FACEBOOK: BEAUTY OR THE BEAST?

Senem Acet Coskun

SENEM ACET-COSKUN

STA 9710

TERM PROJECT

Table of Contents

What is Facebook?

Facebook is a social networking website that is operated and privately owned by Facebook, Inc. Anyone who confirms themselves to be over the age of 13 with a valid e-mail address can become a Facebook user. Users can add friends and send them messages, and update their personal profiles to notify friends about themselves. Additionally, users can join networks organized by workplace, school, or college. Many people happily display their date of birth and details about their educational and employment backgrounds just because the site seems to be some sort of quasi dating site with the potential for job hunting thrown in.

Facebook was launched by Mark Zuckerberg in his dorm room when he was an undergrad at Harvard. His idea was to develop a social networking website for Harvard students only. Amazingly though, Facebook now has over 400 million users all over the world.

In December 2008, the Supreme Court of the Australian Capital Territory ruled that Facebook was a valid protocol to serve court notices to defendants! It is believed to be the world's first legal judgment that defines a summons posted on Facebook as legally binding. Employers (such as Virgin Atlantic) have also used Facebook as a means to keep tabs on their employees and have even been known to fire them over posts they've made. Long story short, even though it started as a fun project in a tiny dorm room, Facebook is the strongest medium to serve for marketing, human resources management and even for justice.

Then one can’t help but ask: is it really safe to share all kinds of private information with a “platform” even though you think that only your friend can “see” your information? Here is an interesting article from Huffingtonpost.com.

“…Business Insider has posted the transcript of an instant message conversation that allegedly took place between Facebook CEO Mark Zuckerberg and a "college friend" while Zuckerberg was still at Harvard. […] In the electronic back-and-forth, Zuckerberg allegedly told his friend "if you ever need info about anyone at Harvard, just ask" and called users who had shared information with him "dumb f--ks."

What private info?

If you really want to go public on Facebook, you can share all kinds of information about yourself such as your name, phone number, date of birth, and place of birth, current address, schools you attended, your employer(s), who you are friends with, where you have been so far, what you think about certain topics…and the list goes on.

And here comes the best part. Facebook “may” use all these information for its own purposes or share your information with third parties. For example, your profile might be sold to a marketing company to develop an online ad campaign for another company. You can’t prevent it. Once you create your account, you automatically accept it. Then another question immediately reveals itself. Do people know all these “terms and conditions” of Facebook or do they see it as an innocent way of contacting their friends and family? Why all the rush to reveal yourself for free?

OK. Facebook is after my private info. Who else?

Most people do not give their private information out to someone on the street but their guard seems to be dropping when it comes to social media. For an identity thief to begin his work, he needs very few pieces of information. Once he has found out a person's full name, date of birth, address, telephone number and social security number, there are many low level crimes that can be committed. Store cards can potentially be applied for or other basic identification documents can be applied for.

By displaying a number of these details online, an individual opens themselves to potential problems. When combined with a little background information which can be found on the site (place of study, hometown, type of job and employer) the task of impersonating a victim becomes much easier.

One way to prevent to be public is to switch off some features in the privacy settings. But here comes another attack story. Sophos, a world leader in IT security and control, published a study[1], in which it fabricated a fake Facebook profile and asked 200 Facebook users at random to give up personal information. The results were amazing. Out of the 200 friend requests, Sophos received 82 responses, with 72 percent of those respondents divulging one or more e-mail address; 84 percent listing their full date of birth; 87 percent providing details about education or work; 78 percent listing their current address or location; 23 percent giving their phone number; and 26 percent providing their instant messaging screen name.

According to the study, Facebook users were all too willing to disclose the names of spouses and partners, with some even sending complete resumes. One Facebook user divulging his mother’s maiden name—the old standard used by many financial and other Web sites to get access to account information.

The Underground Economy of Spam Emails

Spam email is a subset of spam that involves nearly identical messages sent to numerous recipients by e-mail. Spammers collect e-mail addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. They also use a practice known as "e-mail appending", in which they use known information about their target (such as a postal address) to search for the target's email address. Spam averages 78% of all e-mail sent.

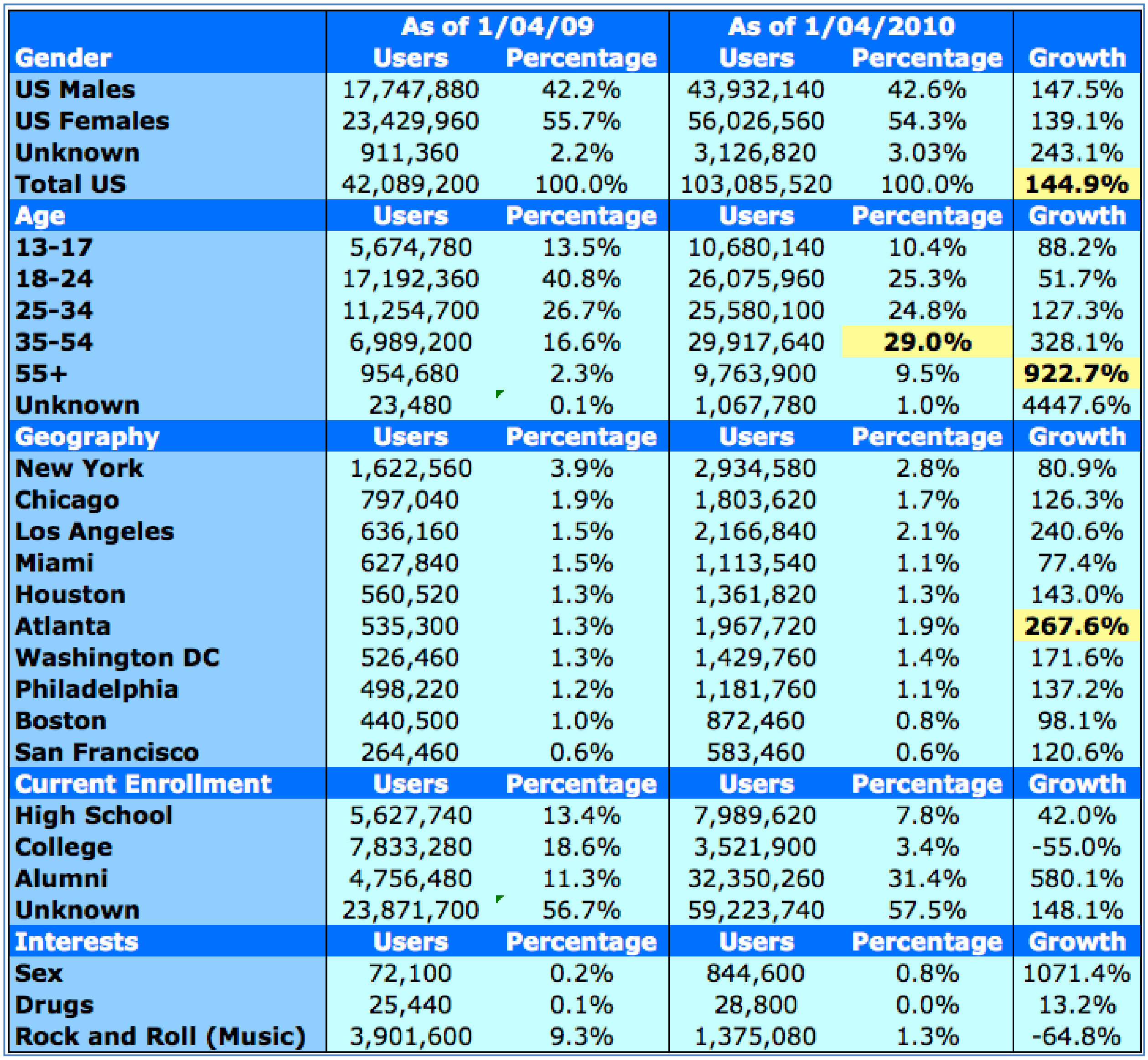
A 2004 survey estimated that lost productivity costs Internet users in the United States $21.58 billion annually, while another reported the cost at $17 billion, up from $11 billion in 2003. In 2004, the worldwide productivity cost of spam has been estimated to be $50 billion in 2005[2]. An estimate of the percentage cost borne by the sender of marketing junk mail is 88%, whereas in 2001 one spam was estimated to cost $0.10 for the receiver and $0.00001 (0.01% of the cost) for the sender[3].

In the underground economy, any kind of personal information has a dollar value. In the paper titled “Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy”, it is claimed that a stolen credit card number is sold for $12. Based on all this information the ultimate question then becomes: what is the contribution of Facebook to this underground economy? In other words, how much are we willing to give away about ourselves for others to make illegal money?

[1] http://www.macworld.com/article/59488/2007/08/facebook.html
[2] http://www.informationweek.com/news/security/vulnerabilities/showArticle.jhtml?articleID=59300834
[3] http://www.emailresults.com/article.asp?ContentID=6

Demographics of Facebook as of January 2010

Please notice the growth rates since 2009. Especially in the 50+ age category, it’s 927%.



	As of 1/04/09		As of 1/04/2010		
Gender	**Users**	**Percentage**	**Users**	**Percentage**	**Growth**
US Males	17,747,880	42.2%	43,932,140	42.6%	147.5%
US Females	23,429,960	55.7%	56,026,560	54.3%	139.1%
Unknown	911,360	2.2%	3,126,820	3.03%	243.1%
Total US	42,089,200	100.0%	103,085,520	100.0%	**144.9%**
Age	**Users**	**Percentage**	**Users**	**Percentage**	**Growth**
13-17	5,674,780	13.5%	10,680,140	10.4%	88.2%
18-24	17,192,360	40.8%	26,075,960	25.3%	51.7%
25-34	11,254,700	26.7%	25,580,100	24.8%	127.3%
35-54	6,989,200	16.6%	29,917,640	**29.0%**	328.1%
55+	954,680	2.3%	9,763,900	9.5%	**922.7%**
Unknown	23,480	0.1%	1,067,780	1.0%	4447.6%
Geography	**Users**	**Percentage**	**Users**	**Percentage**	**Growth**
New York	1,622,560	3.9%	2,934,580	2.8%	80.9%
Chicago	797,040	1.9%	1,803,620	1.7%	126.3%
Los Angeles	636,160	1.5%	2,166,840	2.1%	240.6%
Miami	627,840	1.5%	1,113,540	1.1%	77.4%
Houston	560,520	1.3%	1,361,820	1.3%	143.0%
Atlanta	535,300	1.3%	1,967,720	1.9%	**267.6%**
Washington DC	526,460	1.3%	1,429,760	1.4%	171.6%
Philadelphia	498,220	1.2%	1,181,760	1.1%	137.2%
Boston	440,500	1.0%	872,460	0.8%	98.1%
San Francisco	264,460	0.6%	583,460	0.6%	120.6%
Current Enrollment	**Users**	**Percentage**	**Users**	**Percentage**	**Growth**
High School	5,627,740	13.4%	7,989,620	7.8%	42.0%
College	7,833,280	18.6%	3,521,900	3.4%	-55.0%
Alumni	4,756,480	11.3%	32,350,260	31.4%	580.1%
Unknown	23,871,700	56.7%	59,223,740	57.5%	148.1%
Interests	**Users**	**Percentage**	**Users**	**Percentage**	**Growth**
Sex	72,100	0.2%	844,600	0.8%	1071.4%
Drugs	25,440	0.1%	28,800	0.0%	13.2%
Rock and Roll (Music)	3,901,600	9.3%	1,375,080	1.3%	-64.8%

Source: Facebooks Social Ads Platform

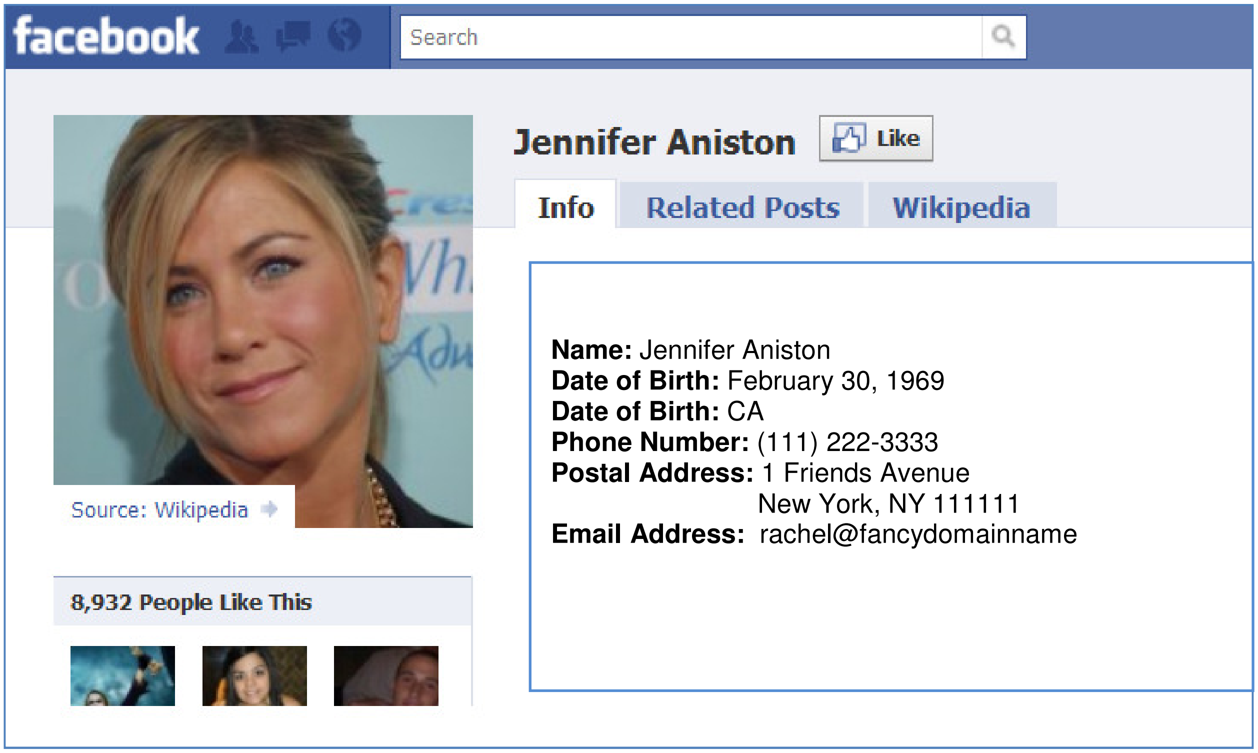
A Study on Facebook: Beauty or the Beast?

Estimating the Proportion of Americans who Provide Key Personal Information

There are over 400 million Facebook users worldwide. Among those, around 103 million are US users[4]. Since identity theft is a bleeding wound especially in the U.S and since identity thieves will be focusing on US accounts mostly, our study should focus on US users only.

In the study, information defined as “key” will include name, date of birth, place of birth, phone number, postal address, email address and a picture of the person. For an account to be considered as US, the owner should be residing in the US during the time of study.

A sample of key information:



Name: Jennifer Aniston
Date of Birth: February 30, 1969
Date of Birth: CA
Phone Number: (111) 222-3333
Postal Address: 1 Friends Avenue
New York, NY 111111
Email Address: rachel@fancydomainname

Disclaimer: Above information is for display purposes only. Does not include real information.

Requirements for the information to be considered in the dataset:

1- Name should include first and last name
2- Date of Birth should include at least year
3- Date of birth should include state
4- Phone number should include area code if postal address is not provided.
5- Postal address should include at least zip code
6- No special requirement for email address
7- Picture – as long as there is any- will be considered as personal information since even a pet or cartoon might reveal clues about personal life or preferences.

[4] Facebook Demographics and Statistics Report 2010 – 145% Growth in 1 Year.

Simple Random Sampling of Facebook Accounts

- Using Graph Crawling Algorithms:

This method has been largely employed by computer scientists. It is a computer based study; there is no real interaction with actual users. In the paper titled "Walking in Facebook: A Case Study of Unbiased Sampling of Online Social Networks", an algorithm was developed to visit each account exactly once until completion. The crawling of the social graph starts from an initial node and proceeds iteratively. In every operation, a node is visited and all its neighbors (friends) are discovered. Using random walk method, a new node is selected to be visited.

An advantage of such study might be the lacking the human interaction and letting the computer do the job. However, a major disadvantage might be the tendency to have a biased sample as the algorithm starts from one account and proceeds to the "friends" of that account. It also does not focus on US accounts.

- Using Human Power

Another approach might be collecting the data by using real people instead of computers. That is, we might hire some number of students, give them necessary equipment, and start collecting data by randomly selecting people. But in this case there are a lot of difficulties.

1- Random is not defined clearly. Is it randomly putting a name in the search box or is it randomly jumping from one friend to friend of a friend? How random is random? Does it depend on student's perception of randomness? What if, for the sake of completion, a student selects the first possible account that appears on his computer?

2- There are people who limit the publicity of their personal information to "friends only". If our data collectors are not friends with them they can't collect data. However, we are looking for the proportion of people who reveal information to Facebook whether it's for friends only or not. So we need to access those "limited access" accounts, too.

3- Revealing information depends on many attributes. Some reveal their information because they don't care; some don't know whether the information will be stolen or used by third parties, some want to attract opposite sex, some look for a job, some want to show off his friends...etc. Knowing that there are potential attribute clusters or attribute strata in the Facebook population, we can't directly apply Simple Random Sampling to the study.

4- How is that possible to actually confirm that a person resides in the US by just checking the address information? What if it's a fake address?

Based on the difficulties mentioned above, I do not suggest Simple Random Sampling as a sampling method.

Stratified Random Sampling of Facebook Accounts

It would be extremely beneficial to leverage the information we have about the demographics of US Facebook users. Since New York has the biggest share in user pie with almost 3 million users, we can take New York as the starting point for our convenience. Assuming that the gender and age proportions for the overall US data will be applicable to NY numbers we can figure out possible strata for our population.

	NY Users	2,934,580
	Female (54.3%)	1,593,476
	Male (42.6%)	1,250,131
	Unknown (3.03%)	88,917
Stratum 1	Female Age 13-17 (10.4%)	165,721
Stratum 2	Female Age 18-24 (25.4%)	403,149
Stratum 3	Female Age 25-34 (24.8%)	395,182
Stratum 4	Female Age 35-54 (29.0%)	462,108
Stratum 5	Female Age 55+ (9.50%)	151,380
Stratum 6	Male Age 13-17 (10.4%)	130,013
Stratum 7	Male Age 18-24 (25.4%)	316,283
Stratum 8	Male Age 25-34 (24.8%)	310,032
Stratum 9	Male Age 35-54 (29.0%)	362,538
Stratum 10	Male Age 55+ (9.50%)	118,762

According to table above, we have 10 non-overlapping strata based on age and gender of Facebook users in NY. If we take random samples from those strata, there will be small variability among measurements which will produce small bounds of errors of estimation. Thus, for instance, if we assume that all the members in Female Age 13-17 stratum think alike on privacy concerns and personal information release, we can obtain a very accurate estimate of the proportion of Female New Yorkers between ages 13-17, who reveal their private data with a relatively small sample. Similarly, another stratum, let's say Male Age 25-34, might have different concerns on privacy than those of Females Age 13-17 but since they will thinking alike in their own stratum, we can again obtain an accurate estimate with a small sample. When results of stratified random samples are combined, the final estimate of the proportion of the New Yorkers who reveal their personal information on Facebook may have a much smaller bound on the error of estimation than would an estimate from a simple random sample of comparable size.

Although stratified random sampling seems a better fit than simple random sampling, it also has its own difficulties. For example, we chose New York as a starting point while creating our strata. While this is a convenient sample for us since we are in NY; we apply an ill assumption that New York would a good representative for the rest of US. That's not true. NY differs from the rest of the US with many aspects.

Another disadvantage of stratified random sampling would be the difficulty of creation of and accessing to a sampling frame. Assume that we created our stratus of Female Age 13-17 Facebook users in NY. How are we going to find members of this stratus?

Cluster Sampling of Facebook Accounts

There is one major difference between the optimal construction of strata and the construction of clusters[5]. Strata are to be as homogenous (alike) as possible within, but one stratum should differ as much as possible from another with respect to the characteristic being measured. With our NY data, we achieve to create homogenous strata for each gender and age categories. Clusters, on the other hand, should be as heterogeneous (different) as possible within, and one cluster should look very much alike another in order for the economic advantages of cluster sampling to pay off.

A very popular block statistics might be useful here. We can define a block and survey households door to door. One problem with this approach might be having potentially homogenous sample while looking for just the opposite. That is, members of the same house might have same privacy concerns or they might influence each other in that way. Hence, although block sampling creates good heterogeneous clusters for age and gender of possible Facebook users, it might be so for attributes.

Proposed Large Scale Survey

In order to estimate the proportion of Americans who reveal their private information on Facebook, I propose a stratified random sampling of Facebook users in NY area. The research might be replicated in other states using the same procedure for creating homogenous strata.

Strata will be created by applying the age and gender demographics of Facebook users in the US to NY State data. There will be 10 strata based on age and gender. A sample will be selected from each strata by using simple random sampling.

Data will be collected by distributing an online survey or distributing a questionnaire. For instance, for strata of male and female between ages 13-17, the best way of collecting data is to go to selected high schools in NY area. Likewise, for strata of female and make between ages 18-24, same questionnaire will be distributed in college and/or grad schools.

A possible questionnaire should be short and include at least the following questions:

Sample Questionnaire:

1- Do you have a Facebook account? – Yes or No. [If No, no further question]
2- Do you have personal information (listed here) in your Facebook account? –Yes or No
3- Gender
4- Age

{Additional questions for further investigation}

5- Have you ever put personal info (listed here) on your account? Yes or No
6- Are considering putting personal info (listed here) on your account in the future? Yes or No

[5] Elementary Survey Sampling, Schaffer, Mandenhall & Ott

Estimation of the Population Proportion

As explained in previous sections, each stratum was determined by applying US demographics for Facebook users into New York data. We have 10 strata with different age and gender specifics. A sample size for each stratum was determined proportional to the population size of the stratum. Below table summarizes the strata data.

All calculations are based on the equations given by "Elementary Survey Sampling", Third Edition, Scheaffer, Mendenhall &Ott, Chapter 5, Stratified Random Sampling.

Summary of Strata

Stratum	Definition	Population Size	Sample Size
Stratum 1	Female Age 13-17 (10.4%)	165,721	100
Stratum 2	Female Age 18-24 (25.4%)	403,149	300
Stratum 3	Female Age 25-34 (24.8%)	395,182	200
Stratum 4	Female Age 35-54 (29.0%)	462,108	300
Stratum 5	Female Age 55+ (9.50%)	151,380	100
Stratum 6	Male Age 13-17 (10.4%)	130,013	100
Stratum 7	Male Age 18-24 (25.4%)	316,283	300
Stratum 8	Male Age 25-34 (24.8%)	310,032	300
Stratum 9	Male Age 35-54 (29.0%)	362,538	300
Stratum 10	Male Age 55+ (9.50%)	118,762	100
	TOTAL	2,815,172	2,100

Proportion of People who Answers Yes to Questionnaire (Reveals info on Facebook*)

Stratum	Population Size	Sample Size	Reveals info on Facebook*	Proportion, $\hat{p}$
Stratum 1	165,721	100	60	0.60
Stratum 2	403,149	300	120	0.40
Stratum 3	395,182	200	75	0.38
Stratum 4	462,108	300	230	0.77
Stratum 5	151,380	100	60	0.60
Stratum 6	130,013	100	30	0.3
Stratum 7	316,283	300	150	0.5
Stratum 8	310,032	300	200	0.67
Stratum 9	362,538	300	65	0.22
Stratum 10	118,762	100	40	0.40

*Numbers for "reveals info on Facebook" are not real, all make up for calculations

Estimated $\hat{p}$ Calculation:

$$\widehat{p}_{st} = \frac{1}{n_{total}}[(n_1.p_1)(n_2.p_2)\dots(n_{10}.p_{10})]$$

$$\widehat{p}_{st} = \frac{1}{2{,}100}[(100x0.6)(300x0.4)\dots(100x.0.4)] = 0.49$$

Variance of $\widehat{p}_{st}$:

$$\hat{V}(\hat{p}_1) = \left(\frac{N_1-n_1}{N_1}\right)\left(\frac{\hat{p}_1\hat{q}_1}{n_1-1}\right)$$

$$\hat{V}(\hat{p}_1) = \left(\frac{165{,}721 - 100}{165{,}721}\right)\left(\frac{0.6x0.4}{100-1}\right) = 0.00242$$

Variance of Estimated Proportions for each Stratum

$\hat{V}(\hat{p}_1)$	0.00242
$\hat{V}(\hat{p}_2)$	0.00080
$\hat{V}(\hat{p}_3)$	0.00118
$\hat{V}(\hat{p}_4)$	0.00060
$\hat{V}(\hat{p}_5)$	0.00242
$\hat{V}(\hat{p}_6)$	0.00212
$\hat{V}(\hat{p}_7)$	0.00084
$\hat{V}(\hat{p}_8)$	0.00074
$\hat{V}(\hat{p}_9)$	0.00057
$\hat{V}(\hat{p}_{10})$	0.00242

$$\hat{V}(\widehat{p_{st}}) = \frac{1}{n_{total}^2}\sum_{i=1}^{10} n_i^2 \hat{V}\,(\widehat{p_i})$$

$$\hat{V}(\widehat{p_{st}}) = \frac{1}{2{,}100^2}\,\{(100^2x0.00242) + \ (300^2x0.0008) + \cdots\} = 0.0001$$

Then, the estimate of proportions of New Yorkers who reveals their private information on Facebook account, with a bound of error estimation, is given by:

$$\widehat{p_{st}}\ \pm 2\sqrt{\hat{V}}(\widehat{p_{st}}) = \ 0.49\ \pm 2\sqrt{0.0001} = (0.47, 0.51)$$

Facebook – Beauty or the Beast?

Based on calculations with make-up numbers, we conclude that around 50% of New Yorker Facebook users reveal their private information on the Facebook. With a simple math; if there were 3 million users in New York as claimed in the Facebook statistics, then it would make 1.5 million potential NY identities waiting to be stolen.

Facebook may be a good way of getting in touch with old friends, keeping up with current ones, finding a date or even a job. For that matter it is a beauty. But it should be remembered that when there is enormous number of information floating freely on a not-so-secure channel, then there is always someone waiting for it to come to daddy. Having all of our information and innocent feelings about getting in touch with friends without a secure platform, Facebook is a Beast.

This study should definitely be conducted to see the real trend in people's security perception and also to set an example to those who either think that revealing information on Facebook is safe or have no idea whatsoever.