ADS IN FACEBOOK

By

LAVANYA PARAVASTU PATTARABHIRAN

A thesis submitted to the

Graduate School – New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Computer Science

written under the direction

of Muthu Muthukrishnan

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2012

ABSTRACT OF THE THESIS

Ads In Facebook

By LAVANYA PARAVASTU PATTARABHIRAN

Dissertation Director:

Muthu Muthukrishnan

Understanding the relevancy of user-targeted ads on social networking websites requires the information about the heterogeneous user data and the corresponding ads that are targeted towards the user. This can be accomplished by performing a large scale collection of ads seen by users online on various social networking sites. For the purpose of this research, Facebook was considered as the social networking website. Since using real user profiles may involve privacy issues, dummy profiles were created and the targeted ads were analyzed. Moreover, in order to limit the amount of data being collected and analyzed, the study targeted Facebook ads from a region, India. An automated script in java and selenium was also developed to collect the ads that were

targeted at the dummy profiles created. The data collected was analyzed using General Linear Modeling to understand the dependency between category of advertisement shown to a user and profile characteristics like age, gender and location.

Acknowledgements

Thanks to all the people whose help and support was essential in making this work possible. None of this would have happened without my advisor MuthuMuthuKrishnan .

I would also like to thank my other committee members: Badri Nath, and Apostolos Gerasoulis for their advice and encouragement at various stages of this process.

Thanks go also to the many friends and family members who have supported me through this process. I wouldn't have survived it without their support. Thanks also to all the supportive people in the CS Department for being so helpful.

Contents

Introduction

Research Objective

Generating targeted advertisements on social networking websites is an interesting problem looking at the phenomenal growth of websites such as Facebook over past decade. The $240 million investment by Microsoft in Facebook asserts that targeted advertising for online social networks as a very lucrative venue. From advertising point of view, Facebook acts medium for self-expression and opinion sharing which lets the users 'declare' their interests and also offers many cues in the form of user generated content, the links in the network, demographic information. Though there has been some research done in this area, there is still a lack of successful methods for combining the user related heterogeneous data to model interest and relevance in the realm of advertising. The current research is an attempt to understand the relevancy of ads for users on social networking websites, based on the heterogeneous user data with distinct but unknown importance using Facebook as an example in India.

About Facebook

Facebook, the largest social networking website in the world was started by Harvard University Students in 2004. Currently Facebook is valued at $102.8 Billion and the initial public offer (IPO) of Facebook is expected to be $44.10 for 150,000 units. [1] The growth of Facebook is unprecedented in the history of Internet. As of September 2011 Facebook had over 800 Million users. [2] With the high growth in user base, Facebook was no longer an alien to the advertising world as its initial days. Facebook overtook search engine Yahoo! Inc. in 2011 and grabbed the biggest share of online display advertising market in United State. Many market researchers claim based on the current growth of Facebook that it might overtake Google in the world market in near future. [3]

Facebook India

Facebook's vice president for mobile partnerships and corporate development Vaughan Smith mentioned that Facebook expects its largest user base to come from India in the near future.[4] According to social bakers, currently there are approximately 45 019 840 Facebook users in India. An interesting fact is that though India has second largest Facebook user count it accounts for only 3.84% of Indian population and just about 55.58% of online population penetration. With Twenty-three Indian cities among the top 120 cities in Facebook user statics, Facebook India is an interesting study. [5]

a. The age distribution of Facebook users from India (source SocialBakers.com) shows that the major chunk of Facebook users from India belong to age group 18-24 and 25-34. Namely,

Age	% of Users
18-24	49%
25-34	28%
16-17	8%
35-44	7%
13-15	4%
45-54	3%
55-64	1%
65-0	0%

a. The demographic data of Facebook users from India would be helpful to understand the geographical targeting of the ads placed. Four cities in India were considered for “location” parameter based on the cities having the highest number of users. Namely, Mumbai has 3,672,500 users, Delhi has 1,566,500 users, Bangalore has 1,526,080 users and Chennai has 1,214,100 users as of February, 2012. [5]

b. The Facebook user data from India (socialbakers.com) shows that males dominate the user base with 73% as compared to female 27%. [5]

Reason for advertising on Facebook

Facebook, the world's largest social networking site, gets action with over 845 million monthly active users, 483 million daily active users on average in December 2011[6], is the most trafficked web site in the world.[7] The availability of wealth of information about the users makes it highly favorable advertising platform.

Considering the fact that an average user spends approximately 46 minutes per day [6] on Facebook the reach of this platform for interaction between other buyers or consumers shows the need for business to concentrate to use Facebook to reach their target audience.

Facebook has grown from a collegiate social network into a marketing site that gets action from 12.1% [9] of the world's population. It is estimated that advertisers spent $4 billion on Facebook advertising in the year 2011. [8]

Chronology of Facebook Advertising evolution

Source [11]

February 2004- Mark Zuckerberg with his college roommates and fellow students Eduardo Saverin, Dustin Moskovitz and Chris Hughes founded Facebook. The Web site's membership was initially limited by the founders to Harvard students.

August 2005- Facebook changed its name from 'thefacebook.com' to 'Facebook' after purchasing the domain name 'facebook.com' in 2005 for $200,000.

September 2005-Facebook expanded by adding high school networks.

October 2005- Photos were added.

May 2006 -Facebook added work Network.

August 2006 - Facebook crossed 100 Million users. Facebook development platform was launched. Facebook announced partnership with J.P.Morgan Chase to promote the Chase Credit Card. In a one-year marketing agreement Facebook members saw banner ads inviting them to join special Chase network-members, to earn reward points for their actions, like paying their bills on time. Facebook and Microsoft formed strategic relationship for banner ad syndication. Microsoft's adCenter became the exclusive provider of banner ads and sponsored link.

September 2006 - Facebook allowed anyone to join. Facebook announced "Election 2006" which allowed anyone to search for and interact with office candidates for the Senate, House and Governorship. News feed is introduced.

November 2006 - Share feature was added.

February 2007 - Virtual gift shop was launched.

May 2007 - Facebook launched marketplace app for classified listing. Facebook platform was launched with about 85 applications.

August 2007 - Facebook offers an opt-out feature that lets advertisers prevent their ads from showing up.

October 2007 - Microsoft purchased 1.6% share of Facebook for $240 million, giving Facebook total implied value of around $15 Billion.

November 2007 - Facebook introduced “Facebook Ads” pages for brands and businesses, Facebook Insights and a controversial ad system called “Beacon” that encouraged the virtual spread of brand messages.

April 2008 - Facebook launched chat.

August 2008 - Facebook launched Engagement ads.

December 2008 - Facebook connect became generally available.

February 2009 - Facebook added the Like button. Facebook transferred ownership of the Marketplace classified listing app to Oodle.

March 2009 - Facebook introduced language and radius-based ad targeting. Facebook relaunched the pages to be more like profiles and included status updates and photos.

June 2009 - Facebook launched self-serve ads for pages and events, giving them engagement capability.

July 2009 - Facebook launched connection targeting, multiple country targeting and birthday targeting.

September 2009 - Facebook began testing ads API. Facebook shut down ad platform

"Beacon", which posted updates to Facebook profile when their owners interacted with its partner sites. The feature inspired a class action lawsuit after privacy advocates rallied against having their actions on sites like Blockbuster, Gamefly and Overstock.com posted to their profiles. Nielsen launched Brand Lift with Facebook at Advertising Week. The product measured the effectiveness of ads on Facebook by polling users.

April 2010 – David Fischer, VP of Advertising and Global Operations joined Facebook.

August 2010 – Facebook launched Places.

September 2010 – Facebook added social context metrics to its performance advertising analytics.

November 2010 – Second Market, Inc valued Facebook for $41 Billion and it became the third largest U.S. Web company after Google and Amazon. Facebook rolled out beta-version of check-in Deals using Facebook Places.

February 2011 – Facebook launched Sponsored Stories in which companies could choose to take certain user actions – such as check-ins or actions within Facebook apps – and feature them in the column on the right side of the News Feed.

March 2011 – Facebook launched Questions for Pages

April 2011 – In an effort to court advertisers, Facebook revealed Facebook Studio, which highlighted interesting work from advertisers. Facebook officially launched Deals, a Groupon competitor.

May 2011 – Facebook introduced a test program that gave credits to users who watch certain ads from third-party ad networks in games.

September 2011 - Facebook crossed 800 Million users mark.

January 2012 – Under the new initiative advertisements will appear in the form of 'stories' or posts about a product in its news feed labeled as 'featured'. Marketers can only pay for these advertisements to appear in a user's feed if the user has already 'liked' the page. And the advertiser does not have the option to add one's own additional message once the post is live.

February 2012 – Graph Science, a Social Marketing company, joined hands with Facebook which aims to create targeted ads on Facebook by analyzing the social data using Facebook API. Also unveiled was Reach Generator, an Advertising solution which would allow the brand to pay the fees on an ongoing basis instead of the existing CPM or CPC model. Also a new form ads know as offers were introduced, which are a form of sponsor posts.

Advertising on Facebook

Facebook overtook search engine Yahoo! Inc. in 2011 and grabbed the biggest share of online display advertising market in United State. Facebook earned around $2.19 billion in display ads sales during the year of 2011, for a 17.7 percent share of the U.S. market, Yahoo was second with 13.1 percent. [3]

Advertising comes in different flavors on Facebook namely:

Targeted Facebook Ads: Facebook pages have between zero and at least six sponsored ads on the right hand side of a page targeted specifically to the user. Sometimes there is information in the source for other ads not displayed to the user.

These ads are in many different forms as Video Ads, Sponsored Stories, Gifts Ads, Event Ads, Page Ads, Website Ads and internal Facebook ads[10] promoting Facebook features.

- Video Ads work somewhat differently on Facebook then on television. On television advertiser pays for a time slot generally 30 seconds, but on Facebook there are no such timing restrictions. Instead advertisers selects target demographic.
- Page ads are created and published to create higher fan base for Facebook Pages. It allows the users to be the fan of the promoted page and posts on the wall for friends to see it.
- Sponsored stories are there to retrieve items in newsfeeds. The possible actions include: Likes, Check-ins, Posts App usage. This format of ads gives advertiser

less control over the message, but may increase trust in the ad. Click through Rates have been 46% higher than standard ads, and Facebook tends to charge 10% less per click.

- Event Ads are integrated with the Facebook Events. Users can "RSVP" the event. Full event details are available to the users and friends of the user can see the RSVP. Friends' responses are also visible to the use.

Sponsored Poll Ads: Originally Sponsored Poll Ads were open to users to define polls and specify target polling audience but Facebook removed this option several years ago, and limits the availability of this type of ad to larger clients.

Literature Review

Research that aimed at studying or evaluating the relationship between the profiles of users on social networking websites and online advertising in them is not new albeit still is a field of great interest. The potential for exploiting the information derived from actions performed by social networking users, in order to target relevant ads towards them is of immense interest to advertisers and researchers. This section reviews four pertinent papers in this area that throw light on the previous research conducted.

The paper "Combining Behavioral and Social Network Data for Online Advertising" [12] (2008) by Bagherjeiran and Parekh measured the relevance of a social network, the Yahoo! Instant Messenger graph, to classes of ads. To improve response predictions, an ensemble classifier that combines existing user-only models with social network features was proposed. The term 'homophily' was coined for the first time, which states how acquainted users and 'friends' tend to have similar interests. This specific study had two sets of data, the first set - classified as behavioral source - is the web browsing data collected from yahoo websites, web searches, and views and clicks on ads. The other set was data collected from the Yahoo IM for a specific duration of 5 weeks, and the number of recorded conversations between users that determine the user neighborhood. For behavioral source data, the user vector is represented using the following vector.

$u = (xu, cu, vu, p^(u), \varepsilon u)$

where,

xu is behavioral feature vector containing the number of occurrences of web-browsing event

cu is the number of clicks on the ad

vu is the number of views of the ad

p^(u) is the output of the user model

εu indicated the confidence in the score.

The paper combined behavioral data p^(u) and social data p’(u) by using a weighted combination of scores, decided by the confidence function α(u).

Final probability p~(u) = α(u) p(u) + (1 - α(u)) p `(u)

where α(u) = (εu)/(εu + ε), ε is the constant capturing the default confidence and is empirically determined as 1. The behavioral p^(u) is determined by the historic activity of the user. The social version combines the user's score with an average of the friends’ scores. This method does not give the best of results and hence author employs machine learning techniques like ensemble classifier. This classifier calculates the social score based on the history of the user, trust factor of the user with respect to the other users, similarity between ads seen by users, gender and age of all users, total clicks and views of an ad. The trust factor mentioned here is calculated by training a logistic classifier on the training dataset. For the social score, gs is trained to predict the click or not click features of the user. A trust model based on the parameters as explained above would give us the

score each friend in the users neighborhood. Only scores with higher or lower values are considered. A high score would score that a user would click the page and a low score would suggest that user would not click the page and hence provide us with definite probabilities of a user clicking the page. The user only classifier is calculated using the below formula.

$gu(xu) = \sigma (\mu 1 \log p'(\mu) + \mu 2)$

where $\mu 1$ and $\mu 2$ are the parameters of the logistic model and σ is the logistic function. The ensemble classifier would help in merging the social and the user models. But when the user-only model is expected to make an error, which is expected as per Boosting, the social model is only employed. Now the gating classifier would select the best classifier for use

$g(gu, gs) = \sigma (\mu 1 \log gu + \mu 2 \log gs + \mu 3)$

$w = g(gu, gs)$

The weight increases as the error in the user only model score increases. The gating classifier is trained to predict clicks based on an independently drawn validation set

$p\sim'(u) = wgu + (1–w)gs$

where $p\sim'(u)$ is the final probability which would tell us the certainty with which the user would click on a given ad. Ensemble classifier never gives out worse performance when compared to any single model and also across several ad categories gives us a 5% improvement. This paper gives us a systematic way of finding the CTR, but this method is tedious. The paper draws probabilities between similar pair users and random pair users and concludes that users have high connectivity in the social circles and that similar

friends tend to see similar ads and have similar ad click patterns. The paper shows that the probability the user would click an ad tends to increase if the friends of that user would have responded to the ad.

On similar lines, later in 2011, Kun Liu, Lei Tang from Yahoo released their paper, "Large-Scale Behavioral Targeting with a Social Twist" [13], which dealt with the potential of leveraging one's friends' activities for behavioral targeting and compared the forecasts derived from such social information and those from standard behavioral targeting (BT) models in terms of accuracy. It concluded that appending social data and behavioral targeting is the most effective and scalable way to go ahead. A wide array of supervised and unsupervised machine learning algorithms using Hadoop MapReduce framework were performed to observe the effects of incorporating the social targeting to standard behavioral targeting.

Extensive experiments were conducted on users with different online activities covering over 60 consumer domains and 180 million users for a period of about two and a half months. The data was mainly divided into behavioral data and social data. The behavioral data was in-turn sub-divided into training data and test data. The training data was collected over the period of 10 weeks (2010/08/23–2010/10/31), with last four weeks (2010/10/4–2010/10/31) for generating targets. The test data was generated over seven weeks (09/20/2010–2010/11/07), with last one week (2010/11/01–2010/11/07) used to form targets. This process enabled authors to produce 13 billion training and test examples and approximately 7 terabyte of data. The social data comprised of social graph

constructed using instant messaging network operated by a large IT company. Removing singleton users and connecting all pairs of users who mutually authenticated each other as buddies resulted in over 390 million nodes and 5 billion edges. Intersection of behavioral data and communication network resulted in about 180 million users. The following formula was developed to ensemble BT with Social Model.

$S_{ensemble} = \alpha S_{behavioral} + (1-\alpha) * S_{social}$

Where $\alpha \in [0, 1]$ is weighting parameter.

But due to high computational costs, this approach was not considered scalable. A network-propagation method was also considered to infer users' BT scores directly from their friends. For this, a three iteration approach was used.

Ite1: $s(t)(u) = (1 -\alpha)\ vN(u)\ s(t-1)(v)d(v) + 1G$

Ite2: $s(t)(u) = (1 -\alpha)\ vN(u)\ s(t-1)(v)d(v) + s(0)(u)$

Ite3: $s(t)(u) = (1 -\alpha)\ vN(u)\ s(t-1)(v)d(v) + 1G + s(t-1)(u)$

But it was found that the propagation does not increase the accuracy of the prediction accuracy of the baseline models but interesting details were observed from the data generated which led to the conclusions. The entire experiment was conducted on Hadoop MapReduce platform and the paper provides pseudo-code for both approaches. Homophily was tested in different profiles using BT quantification and ad clicks and these tests concluded that social data can help Behavioral Targeting. Categories with a strong homophily effect are more likely to benefit from social data, but the degree of

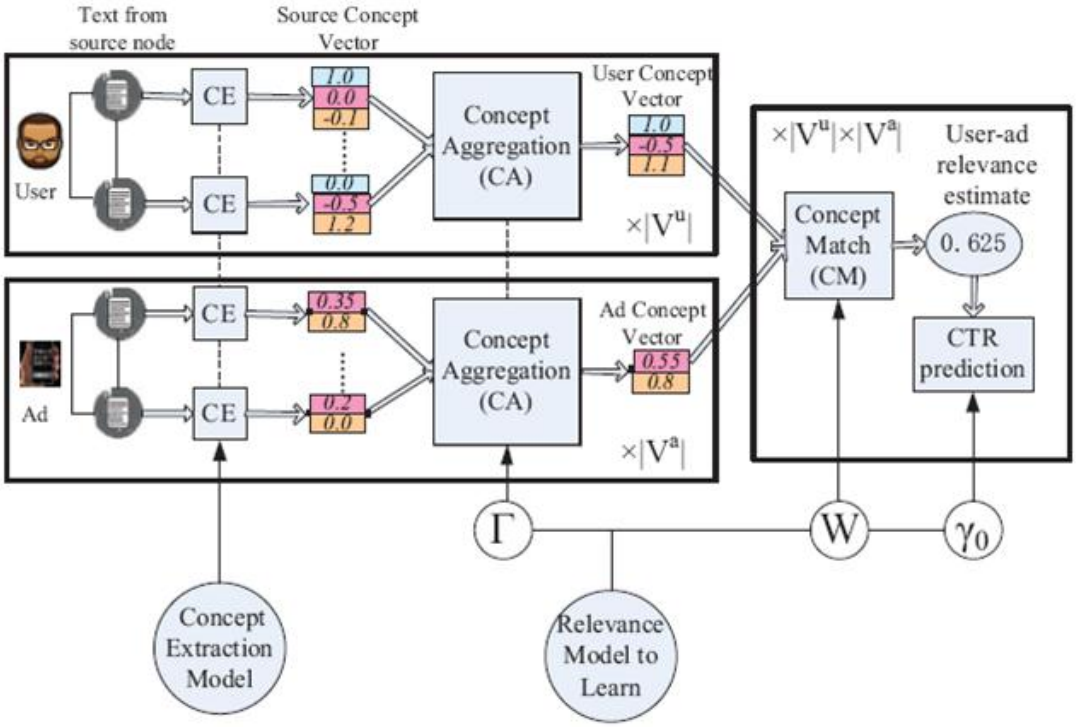
improvement depends on the amount of behavioral information the targeted users have, and how strong the baseline is.

The paper “Learning Relevance from Heterogeneous Social Network and Its Application in Online Targeting” [14] by Chi Wang et. al. introduced a new algorithm for modeling user interests from the heterogeneous data sources such as text in user generated content, links in the network and other demographic information of the user to aid the task of selecting relevant ads for targeting users on Facebook. This algorithm was used in designing a framework for CTR prediction based on a variant vector space model for an improved click through rate in online advertising.

The authors demonstrated that their model could find hidden associations between user concepts and ad concepts. Their hypothesis also stated that their jointly learnt user model is better for targeting than demography and keywords matching. In the vector space model, semantically meaningful concepts were used to abstract the interest of the users and the topic of ads for dimensionality reduction. Both the user and ad were represented by a concept vector where each component was the weight of corresponding topic. The authors utilized a multi typed network to characterize user interaction with pages, groups, and other entities. The nodes in this network had text content from which concepts could be extracted, and served as concept sources to their linked objects. Thereafter user interests and other target content were summarized from the concept classes of linked source nodes. A generic user model was developed to learn the weight for heterogeneous

linked sources and the association between every pair of user concept and target concept.

Conceptual Overview:



The above diagram outlines the pipeline for the relevance prediction. The various terms are defined as below:

- Concept Space. A concept space is a d-dimensional real number space Rd that encodes some cognitive classification system such as ontology. Each dimension represents a concept class in the system. For example, in a human edited Web directory, each category in the directory can be used as a concept class.
- Concept Vector. A concept vector is a vector in the concept space defined above that represents the cognitive property of an object. The component in each dimension indicates how strong the object is associated with the corresponding concept. For example, (Movie: 0.5, Soccer: 0.3) is a short representation of a

sparse concept vector where all the weights are zero except for two concepts. For generality we allow user concept vector and ad concept vector to be in different concept spaces.

- Concept Extraction (CE). Given a source node's associated text, output its concept vector c in certain concept space.
- Concept Aggregation (CA). Given a user or an ad vx, and the set of linked source nodes Sx—for each node vi Є Sx, the corresponding concept vectors are already extracted as ci—output a concept vector c for this user or ad.
- Concept Matching (CM). Given a pair of user concept vector u and ad concept vector a, produce a real value feature f to predict the likelihood of the user clicking the ad.

This pipelining model allowed extraction and aggregation of the user concept vector and ad concept vector offline, and also allows it to be stored in high speed medium.

Pipeline: CE → CA → CM

With the fast access to the concept vector and fast vector matching algorithm, the online feature computation can be efficient. A concept model is first extracted, then parameters are found and the model is then tested on the available click data.

Baseline for CA and CM:

To compute the baseline which would give us the relevance, we have to collect the output of the extraction model and calculate the cosine similarity of the vectors. This is done by taking the unweighted sum of all concept vectors for every user and ad. The formula is as presented below

$$\text{Cos}(U, A) = ((\Sigma i\ Ui)\ T\ (\Sigma j\ Aj\ )) / (\| \Sigma i\ Ui \| \| \Sigma j\ Aj \|)$$

The paper however states several limitations to this baseline feature. The first one being Concept Aggregation - different sources should have different weights. Another improvement in the metric can be achieved by assigning different weights to different concept types. So a better measure should allow one user concept to match different ad concepts and vice versa, and does not require they are in the same concept space.
The paper “Social Networks, Personalized Advertising, and Perceptions of Privacy Control” [15] by Catherine Tucker, MIT, is the first paper to our knowledge, which discussed how the advertising on Facebook was affected based on the privacy controls given to users. The author conducted experiments which confirm that reactance is reduced for highly personal advertising if the consumers perceive they have control over their privacy.

The first analysis in the paper compared the average click through rate before and after the introduction of improved privacy controls. Ads in which content was personalized appeared to have become more effective with a highly significant change (p-value =

0.0047) after the introduction of improved privacy controls. On the other hand, ads that did not personalize the content did not see much change in their effectiveness before and after the privacy controls were introduced (p-value = 0.407). In order to check the robustness of this claim and to check the statistical significance of the results, regression analysis was performed. The click-through rate $ClickRate_{jt}$ for ad j on day t was modeled in the following manner:

$$ClickRate_{jt} = \beta Personalized_j \text{ X } PostPolicyt + \alpha Personalized_j + \theta MediaAttention_{jt} + \gamma k + \delta t + \varepsilon j \quad \textbf{(1)}$$

$Personalized_j$ is an indicator variable which is equal to one if the ad contained personalized content matched to the variable on which it was targeted, and zero if there was no personalized content. $PostPolicy_t$ is an indicator variable equal to one if the date was after the privacy-settings policy change took place, and zero otherwise. The coefficient captures the effect of their interaction. θ captures the effect of various controls we introduce to allow the effectiveness of personalized advertising to vary with media attention. γk is a vector of 39 fixed effects for the 20 different undergraduate institutions and each of the 19 celebrities targeted. These control for underlying systematic differences in how likely people within that target segment were to respond to this charity. The regression coefficients are estimated using ordinary least squares method. The crucial coefficient of interest is Personalized X PostPolicy. This captures how an individual exposed to a personalized ad responds differently to a personalized ad after

Facebook's change in privacy policy, relative to an ad shown to the same people that had generic wording. It suggests a positive and significant increase in the performance of personalized ads relative to merely targeted ads after the introduction of enhanced user privacy controls. The negative coefficient of Personalized, which is marginally significant, suggests that prior to the change in privacy settings, personalized ads were less effective than ads that did not use personalized ad copy.

Next, a logit model was used to model an individual's likelihood of clicking on an ad after the introduction of improved privacy controls. One advantage of an individual-level model is that the non-targeted campaign in the regressions can be included as the baseline. Rather than one observation of a click-through rate of the non-targeted campaign which is collinear with the targeting group fixed effects, there are hundreds of thousands of observations of how individuals responded to that campaign. The probability that an individual i clicked on ad j on day t was modeled as:

$$ClickRate_{jt} = I(\beta_1 Personalized_j \text{ X } PostPolicy_t + \beta_2 Targeted_j \text{ X } PostPolicy_t + \alpha_1 Personalized_j + \alpha_2 Targeted_j + \theta MediaControls_{jt} + \gamma k + \delta t + \varepsilon j) \quad (2)$$

Equation (2) is similar to Equation (1), except for the inclusion of a new indicator variable $Targeted_j$. $Targeted_j$ is an indicator variable for whether the ad was targeted, but had no attempts at personalization. For such ads, it would have been difficult for the consumer to know why they received that ad. Not controlling for media effects, personalized ads performed worse than non-personalized ads before the policy, but

performed twice as well after the policy. There was no significant shift in the efficacy of targeted ads before and after the policy.

One of the explanations provided for the increased efficiency was reduced level of reactance of users to personalized advertising under the improved privacy environment. To explore this hypothesis in an empirical setting, the equation (1) was modified to include additional parameter called Ad-Reach that signifies the reach of the ad or the number of users the ad could have been potentially sent to.

$$ClickRate_{jt} = \beta_1 Personalized_j \text{ X } PostPolicy_t \text{ X } AdReach_k + \beta_2 Personalized_j \text{ X } PostPolicy_t + \beta_3 PostPolicy_t \text{ X } AdReach_k + \alpha_1 Personalized_j + \alpha_2 Personalized_j \text{ X } AdReach_k + \theta MediaControls_{jt} + \gamma k + \delta t + \varepsilon j \quad (3)$$

The negative coefficient on $Personalized_j$ X $PostPolicy_t$ X $AdReach_k$ suggests that the positive effect is smaller for ads that had a larger ad-reach than those that had a smaller ad-reach. In other words, personalization was relatively more successful after the introduction of privacy controls for celebrities who had smaller fan bases or schools with smaller numbers of graduates on Facebook, which can also be seen by the larger point estimate for $Personalized_j$ X $PostPolicy_t$ than estimates from previous models.

Though the paper had a limitations like the data considered was from an NPO with an appealing cause, the authors were not sure how long the positive effects persisted this paper was the first of a kind to examine advertising on social networks by external firms.

Empirical analysis suggested that after changes in privacy policies, Facebook users were roughly twice as likely to react positively to personalized ad content and click on personalized ads.

Methodology

Design of the experiment

Short background

A design of experiment enables designers to evaluate simultaneously individual and interactive effects of many factors that affect the outcome of the experiment. A design of experiment is also an important step to determine the number of trials required to statistically determine the significance of each of the factors (independent variables) and their interactions on the outcome (dependent variable). For example, if there are 3 independent variables x_1, x_2, and x_3 and a variable y dependent on them. Also let the relationship between the dependent and independent variables be determined by the equation

$Y = ax_1+b\ x_2+c\ x_3$

Simple algebra tells us that in order to find the values of a, b and c, we need three different equations. In an experimental terminology, we need three experiments having different values of the independent variables resulting in three different values of the dependent variable. But what if there are different ranges to the independent variables (or three different values the variable can take) and we wanted to also know the most significant range or value of the independent variable. Each of this range or value is called “level” of the factor. For example, if each of the independent variables has 2

levels, then we will need 2x2x2 = 8 experiments in total to statistically evaluate the significance of each of the factor and its levels. This is called a full factorial design. Number of experiments required = l^f where f is the number of factors and l is the number of levels.

Current research

A full factorial design can be implemented if the experiment is inexpensive to conduct. The current research involves running an automated script on dummy Facebook profiles and hence is inexpensive. Therefore, a full factorial design can easily be implemented for the current study. The characteristics of the dummy profiles are the factors of the experiment and the variations in these characteristics are the levels of these factors. The probability of an ad targeted to their profile is the outcome of the experiment.

The factors and their levels considered in this research are as shown in the table below

Factor	Levels
Gender	2 (M,F)
Age	3 (<18 yrs, 18-24 yrs, >24 yrs)
Location	4 (Mumbai, Delhi, Bangalore, Chennai)

So, at least 2*3*4 = 24 experiments are needed to satisfy the all possible combinations. From here a profile would refer to each individual combination of the factors. Hence there are 24 profiles.

Challenges Faced in Profile Creation

1. Facebook rate had limited account creation from IP addresses. Only 1-2 could be created per IP address per day. As a solution an account with a VPN as a proxy with rotating IP addresses was used for each new account.
2. Unique email requirement: Creating a new, unique email for each account was not feasible. Email creation rate is limited and some require verification. As a remedy a loophole was found to make an individual Gmail account appear as different accounts. Example: Email@gmail.com can appear as two separate accounts to Facebook if formatted as Email+1@gmail.com, Email+2@gmail.com
3. Facebook fixed the loophole and to keep the accounts alive, a domain was bought with email support which forwarded any email to an email address in the domain regardless if it had been created or not to one email account, to create unlimited email addresses on the fly.
4. Creating lots of accounts is a long process and Facebook started cracking down on fake profiles and added authentication like a cellphone number in country of profile etc. Hence the number of profiles was limited to 4 for ease of creation and maintenance.

Data Collection

We ran the experiment for 30 days, collecting data twice each day. Each data entry essentially consists of ：

<”URL of ad collection page” ,Email ID, Gender, Location, Time stamp, Title of Ad, External Link, Ad Information, Image Link, Note of Facebook Ad, Inline text of Facebook Ad>

Data Analysis

The ads that we collected are classified, using semantically meaningful concepts to abstract the interest of the users and the target areas of the ads into the following categories:

No	Categories
1	Financial
2	Classifieds
3	Consultancy Ex: Shaadi.com
4	Education
5	Electronics
6	Food
7	Health
8	Miscellaneous
9	Mobile
10	Games/Play
11	Shopping
12	Online Services Ex:Gmail
13	Sports
14	Travel

15	Entertainment

Each set of this classification is an output value which would be influenced by the input factors, which are the user age, location and gender. So for each profile, the ads generated would be placed into their corresponding set. For each profile, 100 successive data entries recorded during the experiment are considered and these are classified into the above output sets.

Keeping the above classification in context, an table has been constructed which holds the profile information including the user email id, age group, location which was set for that specific user when the data was processed and the output values. The output values here mentioned are the number of ads that fall in that set in the run. It can be sufficiently deduced that it would lead to 15 output values for each profile, each value having the number of ads recorded for that particular user during the run. Since 100 entries were used, a row sum of the output values would give a result of 100

Below is the table with 24 profiles, which is a result 2 profiles per each of the 3 age groups and 4 locations.

Output Table

Name	*Age*	*Gender*	*City/Cat*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
abhimehta@uniquedisplayads.com	< 18	M	Delhi	0	0	0	11	3	7	6	6	9	5	30	5	7	2	9
shilpakhandekar@uniquedisplayads.com	< 18	F	Delhi	0	0	0	2	13	13	1	5	8	9	23	0	9	7	10
anvbandi@uniquedisplayads.com	< 18	M	Mumbai	0	0	0	5	0	10	11	3	7	8	43	0	9	1	3
satyasree@uniquedisplayads.com	< 18	F	Mumbai	0	0	0	15	5	6	0	6	9	5	34	0	11	2	7
mougligoda@uniquedisplayads.com	< 18	M	Chennai	0	0	0	3	4	13	8	0	4	8	46	1	6	6	1
deepikaneela@uniquedisplayads.com	< 18	F	Chennai	0	0	0	0	8	8	1	9	9	10	23	2	6	8	16
rahulbakshi@uniquedisplayads.com	< 18	M	Bangalore	0	0	0	5	11	7	6	0	7	9	38	1	7	3	6
bindutama@uniquedisplayads.com	< 18	F	Bangalore	0	0	0	7	1	7	0	2	7	5	32	0	8	11	20
sreeyara@uniquedisplayads.com	18-24	M	Delhi	0	14	0	0	8	0	0	0	1	1	63	1	0	1	11
sreedevirao@uniquedisplayads.com	18-24	F	Delhi	0	9	0	11	8	4	0	6	0	0	57	0	0	0	5
nikhilgupta@uniquedisplayads.com	18-24	M	Mumbai	0	12	0	0	7	0	0	1	0	1	71	0	0	1	7
chuvidakshan@uniquedisplayads.com	18-24	F	Mumbai	0	5	0	0	7	3	0	2	1	0	77	1	0	0	4
dheerajsharma@uniquedisplayads.com	18-24	M	Chennai	0	14	0	1	18	0	0	2	0	7	52	0	0	1	5
madhurinatool@uniquedisplayads.com	18-24	F	Chennai	0	5	0	9	7	1	0	2	1	1	64	3	0	2	5
akshaytalskhi@uniquedisplayads.com	18-24	M	Bangalore	0	10	0	0	15	0	0	7	0	0	67	0	0	0	1
amishaanoo@uniquedisplayads.com	18-24	F	Bangalore	0	5	0	0	6	0	2	4	0	4	79	0	0	0	0
rishi_patel@uniquedisplayads.com	> 24	M	Delhi	15	15	1	9	12	0	0	4	0	5	32	0	0	0	7
kamalapriya@uniquedisplayads.com	> 24	F	Delhi	0	4	0	18	2	0	0	4	0	0	66	0	0	0	6
harinersu@uniquedisplayads.com	> 24	M	Mumbai	7	12	0	6	15	0	0	6	8	8	15	0	5	6	12
kiranpola@uniquedisplayads.com	> 24	F	Mumbai	2	14	0	8	6	0	0	9	0	7	52	0	2	0	0
sathvisadhu@uniquedisplayads.com	> 24	M	Chennai	0	0	1	6	12	9	0	6	10	9	9	0	2	0	36
harshitaawal@uniquedisplayads.com	> 24	F	Chennai	8	13	0	0	14	0	0	5	0	13	47	0	0	0	0
krishnateja@uniquedisplayads.com	> 24	M	Bangalore	7	9	0	7	26	0	0	1	0	13	27	0	10	0	0
ramyanori@uniquedisplayads.com	> 24	F	Bangalore	3	10	0	5	8	3	0	10	0	7	49	0	0	5	0

General Linear Model

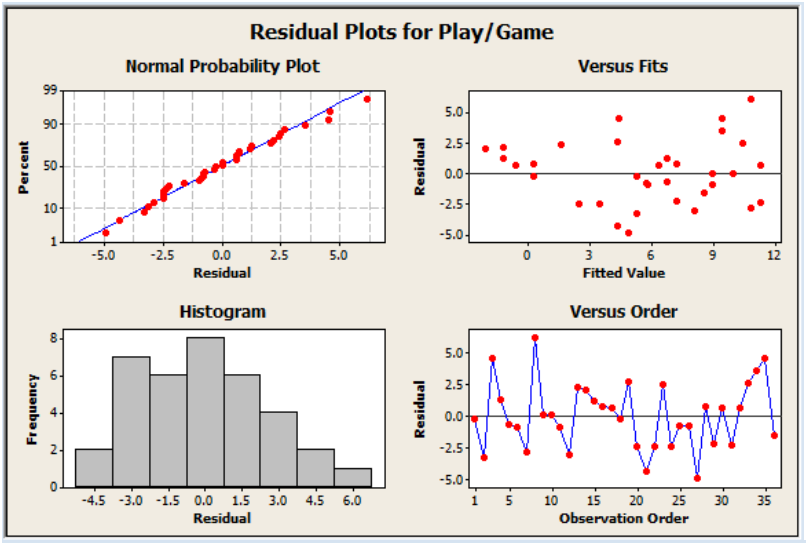
In this work, **general linear modeling** (multiple regression analysis) was performed to understand the relationship between the several independent variables (input variables) and the dependent variable (output term). The independent (predictor) variables considered for the research were Gender (Male and Female), Age (under 18, 18-24 and above 24) and Location (Delhi, Mumbai, Bangalore and Chennai). The dependent variable (output) used is the type of the advertisement. As the ads were classified into 15 types, there are 15 outputs to be considered.

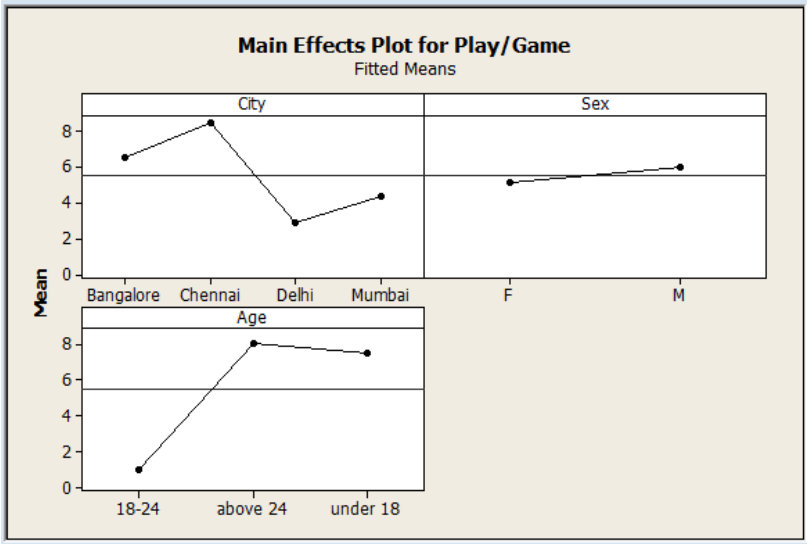
In the research all the 15 possible outputs were considered to find effects on each of them due to the three independent variables used in the experiment.

When the analysis for Play/Game, using Adjusted SS for Tests is performed then the resultant table and graphs are,

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	161.556	161.556	53.852	6.72	0.001
Sex	1	6.125	6.125	6.125	0.76	0.389
Age	2	366	366	183	22.84	0
Error	29	232.319	232.319	8.011		
Total	35	766				

S = 2.83037 R-Sq = 69.67% R-Sq(adj) = 63.40%


Residual Plots for Play/Game
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order


Main Effects Plot for Play/Game
Fitted Means
City
Sex
Age
Mean
Bangalore
Chennai
Delhi
Mumbai
F
M
18-24
above 24
under 18

Checking the Fit of the Model

Residuals

Source [16]

Residuals are calculated by subtracting the observed responses from the predicted responses and so Residual analysis and examination is a key part of all statistical modeling. Since residuals are an estimate of the experimental error, after examining the residuals, we can conclude whether our assumptions are reasonable and the accuracy of our choice of model is within acceptable limits.

Plots of the residuals are constructed to identify any abnormal patterns or atypical data points and we consider the following below:

1) Histogram of residuals versus frequency:

The histogram is a frequency plot obtained by placing the data in regularly spaced cells and plotting each cell frequency versus the center of the cell. Sample sizes of residuals are generally small (<50) because experiments have limited treatment combinations, so a histogram is not be the best choice for judging the distribution of residuals. A more sensitive graph is the normal probability plot.

2) Normal probability plot of residuals:

The normal probability plot should produce an approximately straight line if the points come from a normal distribution. The purpose of the dot plot is to provide an indication the distribution of the residuals. Small departures from the straight line in the normal probability plot are common, but a clearly "S" shaped curve on this graph suggests a bimodal distribution of residuals. Breaks near the middle of this graph are also indications of abnormalities in the residual distribution.

3) Residuals versus fits:

It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

For a well-behaved residual vs. fits plot:

-The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.

-The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.

-No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

4) Residuals versus sequence (time or observation sequence):

It is a scatter plot with residuals on the y axis and the order in which the data were collected on the x axis.

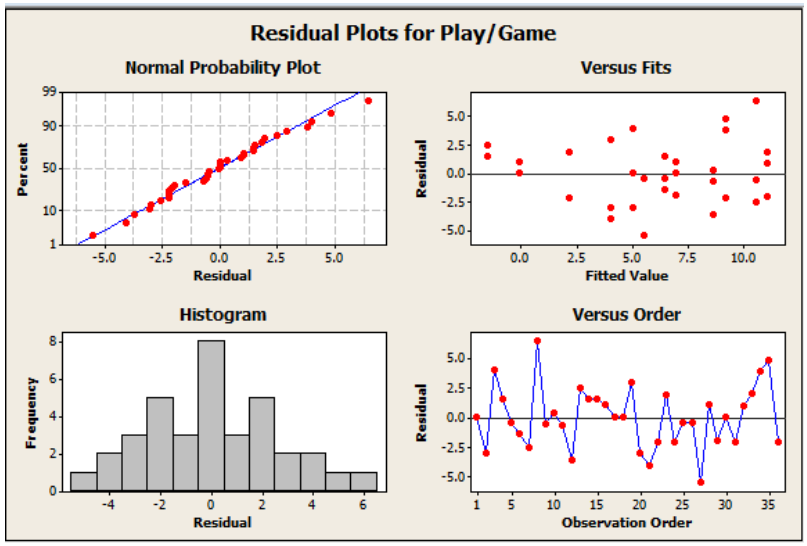
Observation:

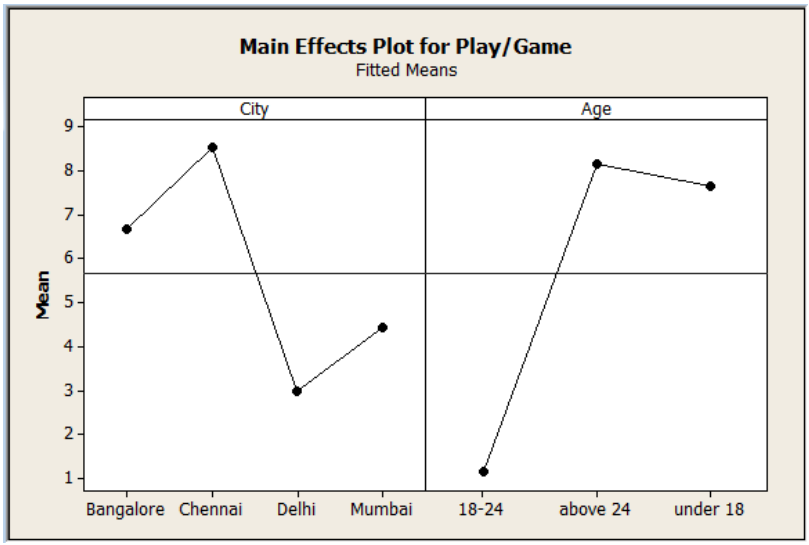
In theory, non-random patterns in the residuals indicate that the model is not adequate. In the current work, the results obtained indicate random patterns for the residual plots, showing that the model considered is appropriate. After the initial data analysis performed considering all the independent variables and then keeping the significant terms in context, a reduction in the model was done to improve the overall prediction rate and the accuracy of the model. Initially the independent variables were considered, i.e. both the two- way and three-way interactions, but these interaction terms observed were negligible and can be quoted as statistically insignificant. Hence a reduced model was employed, where no interaction terms were considered and only the main effects of the process variables are taken into account.

After carefully analyzing the SS table and considering the P values of each of the independent variables we were able to determine that as the P value of sex is greater than 0.05 it is not an important factor. After removing Sex from the list of significant variables and performing the analysis,

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	161.56	161.56	53.85	6.78	0.001
Age	2	366	366	183	23.02	0
Error	30	238.44	238.44	7.95		
Total	35	766				

S = 2.83037 R-Sq = 69.67% R-Sq(adj) = 63.40%





Results

The effects of the 3 design parameters are:

- An analysis of the overall results shows that the characteristic '*Age'* is significant in being targeted by most categories of ads. The categories of ads for which 'Age' is not a significant factor are 'Consultancy' (p-value = 0.135), 'Electronics/Computers' (p-value = 0.412),' Online Services' (p-value = 0.118) and 'Entertainment' (0.171).
- On the other hand, for the category of ads belonging to electronics/computer, health, and shopping, the user characteristic 'Gender' is significant and the relationship between them can also be understood intuitively.
- The characteristic City is significant only for ads belonging to the category of Play/Gaming. This can be justified as all the cities considered in this exercise are metropolitans in India and generally have similar population of Facebook users.

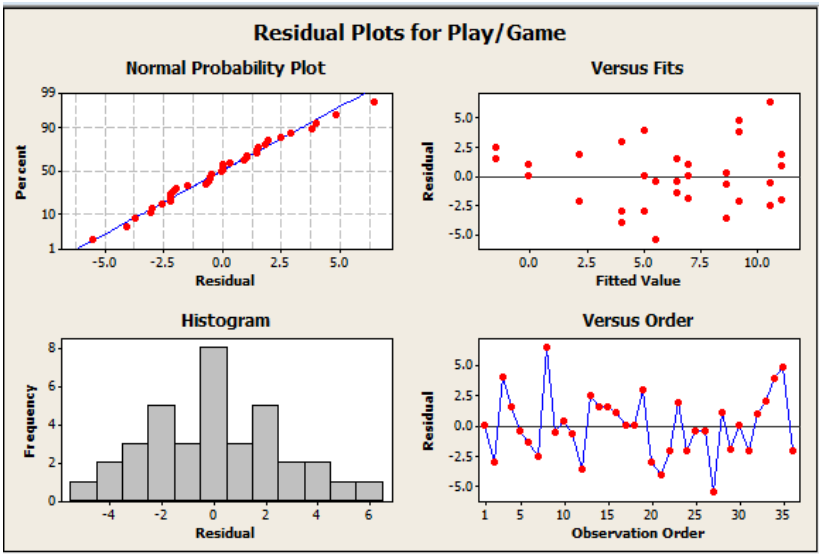
Some of the interesting results observed are discussed in detail below:

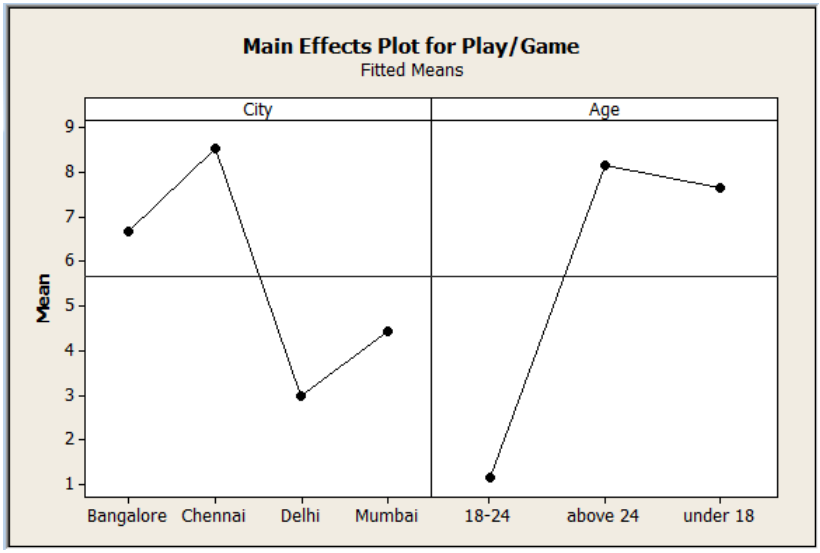
1) The table below shows the relationship between ads relating to **Play/Game** and user profiles. It clearly shows that Age and City with p-values of 0 and 0.001 are highly significant. Age is quite intuitive in nature as people of certain age groups are generally more interested in games or playing than others. But it is surprising to note that the characteristic *City* has such high significance. It shows that users belonging to

certain cities are targeted more with ads relating to games compared to those from other cities.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	161.56	161.56	53.85	6.78	0.001
Age	2	366	366	183	23.02	0
Error	30	238.44	238.44	7.95		
Total	35	766				

S = 2.81925 R-Sq = 68.87% R-Sq(adj) = 63.68%



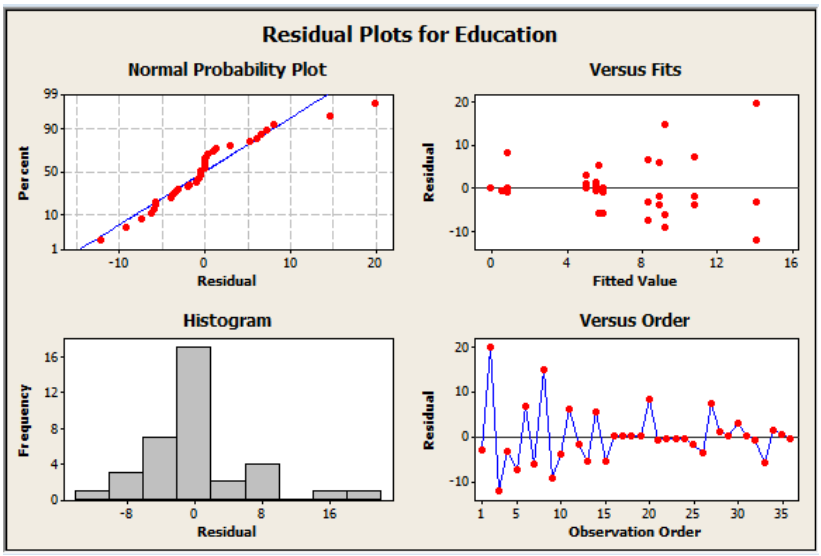


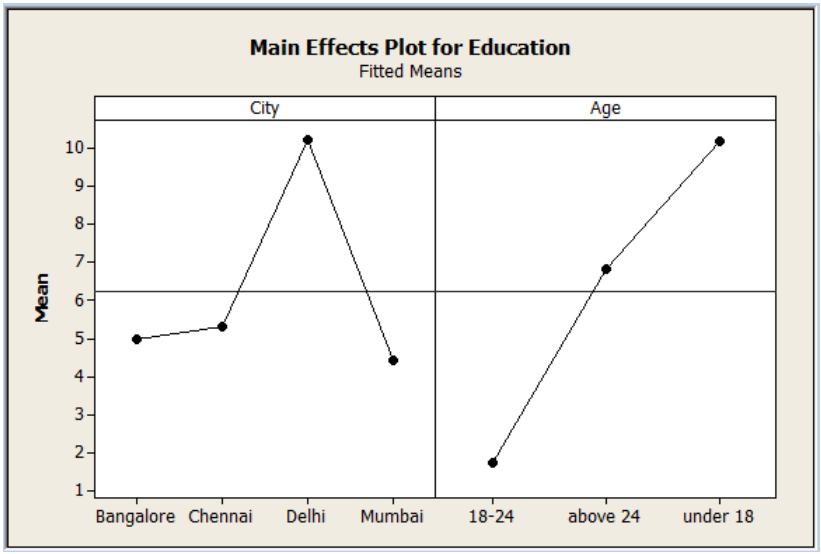
2) The results for ***Education*** related ads for an Indian context look surprising considering the sensitive topic of gender bias with respect to education in India. In India, Percentage of men educated is much higher compared to that of educated women. But the results show a p-value of 1.0 for Gender which implies that gender has no significance when it comes to education-related ads. A deeper look into this suggests that the 4% penetration of Facebook in India is probably assumed by Facebook to consist of a segment of people who have no gender bias with respect to education levels. Location is not a significant factor as expected because all the cities considered for this study are metropolitans in India. Considering smaller towns might have shown different results with respect to targeted ads. Age is only significant at 5% level of significance as a majority of Facebook users in India belong to the age group which is interested in education related ads.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	192.97	192.97	64.32	1.44	0.25

Age	2	431.17	431.17	215.58	4.83	0.015
Error	30	1338.61	1338.61	44.62		
Total	35	1962.75				

S = 6.67985 R-Sq = 31.80% R-Sq (adj) = 20.43%



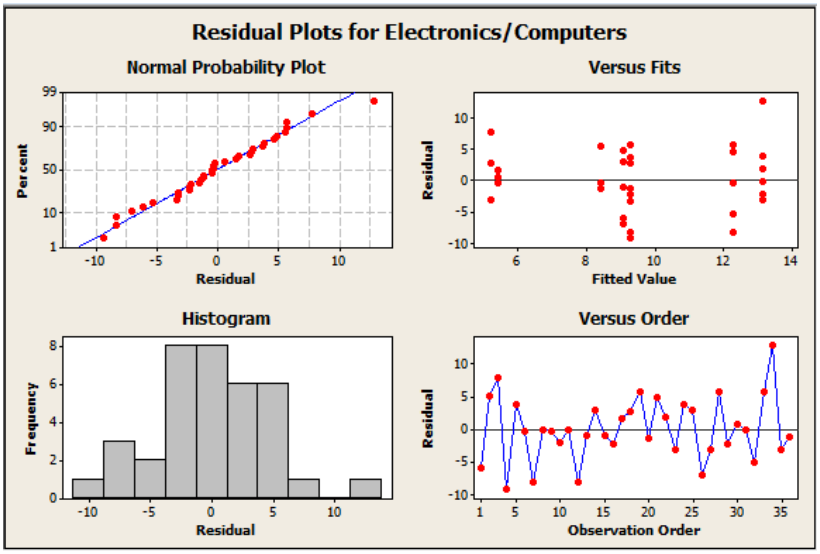


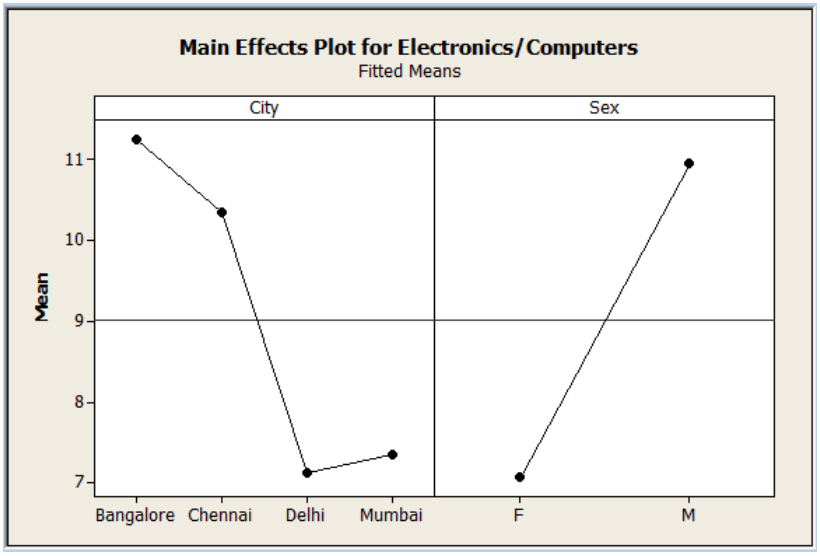
3) Another interesting result comes from the category of **electronics and computers**. It is a common trend that younger generation would be more interested in technologies like electronic gadgets and computers. Therefore it would be expected that Age

would be a significant factor in determining the probability of ads pertaining to this category But surprisingly, the result shows that Age is highly insignificant with a p-value of 0.412. It may be a possibility that in metropolitans of India, people are generally interested in this field irrespective of their age or it might be because a majority of the users are young. Also, it should be noted that the fact that a person uses Facebook in itself implies that the user is at some level interested in computers or at least has basic knowledge in this field. But Gender is a significant factor at 5% significance level which shows that one of the genders (perceived to be males) is more interested in electronics and computers than the other.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	117.56	117.56	39.19	1.48	0.238
Sex	1	120.13	120.13	120.13	4.55	0.041
Error	31	818.32	818.32	26.4		
Total	35	1056				

S = 5.15223 R-Sq = 27.10% R-Sq (adj) = 12.02%





4) But if we look at the category **Health**, it can be again seen that Gender is a significant factor conforming to the perception that women are more interested in health related matters than men.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	39.014	39.014	39.014	5.7	0.023
Age	2	213.722	213.722	106.861	15.62	0

Error	32	218.903	218.903	6.841		
Total	35	471.639				

S = 2.61548 R-Sq = 53.59% R-Sq (adj) = 49.24%

5) One category of ad which is not related any considered characteristic of user is **consultancy**, the results of which can be seen in the table below. The R-sq of the model is also only 0.2.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	0.05556	0.05556	0.05556	1.1	0.301
Age	2	0.22222	0.22222	0.11111	2.21	0.127
Error	32	1.61111	1.61111	0.05035		
Total	35	1.88889				

S = 0.224382 R-Sq = 14.71% R-Sq (adj) = 6.71%

6) Another category of ads that are not related to any characteristic of the user is **entertainment**. Though it can be understood that entertainment ads do not have any correlation to City and Gender, it is surprising to note that Age and entertainment ads are not related.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	105.56	105.56	35.19	0.67	0.579
Age	2	204.5	204.5	102.25	1.94	0.162

Error	30	1583.94	1583.94	52.8		
Total	35	1894				

S = 7.26623 R-Sq = 16.37% R-Sq (adj) = 2.43%

Glossary of Result Tables

1) Analysis of Variance for **Finance**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	9	9	3	0.42	0.74
Sex	1	21.125	21.125	21.125	2.96	0.096
Age	2	338	338	169	23.69	0
Error	29	206.875	206.875	7.134		
Total	35	575				

S = 2.67088 R-Sq = 64.02% R-Sq(adj) = 56.58%

Analysis of Variance for **Finance**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	21.13	21.13	21.13	3.13	0.086
Age	2	338	338	169	25.05	0
Error	32	215.87	215.87	6.75		
Total	35	575				

S = 2.59732 R-Sq = 62.46% R-Sq(adj) = 58.94%

2) Analysis of Variance for **Classified**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	20.44	20.44	6.81	0.52	0.674
Sex	1	2.35	2.35	2.35	0.18	0.676
Age	2	551.06	551.06	275.53	20.9	0
Error	29	382.37	382.37	13.19		
Total	35	956.22				

S = 3.63116 R-Sq = 60.01% R-Sq(adj) = 51.74%

Analysis of Variance for **Classified**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	2.35	2.35	2.35	0.19	0.669
Age	2	551.06	551.06	275.53	21.89	0
Error	32	402.82	402.82	12.59		
Total	35	956.22				

S = 3.54797 R-Sq = 57.87% R-Sq(adj) = 53.92%

3) Analysis of Variance for **Consultancy**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	0.11111	0.11111	0.03704	0.72	0.55
Sex	1	0.05556	0.05556	0.05556	1.07	0.309
Age	2	0.22222	0.22222	0.11111	2.15	0.135
Error	29	1.5	1.5	0.05172		
Total	35	1.88889				

S = 0.227429 R-Sq = 20.59% R-Sq(adj) = 4.16%

Analysis of Variance for **Consultancy**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	0.05556	0.05556	0.05556	1.1	0.301
Age	2	0.22222	0.22222	0.11111	2.21	0.127
Error	32	1.61111	1.61111	0.05035		
Total	35	1.88889				

S = 0.224382 R-Sq = 14.71% R-Sq(adj) = 6.71%

4) Analysis of Variance for **Education**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	192.97	192.97	64.32	1.39	0.265
Sex	1	0	0	0	0	1

Age	2	431.17	431.17	215.58	4.67	0.017
Error	29	1338.61	1338.61	46.16		
Total	35	1962.75				

S = 6.79404 R-Sq = 31.80% R-Sq(adj) = 17.69%

Analysis of Variance for **Education**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	192.97	192.97	64.32	1.44	0.25
Age	2	431.17	431.17	215.58	4.83	0.015
Error	30	1338.61	1338.61	44.62		
Total	35	1962.75				

S = 6.67985 R-Sq = 31.80% R-Sq(adj) = 20.43%

5) Analysis of Variance for **Electronics/Computers**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	117.56	117.56	39.19	1.48	0.242
Sex	1	120.13	120.13	120.13	4.53	0.042
Age	2	48.5	48.5	24.25	0.91	0.412
Error	29	769.82	769.82	26.55		
Total	35	1056				

S = 5.15233 R-Sq = 27.10% R-Sq(adj) = 12.02%

Analysis of Variance for **Electronics/Computers**, using Adjusted SS for Tests after removing Age from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	117.56	117.56	39.19	1.48	0.238
Sex	1	120.13	120.13	120.13	4.55	0.041
Error	31	818.32	818.32	26.4		
Total	35	1056				

S = 5.13784 R-Sq = 22.51% R-Sq(adj) = 12.51%

6) Analysis of Variance for **Food**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	61.417	61.417	20.472	2.22	0.107
Sex	1	0.889	0.889	0.889	0.1	0.758
Age	2	345.722	345.722	172.861	18.78	0
Error	29	266.944	266.944	9.205		
Total	35	674.972				

S = 3.03397 R-Sq = 60.45% R-Sq(adj) = 52.27%

Analysis of Variance for **Food**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	61.417	61.417	20.472	2.29	0.098
Age	2	345.722	345.722	172.861	19.36	0
Error	30	267.833	267.833	8.928		
Total	35	674.972				

S = 2.98794 R-Sq = 60.32% R-Sq(adj) = 53.71%

7) Analysis of Variance for **Health**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	11.639	11.639	3.88	0.54	0.657
Sex	1	39.014	39.014	39.014	5.46	0.027
Age	2	213.722	213.722	106.861	14.95	0
Error	29	207.264	207.264	7.147		
Total	35	471.639				

S = 2.67339 R-Sq = 56.05% R-Sq(adj) = 46.96%

Analysis of Variance for **Health**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	39.014	39.014	39.014	5.7	0.023
Age	2	213.722	213.722	106.861	15.62	0
Error	32	218.903	218.903	6.841		
Total	35	471.639				

S = 2.61548 R-Sq = 53.59% R-Sq(adj) = 49.24%

8) Analysis of Variance for **Misc**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	1.333	1.333	0.444	0.06	0.981
Sex	1	22.222	22.222	22.222	2.92	0.098
Age	2	108.222	108.222	54.111	7.12	0.003
Error	29	220.444	220.444	7.602		
Total	35	352.222				

S = 2.75709 R-Sq = 37.41% R-Sq(adj) = 24.46%

Analysis of Variance for **Misc**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P

Sex	1	22.222	22.222	22.222	3.21	0.083
Age	2	108.222	108.222	54.111	7.81	0.002
Error	32	221.778	221.778	6.931		
Total	35	352.222				

S = 2.63259 R-Sq = 37.03% R-Sq(adj) = 31.13%

9) Analysis of Variance for **Mobile/ISP**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	18.889	18.889	6.296	0.97	0.42
Sex	1	5.014	5.014	5.014	0.77	0.387
Age	2	484.722	484.722	242.361	37.33	0
Error	29	188.264	188.264	6.492		
Total	35	696.889				

S = 2.54791 R-Sq = 72.99% R-Sq(adj) = 67.40%

Analysis of Variance **for Mobile/ISP**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	5.01	5.01	5.01	0.77	0.385
Age	2	484.72	484.72	242.36	37.44	0
Error	32	207.15	207.15	6.47		

Total	35	696.89				

S = 2.54431 R-Sq = 70.27% R-Sq(adj) = 67.49%

10) Analysis of Variance for **Play/Game**, using Adjusted SS for Test

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	161.556	161.556	53.852	6.72	0.001
Sex	1	6.125	6.125	6.125	0.76	0.389
Age	2	366	366	183	22.84	0
Error	29	232.319	232.319	8.011		
Total	35	766				

S = 2.83037 R-Sq = 69.67% R-Sq(adj) = 63.40%

Analysis of Variance for **Play/Game**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	161.56	161.56	53.85	6.78	0.001
Age	2	366	366	183	23.02	0
Error	30	238.44	238.44	7.95		
Total	35	766				

S = 2.81925 R-Sq = 68.87% R-Sq(adj) = 63.68%

11) Analysis of Variance for **Shopping**, using Adjusted SS for Test

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	618.1	618.1	206	1.39	0.266
Sex	1	512	512	512	3.45	0.073
Age	2	11456.2	11456.2	5728.1	38.61	0
Error	29	4302.5	4302.5	148.4		
Total	35	16888.8				

S = 12.1804 R-Sq = 74.52% R-Sq(adj) = 69.25%

Analysis of Variance for **Shopping**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	512	512	512	3.33	0.077
Age	2	11456.2	11456.2	5728.1	37.25	0
Error	32	4920.6	4920.6	153.8		
Total	35	16888.8				

S = 12.4003 R-Sq = 70.86% R-Sq(adj) = 68.13%

12) Analysis of Variance for **Online Services**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	2	2	0.667	0.61	0.615

Sex	1	0.056	0.056	0.056	0.05	0.823
Age	2	5.056	5.056	2.528	2.31	0.118
Error	29	31.778	31.778	1.096		
Total	35	38.889				

S = 1.04680 R-Sq = 18.29% R-Sq(adj) = 1.38%

Analysis of Variance for **Online Services**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	2	2	0.667	0.63	0.602
Age	2	5.056	5.056	2.528	2.38	0.11
Error	30	31.833	31.833	1.061		
Total	35	38.889				

S = 1.03010 R-Sq = 18.14% R-Sq(adj) = 4.50%

13) Analysis of Variance for **Sports Wear**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	6.889	6.889	2.296	0.31	0.818
Sex	1	2	2	2	0.27	0.608
Age	2	224	224	112	15.1	0
Error	29	215.111	215.111	7.418		

Total	35	448				

S = 2.72353 R-Sq = 51.98% R-Sq(adj) = 42.05%

Analysis of Variance for **Sports Wear**, using Adjusted SS for Tests after removing City from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Sex	1	2	2	2	0.29	0.595
Age	2	224	224	112	16.14	0
Error	32	222	222	6.937		
Total	35	448				

S = 2.63391 R-Sq = 50.45% R-Sq(adj) = 45.80%

14) Analysis of Variance for **Travel**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	40.33	40.33	13.44	1.31	0.29
Sex	1	0.68	0.68	0.68	0.07	0.799
Age	2	112.89	112.89	56.44	5.51	0.009
Error	29	297.32	297.32	10.25		
Total	35	451.22				

S = 3.20194 R-Sq = 34.11% R-Sq(adj) = 20.48%

Analysis of Variance for **Travel**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	40.333	40.333	13.444	1.35	0.276
Age	2	112.889	112.889	56.444	5.68	0.008
Error	30	298	298	9.933		
Total	35	451.222				

S = 3.15172 R-Sq = 33.96% R-Sq(adj) = 22.95%

15) Analysis of Variance for **Entertainment**, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	105.56	105.56	35.19	0.65	0.591
Sex	1	6.12	6.13	6.13	0.11	0.74
Age	2	204.5	204.5	102.25	1.88	0.171
Error	29	1577.82	1577.82	54.41		
Total	35	1894				

S = 7.37615 R-Sq = 16.69% R-Sq(adj) = 0.00%

Analysis of Variance for **Entertainment**, using Adjusted SS for Tests after removing Sex from the significant variables.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
City	3	105.56	105.56	35.19	0.67	0.579

Age	2	204.5	204.5	102.25	1.94	0.162
Error	30	1583.94	1583.94	52.8		
Total	35	1894				

S = 7.26623 R-Sq = 16.37% R-Sq(adj) = 2.43%

References

[1] http://www.bloomberg.com/news/2012-03-30/facebook-valued-at-102-8-billion-in-final-sharespost-auction.html

[2] http://en.wikipedia.org/wiki/Facebook

[3] http://www.bloomberg.com/news/2011-06-20/facebook-surpasses-yahoo-as-top-u-s-display-ad-seller-in-study.html

[4] http://articles.economictimes.indiatimes.com/2011-11-09/news/30373933_1_facebook-users-social-networking-internet-users

[5] http://www.socialbakers.com/facebook-statistics/india

[6] http://newsroom.fb.com/content/default.aspx?NewsAreaId=22

[7] http://money.cnn.com/2010/03/16/technology/facebook_most_visited/

[8] http://www.bloomberg.com/news/2011-09-20/facebook-revenue-will-reach-4-27-billion-emarketer-says-1-.html

[9] http://www.internetworldstats.com/facebook.htm

[10] http://www.insidefacebook.com/2008/10/30/facebook-advertising-resources-the-6-types-of-ads-on-the-new-home-page/

[11] http://mashable.com/2011/06/28/facebook-advertising-infographic/

[12] Combining Behavioral and Social Network Data for Online Advertising By Abraham Bagherjeiran, Rajesh Parekh Yahoo! Strategic Data Solutions: Data Mining & Research {abagher, rparekh}@yahoo-inc.com, 2008

[13] Large-Scale Behavioral Targeting with a Social Twist By Kun Liu Yahoo! Labs 4301 Great America Pkwy Santa Clara, CA 95054 kun@yahoo-inc.com, Lei Tang Yahoo! Labs 4301 Great America Pkwy Santa Clara, CA 95054 ltang@yahoo-inc.com, 2011

[14] Learning Relevance from Heterogeneous Social Network and Its Application in Online Targeting by Chi Wang UIUC chiwang1@illinois.edu, Rajat Raina Facebook rajatr@fb.com, David Fong Stanford University clfong@stanford.edu, Ding Zhou Facebook dzhou@fb.com, Jiawei Han UIUC hanj@illinois.edu, Greg Badros Facebook badros@fb.com, 2011

[15] Social Networks, Personalized Advertising, and Perceptions of Privacy Control By Catherine Tucker, 2011

[16] http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm