferences can also be used to create more inferences such as tying internet activity to a particular user. This can have adverse effects for users as they are potentially giving up a lot more information to sites and online services than what they anticipated giving. We conjecture, that given these inferences and the raw data that we collected from our leaked datasets - that we could de-anonymize several participants in our datasets. Further, that not only can we de-anonymize them, but we can build a comprehensive picture of their online user profile that is accurate. We were unable to do this ourselves due to our limitations with hashed data however.

B. Datasets

The results of aggregate inferencing were inconclusive. We started aggregating datasets starting with the sets we believed to be most interesting and would be successful in adding additional fields to the user profile. We determined this order to be, Ashley Madison, MySpace, Twitter, NulledIO, Webhost, FBI - Atlanta, and finally X-Factor. Even though the largest, most similar sets were matched first, there were still insignificant amounts of information being discovered for entries because several of these datasets were not diverse in terms of data fields. If each discovered profile had matches from every dataset that was analyzed, there would have only been a maximum of 6 unique attributes discovered.

When looking at all unique values discovered across all attributes, results were still insignificant. Profiles discovered an average of 5.59 unique values added to their information, from a minimum of 5 unique values to a potential maximum of 6.64 unique values. The minimum is determined by the two datasets that determined the base set of profiles that was created from matching our two largest datasets. The maximum is determined by averaging possible unique values for our total aggregate matches from the perspective of only our username-identified profiles. If it were possible to match emails to usernames, the maximum increases to a potential 7 possible values discovered for each profile.

One intriguing aspect of our matching process, was that we received significantly more matches in some cases than we expected. A significant amount of datasets do not contain information about email addresses and instead use usernames. We conjecture that our match rate is inflated due to the re-use of usernames from un-related users across many domains. A good example of this was the Twitter dataset, which contained both usernames and emails. When usernames were matched, 17.02% of entries were matched; subsequently, when emails were matched, only 0.429% of entries were able to be matched.

We calculated the Jaccard Index across our dataset matches using only a depth of 2 matches. And found that while some datasets seemed to be very closely related, others were disjoint sets. For example, one factor of concern is that

the FBI had a relatively high match rate with the Twitter database. This raises additional concerns over how these leaked datasets could be used to attack not only users but also the organizations that employ them.

It is not surprising that we saw zero matches between the FBI dataset and the WebHost, XFactor and NulledIO datasets. The FBI dataset was rather small and for a completely different domain than the other datasets. The matches between the FBI dataset and the MySpace and Twitter datasets are significant because of the risk that it presents to the users that were represented in the FBI database being able to connect them to their public online profiles can create risk for someone in the FBI.

TABLE IV
DATASET STATS

Dataset	Entries	Notable identifiers
Myspace	358,066,582	email, username
Twitter (emails)	56,204,754	email
Ashley Madison	36,396,204	username
Twitter (usernames)	15,440,003	username
WebHost	15,152,237	email, username, ip
NulledIO	159,298	email, username
XFactor	73,727	email, legal_name, date_of_birth, phone
FBI	137	email username, legal_name

There were additional fields in some of these datasets such as Ashley Madison, however we chose to explore this set of identifiers.

TABLE V
PAIR INFERENCING - MATCH COUNT

	Twitter	Myspace	XFactor	WebHost	AshleyM	FBI	NulledIO
Twitter		5225071	1997	497922	X*	6	3812
MySpace	5225071		12786	311455	26	6	2244
XFactor	1997	12786		244	X*	0	4
WebHost	497922	311455	244		32	0	15585
AshleyM	X*	26	X*	32		X*	24
FBI	6	6	0	0	X*		0
NulledIO	3812	2244	4	15585	24	0	X*

X* denoting a section where the datasets could not be merged due to not having matching fields.

TABLE VI
PAIR INFERENCING - JACCARD INDEX

	Twitter	Myspace	XFactor	WebHost	AshleyMadison	FBI	NulledIO
Twitter		0.012159	0.0000278	0.005737	X*	0.0000000837	0.0000531
MySpace	0.012159		0.0000357	0.000835	0.0000000659	0.0000000168	0.00000626
XFactor	0.0000278	0.0000357		0.0000160	X*	0	0.0000172
WebHost	0.005737	0.000835	0.0000160		0.000000621	0	0.0010179
AshleyMadison	X*	0.0000000659	X*	0.000000621		X*	0.000000657
FBI	0.000000837	0.000000168	0	0	X*		0
NulledIO	0.0000531	0.00000626	0.0000172	0.0010179	0.000000657	0	X*

X* denoting a section where the datasets could not be merged due to not having matching fields.

TABLE VII
AGGREGATE INFERENCING

Combined Dataset	cumulative entries	unique emails	unique usernames
Ashley Madison + MySpace	8,890,034	8,890,030	4,445,005
Ashley Madison + MySpace + Twitter	11,759,407	11,518,417	4,685,991

VI. LIMITATIONS

We had a number of limitations that narrowed the scope of our work and final findings. The first and perhaps most detrimental limitation was that we did not have access to an expert system, nor the resources or time to build one. This prevented us from making more advanced inferences based on contextual or expert information. Another important limitation on our work was that we had only limited access to a limited number of datasets. Some of the ones we acquired were only partial leaks, and we only had access to a small subset of total leaked datasets. Further, we were not able to access any leaked databases that were purchasable. Thus, our data was from a very small sample pool in comparison to the possible population we could have obtained. This greatly limited the scope of our analysis and the possible matches and inferences we could have made.

Another condition on our work that hindered our ability to make inferences, was the constraint of using encoded attributes. This was to be in compliance with an expedited IRB proposal, and meant that we could not perform partial string matching, pseudonym generation, nor cross field matching. This greatly restricted the cross-referencing step of our analysis because we could only rely on exact matches on the same field between data sets - meaning that we are greatly under-utilizing our datasets and not fully exploring them. Further, we could not find useful trends or analysis on any interesting database fields because these were hashed. One example of a inference we wanted to make, would be if people had addresses close to each other, and if this implied any relationship between their user accounts in the various datasets.

Lastly, even though we only gathered a limited number of data and could only perform basic matching and inferencing, we still had a significant amount of records. Available computing and processing power was very limited. We spent a considerable amount of time cleaning and inferencing the datasets as well as determining efficient methods for making the inferences.

VII. FUTURE WORKS

A. New Data Sources - Social Media

For future work, we suggest an exploration into public social media websites and people search services. By combining leaked database information with publicly available information more comprehensive user profiles can be created that further diminish the anonymity set. This process may also be used in reconnaissance attacks as currently most doxing attacks are performed using services such as Facebook, PIPL, and other public services rather than relying on leaked datasets. However, by combining these with leaked data sets may provide better and more fruitful results.

Another advancement in information gathering could be collecting unformatted data such as tweets, facebook likes, and other dynamic information being generated online. These data sources would have to be processed into a format for large data processing and then linked to various social media profiles, but could might be used by expert systems to create some unique kinds of inferences rather than just static social media profile matching.

B. Data Expansion and Code Book Expansion

Additionally, we believe a study with more datasets and complete datasets would greatly improve the number of credible matches that could be gained through cross-referencing, and that this would grow even more if the data was unhashed. Thus allowing for more complex matching procedures to take place. With these more complex matching procedures and richer data coming in, expert systems could be used to help generate more inferences.

These new inferences could be used in a much larger codebook in order to help quantify more risk scenarios for users and database managers. By focusing on an expanding code book, more users and security members will be made aware of common inferences that could be made and work to limit their impact and knowledge gained.

Another interesting topic that could provide fruitful results, would be disinformation campaigns inside databases. Where users or security members elect to provide false information in the datasets and only the application software know which information is true or not, thus complicating inference attacks much further as the anonymity set could grow widely with the additional fake fields.

C. IOT

Another aspect that could be explored, would be the use of leaked IOT information. For example, in a few of our databases we found IP address information. This could be combined with leaked datasets of IOT vendors to infer which IP addresses may have unpatched/vulnerable IT devices. Further, this could be coupled with public datasets such as Shodan. Meaning that along with being identified, users may also be attacked - making this a primary concern as IOT continues to develop in today's culture.

D. Data Obfuscation

The previous future works explore how to make these attacks more damaging. However, there is also the availabilty for work to help lessen the effects of database leakage. One such work that we believe would be helpful is database obfuscation. Database obfuscation would provide an additional

layer of security to protect users from purely data leaks. We propose that the de-obfuscation process happen not at the database level but in a different security domain. Thus an attacker would have to compromise an additional domain in order to de-obfuscate any data collected.

APPENDIX

The code and codebooks we used for our work can be found at the following git repo.
`github.com/Dreadwall/Database-Inference-Attacks`
Databases are not provided as they are already public domain knowledge.

Fig. 2. A heat map for the Jaccard Index

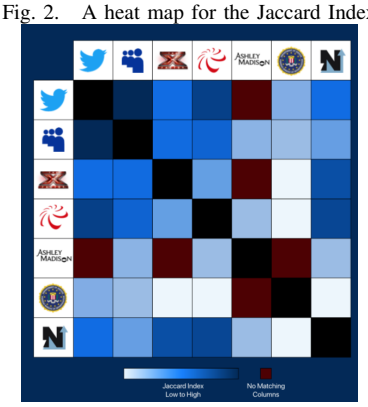


TABLE VIII
DATASET FIELDS

ATTRIBUTE	Value Label	Weighting	Number of Datasets it appears in
legal_name	Detrimental Risk	3	3
ip_address	Small Risk	1	1
race	No Risk	0	1
gender	Small Risk	1	2
date_of_birth	Small Risk	1	2
email	Small Risk	1	5
username	No Risk	0	3
password	No Risk	0	4
zip_code	Small Risk	1	1
city	Small Risk	1	1
employer	small Risk	1	1
interests	No Risk	0	1
relationship_status	No Risk	0	1
looking_for	No Risk	0	1
	AVG:	0.7142857143	

ACKNOWLEDGEMENT

Special thanks to our two faculty advisors, Nicolas Christin and Norman M. Sadeh. Professor Christin met with us several times and helped point us in the right direction for our research. Professor Sadeh provided important feedback and direction during our presentations to the class.

REFERENCES

[1] "The Reality of Data Breaches." Breach Level Index, breachlevelindex.com/assets/Breach-Level-Index-Infographic-H1-2017-Gemalto-1500.jpg.
[2] The City Privacy Attack: Combining Social Media and Public Records for Detailed Profiles of Adults and Children
[3] C. Jernigan and B. F. Mistree. Gaydar: Facebook friendships expose sexual orientation. First Monday, 14(10), 2009.
[4] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology, 9(2):69–83, 2009.
[5] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How unique and traceable are usernames? In Privacy Enhancing Technologies, pages 1–17. Springer, 2011.
[6] Delugach, Harry S., and Thomas H. Hinke. "Using Conceptual Graphs to Represent Database Inference Security Analysis." CIT. Journal of Computing and Information Technology, cit.fer.hr/index.php/CIT/article/view/3067/1929.
[7] McCallister, E., Grance, T., & Scarfone, K. A. (2010). Guide to protecting the confidentiality of personally identifiable information (PII). Special Publication (NIST SP)-800-122.