To demonstrate weaknesses in the k-Anonymity approach to data anonymization by extracting information from the Social Web

Vinay Bharadwaj
Georgia Institute of technology
vbharadwaj6@gatech.edu

Nishith Agarwal
Georgia Institute of Technology
nagarwal36@gatech.edu

ABSTRACT

Data mining and privacy is a growing field. Anonymized private datasets (e.g., list of diseases and demographics affected by them) are released for research and analysis purposes. Even though the data in the anonymized dataset is anonymized, with the huge amount of private data that is available easily from social networks we can perform various analyses on them to infer very interesting and concrete results, such as the identity of the person and other private details that are not meant to be released. There has to be ways in which one can protect his/her data. Analyzing several approaches put forth to provide such protection, we enlighten ourselves with loop holes and flaws that might be present in such data anonymization techniques. k-Anonymity is one such anonymization technique whose effectiveness we're trying to question.

The web data is one of the most interesting data mines that renders itself to such analyses. The web (especially social networks) actually contains information which, if structured and compared, does reveal itself to be a knowledge base rather than just lines of information. Techniques like L – diversity and t – closeness have been proposed which question the fallacies of k – anonymity. They concentrate on the assumptions of k-Anonymity, marking them as flaws. But little light has been thrown on the fact that even with improved techniques if knowledge can be derived from the available web data, this background knowledge helps us weaken the guarantees provided by such privacy techniques.

GOAL

Our goal in the project is twofold. The first aim is to expose weaknesses in the K-anonymity approach to privacy. Secondly, we intend to study the patterns of private social data on the web and get concrete results.

I. INTRODUCTION

With the extensive use of the web, people are socializing more and more online. Before signing up on an online social community, the user is required to fill in a large amount of personal details. When using the social network, users share their thoughts, activities, comments, favorites, personal information and lots of other useful information. Little have people realized that such a huge amount of data is actually a potential privacy threat.

The K – anonymity concept on privacy of data proposes that if one has a set of information, deriving particular information about a person is difficult since every data is k – anonymous, meaning that there are k-1 similar data to the one being targeted. But with sufficient background knowledge, one can challenge this technique. With the huge amount of data on the web, crawling through it can build substantial database of background knowledge which can be used in an attack. A k – anonymous dataset can be joined with this concrete data to highlight important information which is not intended to be disclosed.

II. APPROACH

In this project, we targeted the social networking giant, facebook to collect public data about people. We specifically targeted the groups for various diseases on facebook like “Cancer Survivors”, “Heart disease patients”, “Breast cancer” etc. We collected data about members on these groups which included their public personal information such as gender, birthdate, location, activities and interests. We also extracted other connections from their profile like family members and friends.

In parallel with this, we extracted other publicly available data (eg. Voter database) that could be used in conjunction with our facebook data to help strengthen our stand. The voter database provides information like birthdate, address, location etc. We collected data about 20000 users on facebook and corresponding voter information from the voter databases. We then constructed k-anonymized datasets in accordance with the cancer statistics of Washington state and compared our data with the anonymized data to uniquely identify people, thus weakening k-Anonymity approach to privacy.

III. TECHNOLOGIES USED:

1) Javascript Obect Notation (JSON) parser – org.json library

   org.json is a library that can be used to parse JSON objects and arrays. The Facebook Graph API returns profile information in JSON format.

2) Apache HttpClient and Gargoyle HtmlUnit

   HttpClient and HtmlUnit are used to login to facebook to get additional information from profile pages.

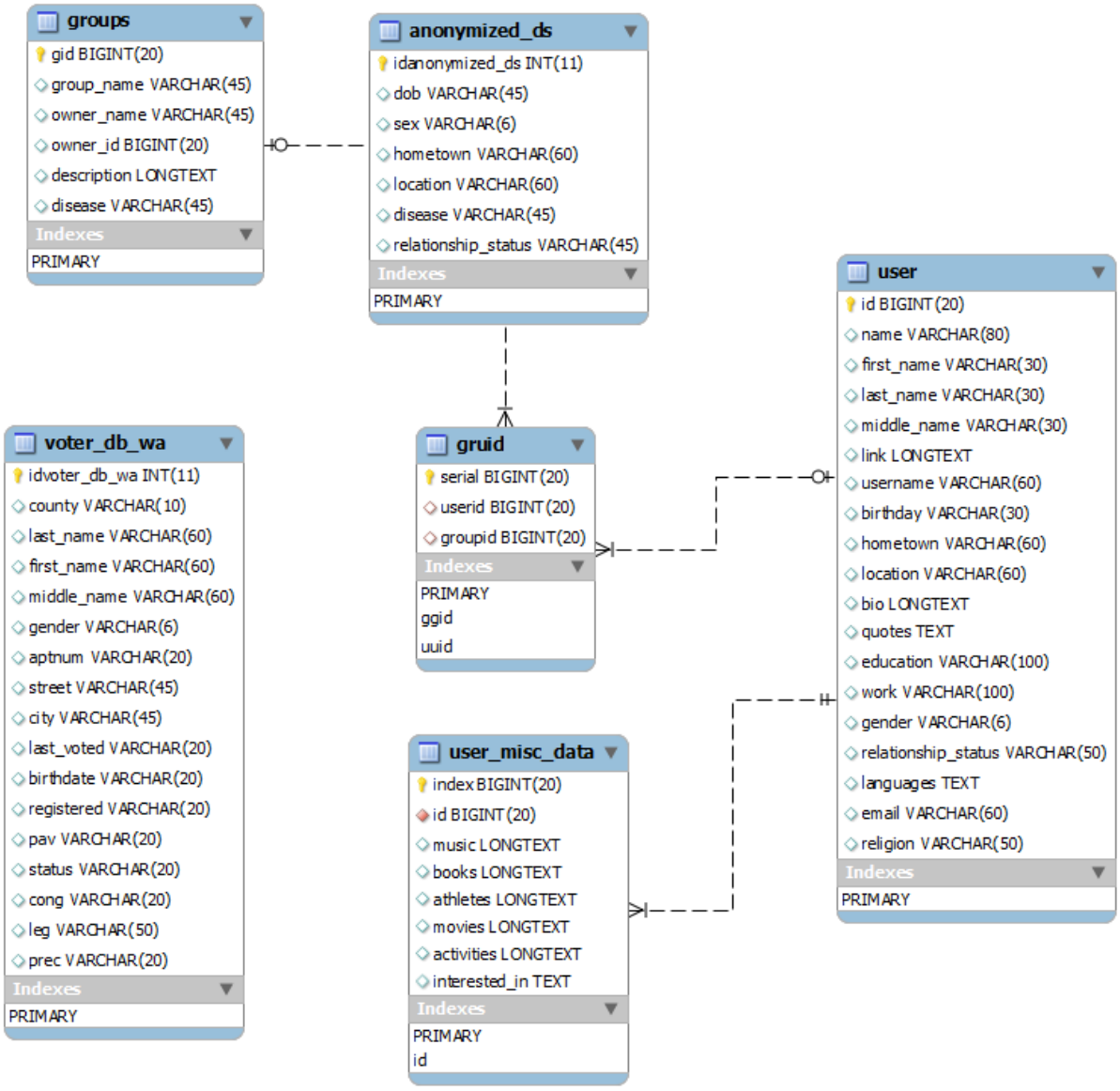
3) JSOUP – HTML parser

   JSOUP library is used to parse HTML pages returned by HtmlClient.

4) MySQL Workbench and JDBC client

   MySQL is our database and JDBC client provides database connectivity from JAVA.

IV. DATABASE SCHEMATA – EER diagram:



V. DATA COLLECTION:

i) Facebook Graph API

ii) Groups and Profiles

iii) Voter Database

Facebook Graph API

The Facebook Graph API returns information about a person or group in the form of JSON objects. This is meant as a utility for applications running on facebook that need to collect user information or user permissions, but it renders itself as a loophole that an attacker can exploit. An example of the JSON objects array returned by the Graph API is presented below:


```
{
   "id": "677820261",
   "name": "Vinay Bharadwaj",
   "first_name": "Vinay",
   "last_name": "Bharadwaj",
   "link": "http://www.facebook.com/thejedivind",
   "username": "thejedivind",
   "birthday": "02/23/1989",
   "hometown": {
      "id": "106377336067638",
      "name": "Bangalore, India"
   },
   "location": {
      "id": "107991659233606",
      "name": "Atlanta, Georgia"
   },
   "bio": "http://memoirsduvind.blogspot.com\r\n\r\nAlmighty Freedom,\r\nAlmighty freer of the soul,\r\nBe free,\r\nWalk with me,\r\nThrough the golden fields,\r\nSo
lovely..",
   "quotes": "God proposes, man disposes!\r\n\r\n\"When life becomes an irony...\" - via Shishir Ramesha",
   "work": [
      {
         "employer": {
            "id": "103128663061181",
            "name": "Fedora Project"
         },
         "location": {
            "id": "106377336067638",
            "name": "Bangalore, India"
         },
         "position": {
            "id": "164818406883966",
            "name": "Brand Ambassador & Software Packager"
         },
         "description": "RPM packaging from source tarballs.\nTesting package quality with rpmlint.\nPromote Fedora project.",
         "projects": [
            {
               "id": "207358022614725",
               "name": "SkyViewer - Packaging",
               "description": "SkyViewer is an OpenGL-based program written by Nicholas Phillips to display HEALPix-based sky maps from FITS format files. The program
will display sky maps on a 3D sphere or a 2D Mollweide projection. Real time panning and zooming are supported, as are rotations of the 3D sphere (if you have a fast
graphics card).\t"
```


We extracted all the publicly available information about the person/group from the Graph API and stored it in our database. We use org.json library to parse the JSON objects.

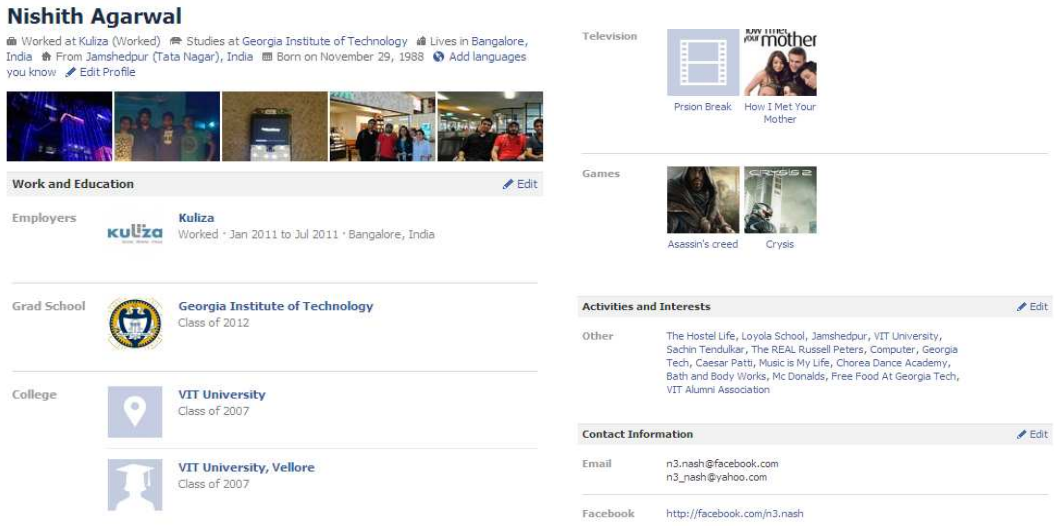
Groups and profiles

We crawled more than 30 groups and 20000 profiles. In deciding which groups to crawl, we chose the health related groups since hospitals and associations usually release health statistics and anonymized data sets for medical research and demographics. Our approach here was to get the members of the each group from the Graph API in a JSON array and then recursively crawl the members' pages using HttpClient library and HtmlUnit library.

Below are some of the groups that we crawled:

1) Cancer survivors
2) Breast cancer
3) Diabetes
4) Type I diabetic children
5) Heart disease
6) Kidney cancer survivor group
7) AIDS
8) People with diabetes
9) Lung cancer people
10) Fallen leaves and cancer survivors

We also crawl individual user profiles apart from the information we got from the Graph API because the profile pages provide extra information about activities and interests.



Screenshots of the facebook database:

Overview | Output | Snippets | Query 1 Result

Fetched 1000 records, more available. Duration: 0.000 sec, fetched in: 0.094 sec

id	name	first_name	last_name	middle_name	link	username	birthday	hometown	location
116771	Jill Muller	Jill	Muller		http://www.facebook.com/jecmuller	jecmuller	null	null	null
934499	Graham Lubinsky	Graham	Lubinsky		http://www.facebook.com/graham.lubinsky	graham.lubinsky	null	Irvine, California	Grand Rapids, Michigan
1009241	Scott Johnston	Scott	Johnston		http://www.facebook.com/happyinwater	happyinwater	null	Ann Arbor, Michigan	Darien, Connecticut
1801473	Jeremy Zitner	Jeremy	Zitner		http://www.facebook.com/jeremy.zitner	jeremy.zitner	null	Cold Spring Harbor, New York	Laurel Hollow, New York
1911073	Morgan Ore	Morgan	Ore		http://www.facebook.com/morgan.ore	morgan.ore	null	null	null
1938562	Matt Wontroba	Matt	Wontroba		http://www.facebook.com/OmegaLolrus	OmegaLolrus	null	null	null
2220351	Matt Schaar	Matt	Schaar		http://www.facebook.com/schaar	schaar	null	Rapid City, South Dakota	Detroit, Michigan
2534533	Jen Levine	Jen	Levine		http://www.facebook.com/profile.php?id=2534533	null	null	null	null
2737799	Matt Haha Heidel	Matt	Heidel	Haha	http://www.facebook.com/profile.php?id=2737799	null	null	Cookeville, Tennessee	Asheville, North Carolina
2908865	Lubomir Martin Ondrasek	Lubomir	Ondrasek	Martin	http://www.facebook.com/lmondrasek	lmondrasek	null	Martin, Slovakia	Chicago, Illinois
3318165	Kai Peter Chang	Kai	Chang	Peter	http://www.facebook.com/kai.chang	kai.chang	null	Taipei, Taiwan	Oakland, California
5028029	Claudia Ramirez Lagos	Claudia	Lagos		http://www.facebook.com/profile.php?id=5028029	null	null	null	null
5219890	Camilo Forero	Camilo	Forero		http://www.facebook.com/Cforclev	Cforclev	null	null	null

Overview | Output | Snippets | Query 1 Result

Fetched 1000 records, more available. Duration: 0.000 sec, fetched in: 0.078 sec

	movies	activities	interested_in
	null	null	null
	Dirty Dancing Pretty Woman Just Friends Official Grease Movie Food Inc	Stone Brewing Company Peyton Manning	null
ing (Steve Borden) Tim Lincecum Joe Thornton	Distribute Ayrton Senna Movie in the US, please	Junior League	null
	Documentaries Edward Scissorhands Lunch Line P.A.M (Post Apocalyptic Men) The Love Section (2012 Movie)	S. N. Goenka	Men
nny Powers of Eastbound & Down	Back to the Future Tron Monty Python s Quest for the Holy Grail Rocky Indiana Jones	Eating Treasure hunting	Men
	50/50 The Royal Tenenbaums Best in Show A Mighty Wind The Faculty	Eating Treasure hunting	Men
	Friday After Next Next Friday Baby Boy Coach Carter	Iota Phi Theta Fraternity, Inc. B.S.A.	Men
	Hope Floats For Love of the Game Eat Pray Love In Her Shoes Pretty Woman	Iota Phi Theta Fraternity, Inc. B.S.A.	Men
nn Marshawn Lynch	Harry Potter Serenity Attack The Block	Watching Movies Pembroke Welsh Corgis	Men
nn Marshawn Lynch	Harry Potter Serenity Attack The Block	Watching Movies Pembroke Welsh Corgis	Men
	James Bond Captain America Official Water for Elephants Movie Dirty Dancing Black Swan	Karate I Am Also Trying to Learn to Play the Guitar	Men
	Mrs. Doubtfire Most Anything Disney Disney Pixar Disney	Computer-Games Watching TV	Men and Women
ordan Floyd Mayweather	Why Did I Get Married Too (2010) Dear John Skank Robbers Titanic	Internet marketing Unemployment	Women

Voter Database

All U.S states release voter information about the voters in the respective state. This database has information about the voters such as names, address, date of birth, voter status etc. We queried information about people in our database in the Washington state voter database and extracted their addresses and other information. We later used this information to uniquely identify people with a particular disease and their addresses. Availability of such information is a serious threat to the privacy of such individuals because not only does it uniquely identify the person, but can reveal their addresses and other sensitive information too.

Below is a screenshot of a query from the voter database:

Sound Politics Washington State Voter Database

County	Last Name	First / Middle Name	M/F	Number	Street	City	Last Voted	Birthdate	Registered	PAV	Status	Cong.	Leg.	Prec.
FR	GOMEZ	TERESA	U	192e	ROAD 32	PASCO		1978-DEC-10	1998-JUN-10	V	C	4	16	14
KI	GOMEZ	TERESA B	F	923e	36TH AVE S	SEATTLE	2010-NOV-02	1946-MAY-31	2002-MAY-28	P	A	7	37	3159
CR	GOMEZ	TERESA M	F	1070e	NE 212TH AVE	VANCOUVER	2010-NOV-02	1989-NOV-16	2008-FEB-02	P	A	3	18	620
BE	GOMEZ	TERESA MARGUERITE	F	50e	S HARRISON ST	KENNEWICK		1979-AUG-30	2010-JUL-10	V	A	4	8	1650

Data is from Secretary of State's Voter Registration Database public release of Aug. 31, 2011

We were able to get information about all the people in our database who reside in the state of Washington.

VI. DATA ANALYSIS

i.) Weakening k-Anonymity by comparing k-anonymized dataset with our established database.

Our prime goal for this project was to question the effectiveness of the k-Anonymity approach. We constructed an anonymous table in accordance with the cancer statistics we got from the Washington state cancer registry and the Washington state Department of health. The distribution of cancer patients according to age groups in Washington state are as follows:

```
2008- WA State cancer stats
Total Female cases – 18014
Age gp  Number of reported cases  Percentage of total cases
0-4                42                      0.02%
5-9                25                      0.13%
10-14              31                      0.17%
15-19              56                      0.3%
20-24              104                     0.5%
25-29              188                     1%
30-34              290                     1.6%
35-39              499                     2.7%
40-44              857                     4.7%
45-49              1360                    7.5%
50-54              1843                    10.2%
55-59              1962                    10.8%
60-64              2182                    12.1%
65-69              2004                    11.1%
70-74              1888                    10.4%
75-79              1688                    9.3%
```

80-84	1515	8.4%
85+	1480	8.2%

Total Male cases - 17221

0-4	57	0.3%
5-9	20	0.1%
10-14	32	0.18%
15-19	61	0.3%
20-24	74	0.4%
25-29	125	0.7%
30-34	191	1.1%
35-39	223	1.2%
40-44	392	2.2%
45-49	709	4.1%
50-54	1353	7.8%
55-59	1998	11.6%
60-64	2438	14.1%
65-69	2546	14.7%
70-74	2321	13.4%
75-79	1954	11.3%
80-84	1505	8.7%
85+	1222	7%

The anonymized dataset that we constructed is as follows:

Overview Output Snippets Query 1 Result

Fetched 42 records. Duration: 0.000 sec, fetched in: 0.047 sec

idanonymized_ds	dob	sex	hometown	location	disease	relationship_status
6001	1978-SEP-27	female		Seattle, Washington	cancer	
6002	1962-MAR-25	male		Everett, Washington	cancer	
6003		male	Seattle, Washington	Winona, Minnesota	cancer	
6004		female	Seattle, Washington	Honolulu, Hawaii	cancer	
6005	1965-SEP-16	male	Bakersfield, California	Tacoma, Washington	cancer	
6006		female	Port Angeles, Washington	Scottsdale, Arizona	breast cancer	
6007		male	Wellpinit, Washington	Pablo, Montana		Single
6008	1965-SEP-16	male	Smyrna, New York	Tacoma, Washington	cancer	Single
6009		male	College Place, Washington		cancer	Married
6010	1978-SEP-27	male	Council Bluffs, Iowa	Seattle, Washington	cancer	
6011	1978-SEP-27	female	Tulsa, Oklahoma	Seattle, Washington	cancer	
6012	1978-SEP-27	male	Westport, Connecticut	Seattle, Washington	cancer	
6013		female	Olympia, Washington	Berkeley, California		
6014	1952-NOV-02	female	Longview, Washington	Longview, Washington	cancer	

By querying our database with the date of birth, gender, location, we can find a match for each of the above anonymized entries and uniquely identify the person and get their details.

county	last_name	first_name	middle_name	gender	aptnum	street	city	last_voted	birthdate	registered	pav	status	cong	leg	prec
CZ	JORDAN	SARAH	L	F		CANYON VIEW DR E	LONGVIEW	2011-FEB-08	1952-NOV-02	1975-AUG-05	V	A	3	19	52

name	first_name	last_name	middle_name	link	username	birthday	hometown	location	bio	quotes
Sarah Jordan	Sarah	Jordan		http://www.facebook.com/kale...	kaleighslover69	null	null	Washington	null	Life is too short to wake up in the morning with regrets

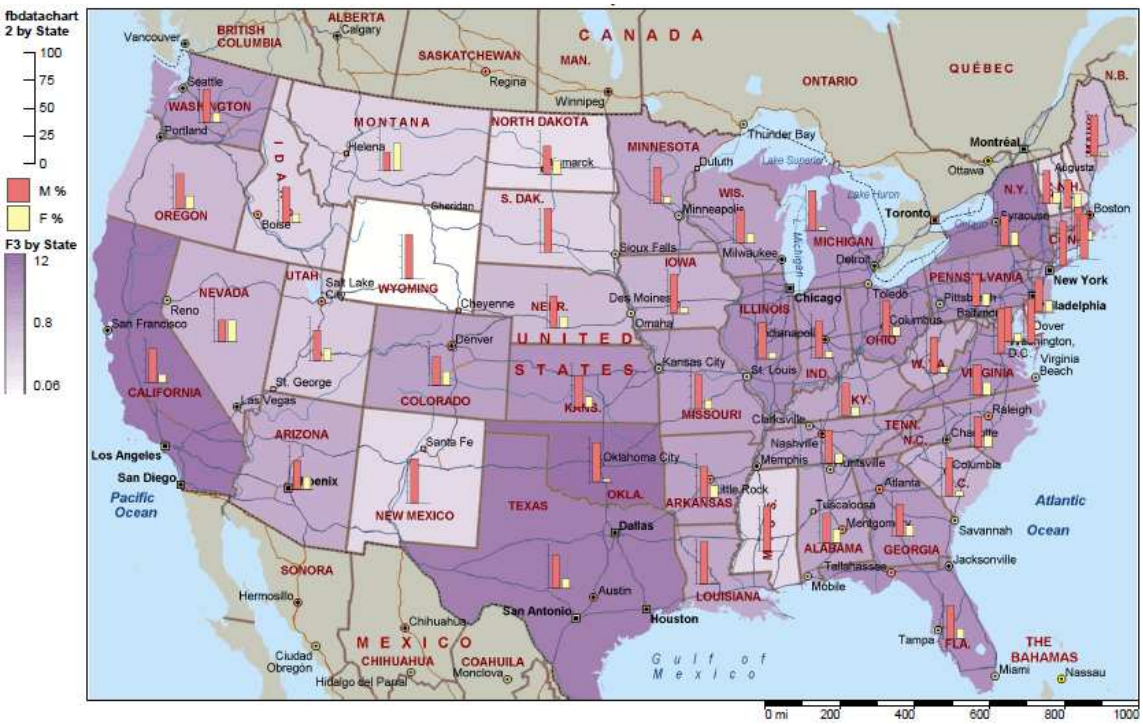
Thus, as shown above, we could uniquely identify the person with a particular disease and also get other details about them like their address. This kind of identification is possible with 90% of the data that we have. Just getting information from a voter's database made the probability of identification much higher. An attacker may have access to much more information, which makes this a significant privacy threat to people on social networking platforms.

Another thing we noticed is that people very freely post sensitive information about themselves on open groups which makes k-Anonymity even weaker and easier for an attacker to exploit.

One of the arguments that we put forth is "What use is anonymizing data to protect privacy when the people themselves give up their information so freely?" The answer that we got after some thinking is that most people do not know the threats of posting such information on social networking sites. They believe that the information that they post would only be shared among friends and family. But this is seldom the case because even a user who's not in the person's friend list can view significant information about them. And posting on open groups makes the posts accessible to the public. Sharing sensitive information on such groups makes it easier for an attacker to gather such information.

ii.) Statistics & graphs

Below, we post some of the graphs that show the facebook users with cancer who revealed information (by state) and male-female ratio of cancer in different states of U.S based on the data we collected.



iii.) Other interesting phenomenon

We also collected miscellaneous data about people like activities and interests, music, movies, sexual orientation, work place, education.

These fields may be exploited to gain more understanding of the individual and their psychology as well as their background.

music	books	athletes	movies	activities	interested_in
Muse deadmau5 Daft Punk Justice	null	null	null	null	null
Jimmy Buffett John Mellencamp Coun...	Jodi Picoult	Ben Bostrom Fan Page Jason Farrel...	Dirty Dancing Pretty Woman Just Fri...	Stone Brewing Company Peyton Manning	null
Liz Downing	Chuck Palahniuk Harry Potter August...	Brian Wilson Rated R Superstar Ed...	Distribute Ayrton Senna Movie in the...	Junior League	null
Ganesh H Hegde	Marion Nestle Food Politics Siddharth...	Derrick Rose	Documentaries Edward Scissorhand...	S. N. Goenka	Men
Toots & the Maytals CCR The Smashi...	Common Sense Pillars of the Earth Cl...	Eli Manning Babe Ruth Curtis Grand...	Back to the Future Tron Monty Pyth...	Eating Treasure hunting	Men
Phish Grateful Dead The String Chee...	Common Sense Pillars of the Earth Cl...	Ben Cohen	50/50 The Royal Tenenbaums Best...	Eating Treasure hunting	Men
Jazz Contemporary R&B Gospel musi...	The One Im Reading at the Time	Kevin Durant	Friday After Next Next Friday Baby B...	Iota Phi Theta Fraternity, Inc. B.S.A.	Men
Jason Aldean Miranda Lambert Zac B...	Heaven Is for Real	Kevin Durant	Hope Floats For Love of the Game ...	Iota Phi Theta Fraternity, Inc. B.S.A.	Men
Easy to Please	The Harry Potter Series Ender s Gam...	Jerome "Mighty Mouse" McGee Da...	Harry Potter Serenity Attack The Blo...	Watching Movies Pembroke Welsh Corgis	Men
Easy to Please	The Harry Potter Series Ender s Gam...	Jerome "Mighty Mouse" McGee Da...	Harry Potter Serenity Attack The Blo...	Watching Movies Pembroke Welsh Corgis	Men
Take That Black Eyed Peas Girls Alo...	The White Queen The Red Queen ...	Michelle Kwan Austin Collie	James Bond Captain America Officia...	Karate I Am Also Trying to Learn to Play th...	Men
Queen Bach Beethoven Mozart	The Hobbit The Lord of the Rings Th...	Michelle Kwan Austin Collie	Mrs. Doubtfire Most Anything Disney...	Computer-Games Watching TV	Men and Women
Gifted Pastor Clint Brown Ian Von Larr...	Heaven is so Real Think and Grow R...	Terrell Owens Chad Ochocinco Roy...	Why Did I Get Married Too (2010) D...	Internet marketing Unemployment	Women

CONCLUSION

The emergence of social networks and the fact that people post sensitive information unaware of the consequences make anonymization of private data harder to achieve. K-Anonymity by itself would have been a good anonymization technique but with a lot of sensitive private information available on social networks and on the web, any anonymization technique would fail to achieve significant efficacy.

In this study, we have shown the ineffectiveness of k-anonymity technique for privacy. In our further work, we intend to collect more data from various sources and conduct more analysis.

**All data collected during this study are for research and statistical analysis purposes only. We do not intend to breach the privacy of individuals nor release any data. Our goal is to show the weaknesses of anonymization techniques only.*

REFERENCES

[1] *k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY*, Latanya Sweeney
School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[2] *Applying l – diversity in anonymizing collaborative social network*. G.K.Panda , L.Mitra.

[3] *Worst-Case Background Knowledge for Privacy-Preserving Data Publishing*, David J. Martin, Daniel
Kifer, Ashwin Machanavajjhala, Johannes Gehrke, Joseph Harlpern

[4] Modeling and Integrating Background Knowledge in Data Anonymization, Tiancheng Li, Ninghui Li, Jian Zhang