Determining User Influence on a Social Network

Chaitanya Khurana

University of Eastern Finland

School of Computing

Master’s Thesis

25.4.2015

ABSTRACT

In recent years, social network services have become very popular across the globe. These include Facebook, Twitter, Flickr, Pinterest, Tumblr, LinkedIn and Google+. Facebook, with more than 1 billion users[1] is most popular. It allows integration with third party applications, including games.

Mopsi[2] is a location based application which provides services such as finding the location of friends, photo sharing, tracking and chatting. It is also integrated with Facebook and enables users to upload their content (photos and routes) on Facebook. Since a large number of photos are uploaded on Facebook via Mopsi, such photos can be used as a means of advertising on Facebook.

Influence of users in a social network has been studied by various researchers. Influence affects the diffusion of information in the network, which is of interest to the social media strategy of the business world. In our research, we rely on users' activity data such as likes and comments of users' photos and develop a method for identifying influential users in Facebook. We design predictive model for estimating user influence, conduct temporal analysis of user response, and build a predictive popularity model to determine best time and day for uploading photo. We validate our results by online survey.

The results show that average likes received by a female user is greater than a male user. A user who has relatively more male friends are more influential than others, on average. A female user with male dominant friend list seems to be best combination for attracting likes and comments. A user with more friends is relatively more influential than those with less friends, on average. About 80% of total comments on a popular photo are received within 24 hours of uploading it. Early comment makers are more influenced than late comment makers. Most of the popular photos are uploaded on Sunday and Monday, and during morning and night time. Popularity of a photo correlates with day and time of uploading the photo, user's gender, friends and gender of user's friends.

[1] http://www.usatoday.com/story/tech/2012/10/04/facebook-tops-1-billion-users/1612613/
[2] http://cs.uef.fi/mopsi

Using these results, we outline a new business model for Mopsi by enabling companies target their advertisements at users with favourable characteristics.

Keywords: Social network analysis, influence in social network, influential users, advertising on Facebook, new advertising model, product marketing.

ACKNOWLEDGEMENTS

I am grateful to the University of Eastern Finland for giving me an opportunity to study excellent computer courses under the IMPIT programme. I would like to express thanks to all the faculty members who helped me gaining both theoretical knowledge and practical experience in different areas of computer science.

I wish to express my deepest gratitude to my supervisor, Professor Pasi Fränti, for his guidance on research work and openness for encouraging new ideas. His continued support helped me in improving problem solving, time management and presentation skills. In addition, I would like to thank all Mopsi members for making me feel as an important part of a team.

I want to thank my family members and friends for their moral support and encouragement. They always stood by me in all the circumstances.

In the end, I would like to dedicate this thesis to my late grandparents who always encouraged me to pursue higher studies. They are still alive in my heart and memories.

Contents

1 Introduction

A *social network* [1] is a structure made up of individuals and connections between them. In the last decade, a number of social networking platforms have emerged on the Internet. Examples include Facebook, Twitter, MySpace, Flickr, LinkedIn and Google+. These platforms enable generation of a large amount of data due to user activities. Such data can be analyzed to understand user behaviour and its impact on e-commerce.

Facebook has a large number of users who connect with others by adding them as friends and share their information such as status updates, photos, videos, places visited, life events, interests and emotions. The users communicate by likes and comments on the content of other users. According to a social media survey[3], a user spends almost seven hours each month on Facebook, on average. These factors enabled Facebook to earn estimated revenue of 1.86 billion dollars[4] (2010) from advertising alone. Its advertising model enables business enterprises to target their advertisements based on user's characteristics such as location, gender, age, likes and interests, relationship status, workplace and education. The main motivation of integrating applications with Facebook is to get easy and quick access to its large user base. Location based applications which are integrated with Facebook include *Foursquare*, *Mopsi*, *OpenTable* and *Fandango*. However, none of these applications use sharing of photos for advertisements.

In this thesis, we study how the photos uploaded on Facebook can be used for advertisements. For this, we use Mopsi and implement tools for extracting user information from Facebook.

Mopsi users can login using the Facebook account details or link their Mopsi and Facebook accounts explicitly. Such users (User of both) are shown as in Figure 1. These users can share their photos and routes via Mopsi to Facebook. The other Mopsi users might still be Facebook users, but since they have not logged in using Facebook credentials or linked their Mopsi-Facebook accounts, we do not have access to their Facebook network. Example of a photo uploaded by Mopsi user to Facebook is shown in Figure 2.

By studying the network, we aim at learning about users via their photo sharing. In specific, we

[3] http://mashable.com/2012/11/28/social-media-time/
[4] http://mashable.com/2011/01/17/facebooks-ad-revenue-hit-1-86b-for-2010/

want to perform:

- Identification of influential users
- Temporal analysis of user response, and
- Popularity of photos

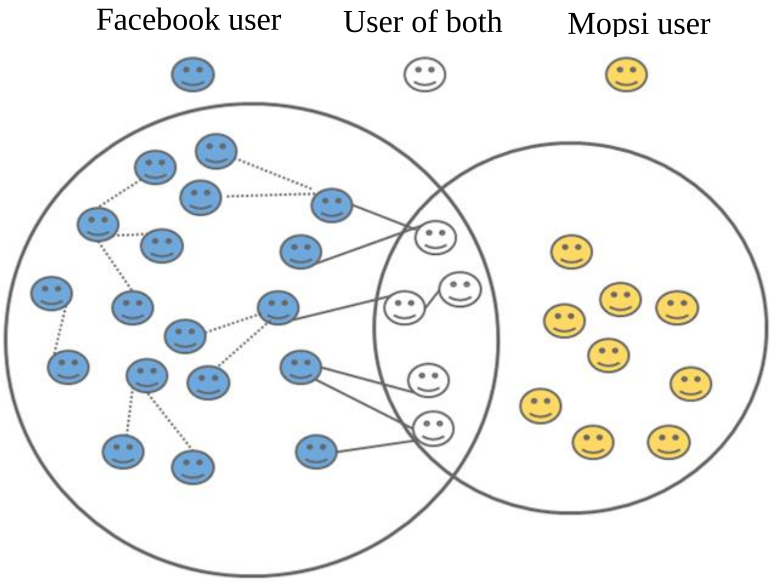
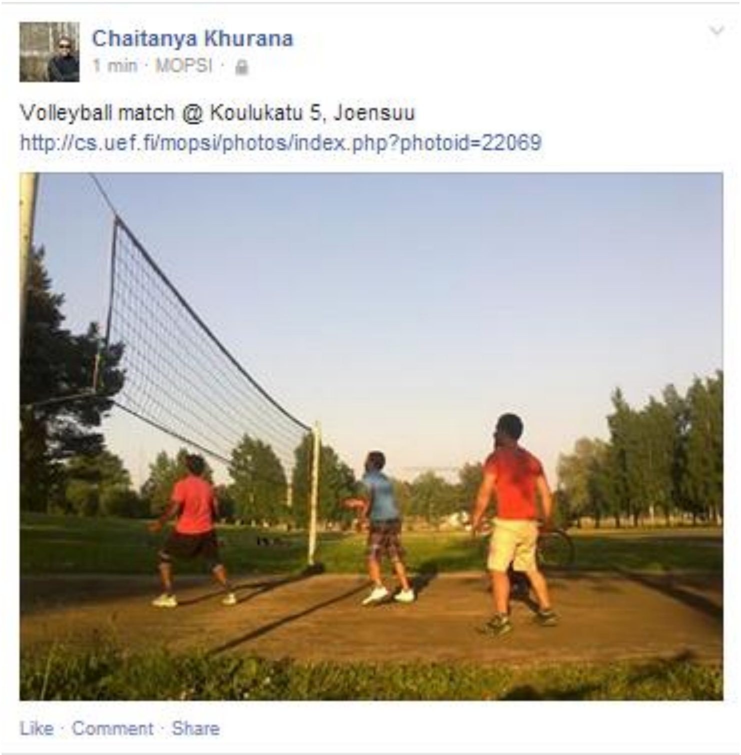


Figure 1: Network connections (solid lines) between Mopsi and Facebook users



Chaitanya Khurana
1 min · MOPSI ·

Volleyball match @ Koulukatu 5, Joensuu
http://cs.uef.fi/mopsi/photos/index.php?photoid=22069

Figure 2: Mopsi photo uploaded on Facebook by Mopsi user.

`

Identification of influential users

Influential people are those who excel in persuading others [2]. Using this definition, we determine influential users on Facebook, by considering the likes and comments made by their friends on their photos.

Temporal analysis of user response

In this thesis we will analyze the number of comments made on photos with respect to time. We select a set of popular and random photos for this analysis. This analysis identifies the percentage of comments received in first 24 hours of uploading a photo. Besides, it will identify the relationship between the influence[5] on user and the duration before making a comment.

Popularity of photos

Popularity of the photos uploaded on Facebook by Mopsi users will be computed using likes and comments on a photo. Its equation is discussed with an example in Chapter 5. In this, we try to find relationship between the photo's popularity value and the attributes including day and time of uploading, number of friends and gender of a user, and gender of user's friends. Based on these, we make a predictive model for photo popularity. We generate a popularity matrix which predicts the best day and time to upload a photo to gain maximum popularity.

A web-based Social Network Analysis tool (SNA) is developed to conduct this research. It will enable introduction of a new advertising model that would help Mopsi to increase its popularity and generate revenue.

[5] M. Trusov, A.V. Bodapati, R. E. Bucklin, "Determining Influential Users in Internet Social Networks", *Journal of Marketing Research*, pp. 643-658, 2010

2 Mopsi

Mopsi is a location-based application which provides services such as finding the location of friends, photo sharing, users tracking, chatting, search, recommendation and action notifications. It has been developed by the Speech and Image Processing Unit, School of Computing, University of Eastern Finland and can be found on web at http://cs.uef.fi/mopsi. The mobile version of Mopsi can be downloaded for most platforms from http://cs.uef.fi/mopsi/mobile.php.

To achieve the goals of this thesis, we used Mopsi application hosted on Facebook to enable two-way communication between FB and Mopsi. Facebook provides an application id and a secret key to the Mopsi application that allows the application to access data from Facebook and share the content of Mopsi users on Facebook. A description of how the access to Facebook data is enabled by this integration is shown in Figure 3.

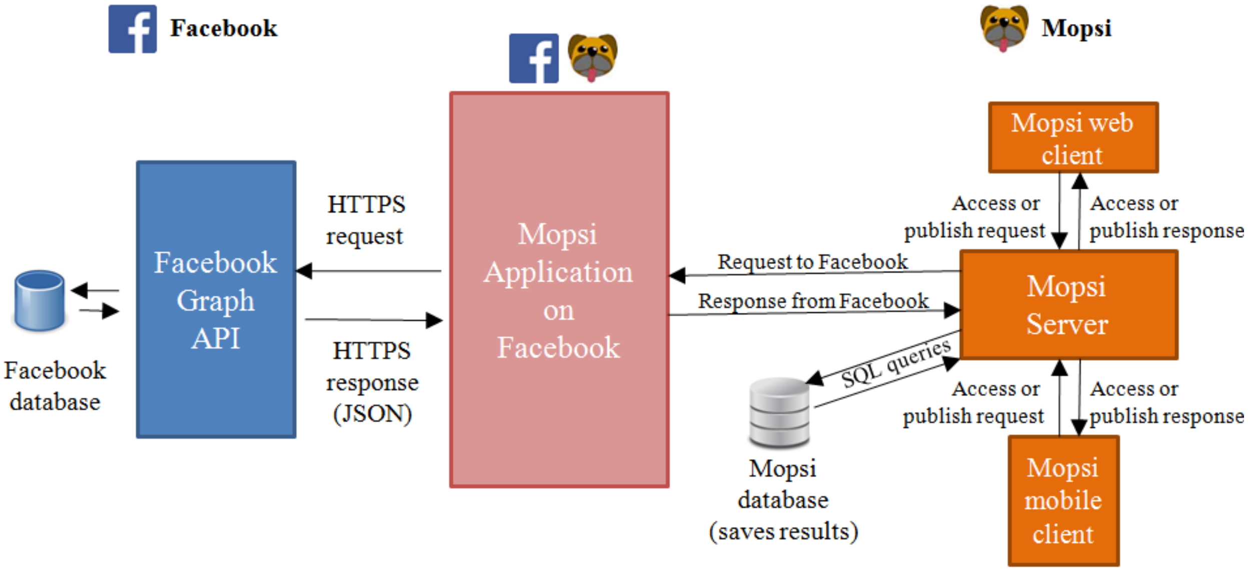


Figure 3: Mopsi application on Facebook integrates Mopsi and Facebook.

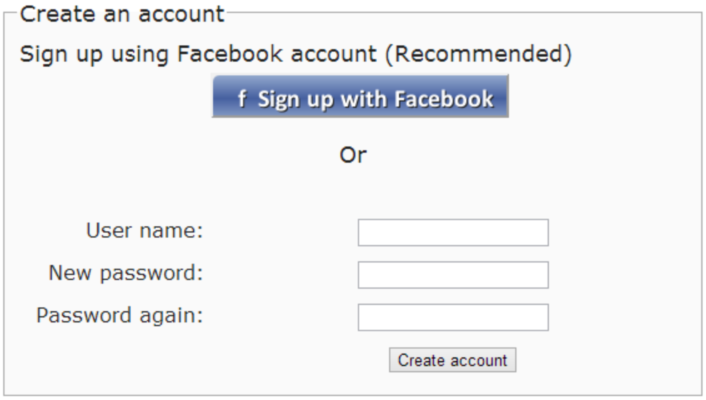
2.1 System overview

To access user's data available on Facebook, Mopsi application sends HTTPS request to Facebook Graph API. This includes Facebook user id, required fields and access token of user. Graph API responds in JSON format, which is decoded and stored in Mopsi database using SQL queries. To share user's content (photos and routes) on Facebook, application sends a request to

Facebook Graph API. The detailed features of application are described in the following sub sections.

2.2 Registration and authentication

To register on Mopsi, users may sign up with Facebook or create own Mopsi username, see Figure 4. If user chooses Facebook sign up, Facebook login window is displayed, see Figure 5, and then a login dialog appears, see Figure 6. It notifies the user that Mopsi will get access to user's information (public profile, friend list, email address and photos) and post publicly to Facebook on behalf of the user, if he presses the ok button. Access token is generated and saved in Mopsi database, when the user accepts Mopsi. Figure 7 shows the workflow for generation of access token. It describes the workflow at two levels: System level and UI level. At system level, Mopsi server sends request (including app id, app secret key and set of permissions) to Facebook Graph API for generating login dialog. At UI level, user accepts login dialog and generated access token is returned to Mopsi.



Create an account

Sign up using Facebook account (Recommended)

f Sign up with Facebook

Or

User name:

New password:

Password again:

Create account

Figure 4: Interface for creation of a new Mopsi account using existing Facebook account.

Example of user's Facebook data stored in Mopsi is shown below. Facebook user id and email are used for uniquely identifying a user. User's access token allows Mopsi to retrieve or publish data to Facebook on behalf of user.

Facebook user id:	1546381264
Email id:	chaitanyakhurana@ymail.com
Access token:	CAABtc4NJzcgBAPe73TU0wPWjp5VXg9747NO2p…..
Full name:	Chaitanya Khurana

There are two cases for authentication when a user logs in using Mopsi, see Figure 8:

1. User maintains a username and password on Mopsi: These are matched with those in the Mopsi database to authenticate the user.

2. Facebook login details: Mopsi checks the Facebook user id or email id in its own database for authentication, see Figure 5. The complete workflow for authentication of the user in Mopsi is shown in Figure 9.

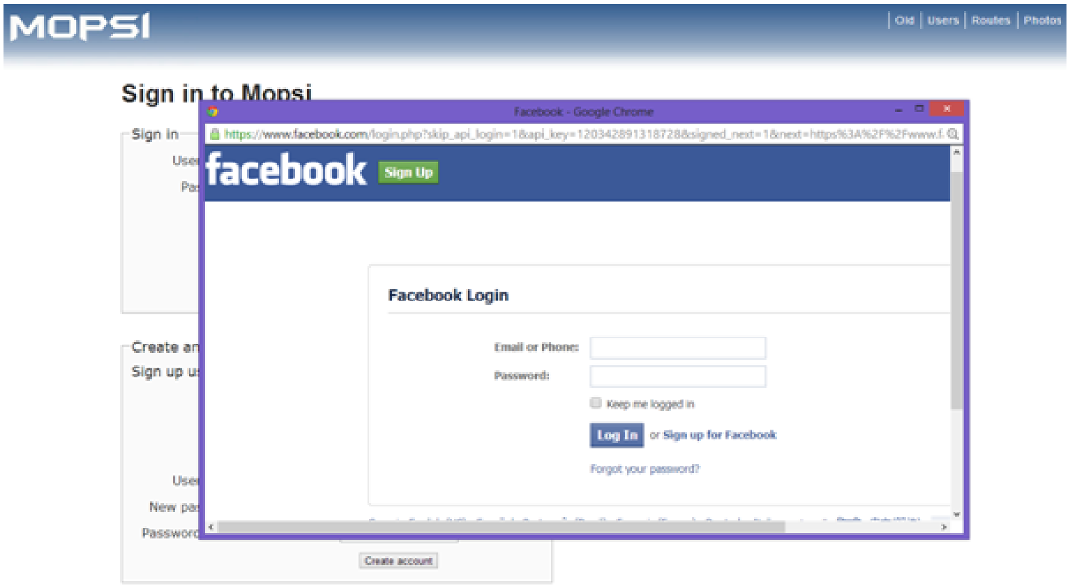
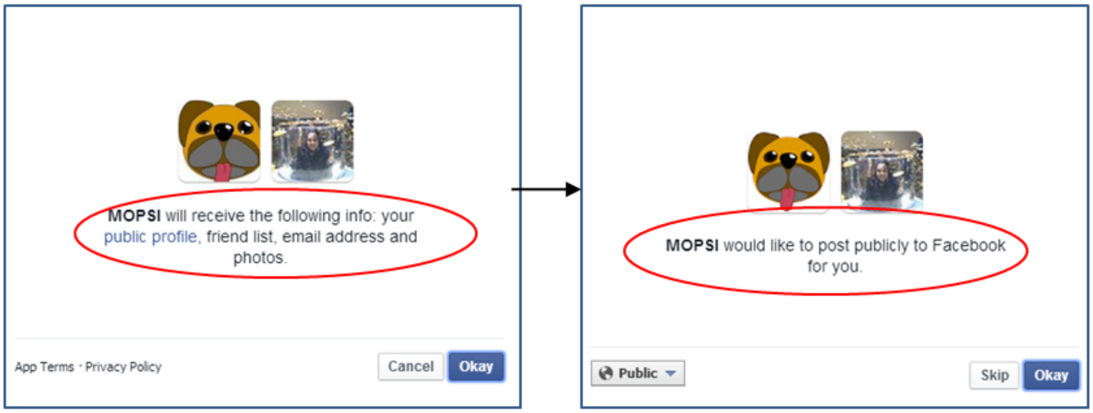


Figure 5: Facebook login details can be used in Mopsi


MOPSI will receive the following info: your public profile, friend list, email address and photos.

MOPSI would like to post publicly to Facebook for you.

Figure 6: Mopsi application login dialog asking for user's permissions on Facebook

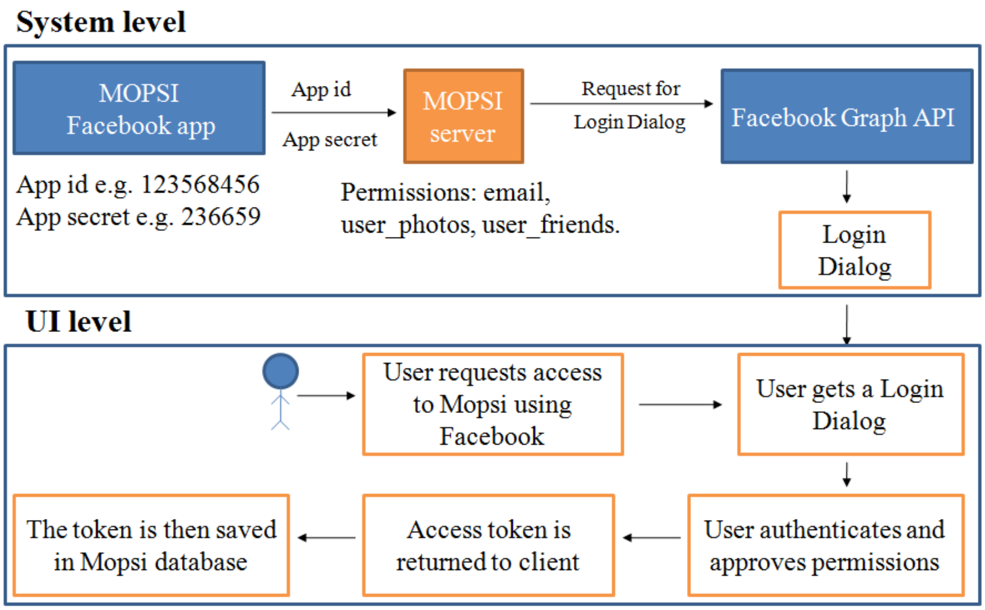

System level
MOPSI Facebook app
App id
App secret
MOPSI server
Request for Login Dialog
Facebook Graph API
App id e.g. 123568456
App secret e.g. 236659
Permissions: email, user_photos, user_friends.
Login Dialog
UI level
User requests access to Mopsi using Facebook
User gets a Login Dialog
The token is then saved in Mopsi database
Access token is returned to client
User authenticates and approves permissions

Figure 7: Workflow for generation of access token

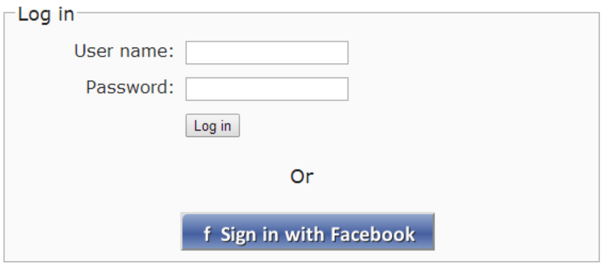


Figure 8: Login interface on Mopsi

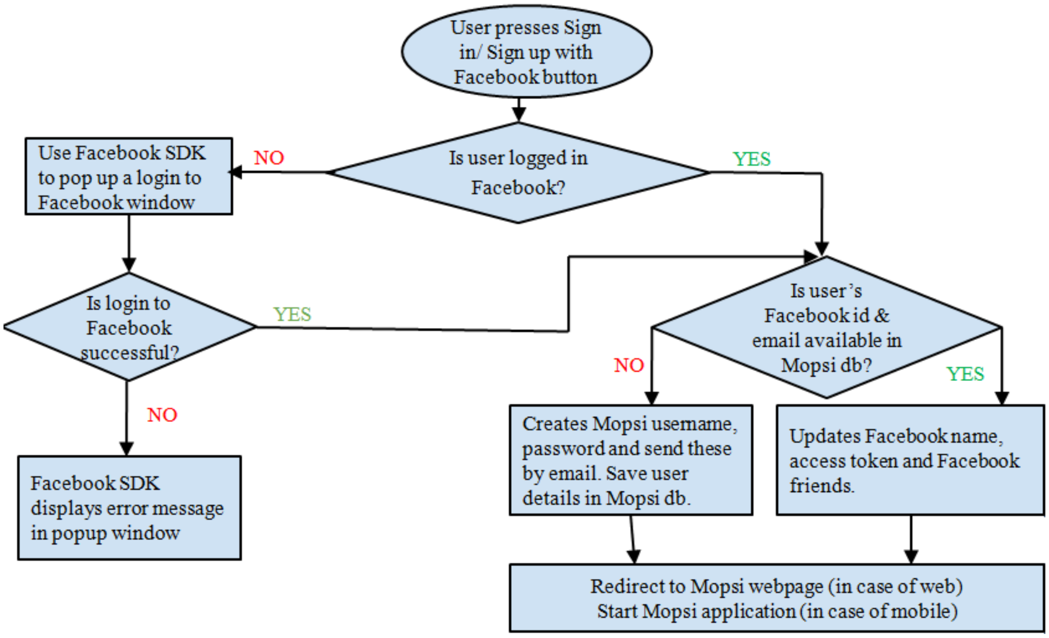
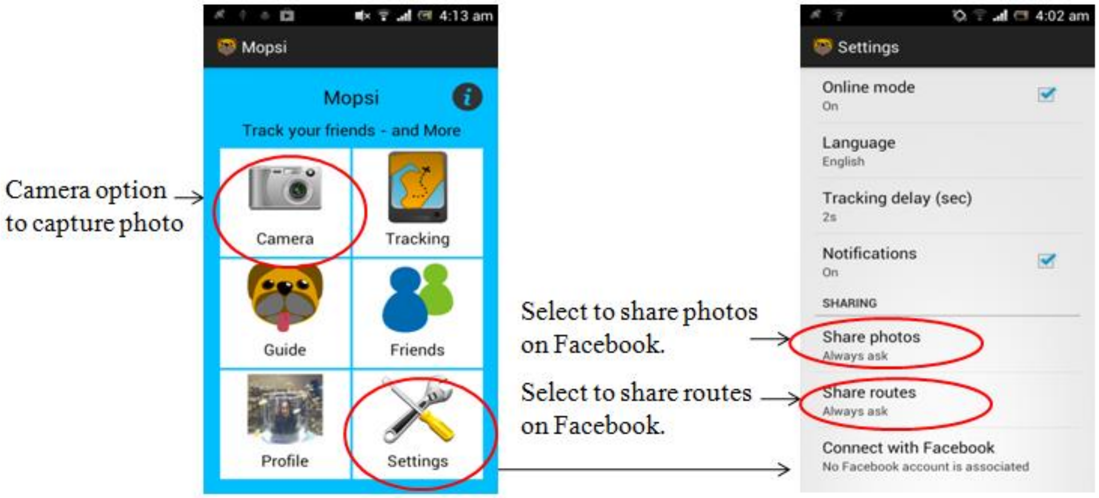

User presses Sign in/ Sign up with Facebook button
Is user logged in Facebook?
NO
YES
Use Facebook SDK to pop up a login to Facebook window
Is login to Facebook successful?
YES
NO
Facebook SDK displays error message in popup window
Is user's Facebook id & email available in Mopsi db?
NO
YES
Creates Mopsi username, password and send these by email. Save user details in Mopsi db.
Updates Facebook name, access token and Facebook friends.
Redirect to Mopsi webpage (in case of web)
Start Mopsi application (in case of mobile)

Figure 9: Workflow for user authentication in Mopsi (with Facebook details)

2.3 Photo sharing

Photo sharing on Mopsi involves use of Mopsi mobile application which has versions for Android, iOS, Windows and Symbian. In this thesis, we use screenshots of Mopsi Android application to demonstrate the process.

To capture and upload photo on Mopsi, a user will select the *Camera* option on the welcome screen, see Figure 10. This photo will be visible on the Map in Mopsi web. Uploaded photo will automatically get shared on Facebook if the user has selected the option *Share photos* in the settings, see Figure 10. If not, then the photo can be shared also from the Mopsi web, by pressing the Facebook button available below the photo, see Figure 11. Technical details of this photo sharing are given below.



Camera option to capture photo

Select to share photos on Facebook.

Select to share routes on Facebook.

Settings

Online mode ☑
On

Language
English

Tracking delay (sec)
2s

Notifications ☑
On

SHARING

Share photos
Always ask

Share routes
Always ask

Connect with Facebook
No Facebook account is associated

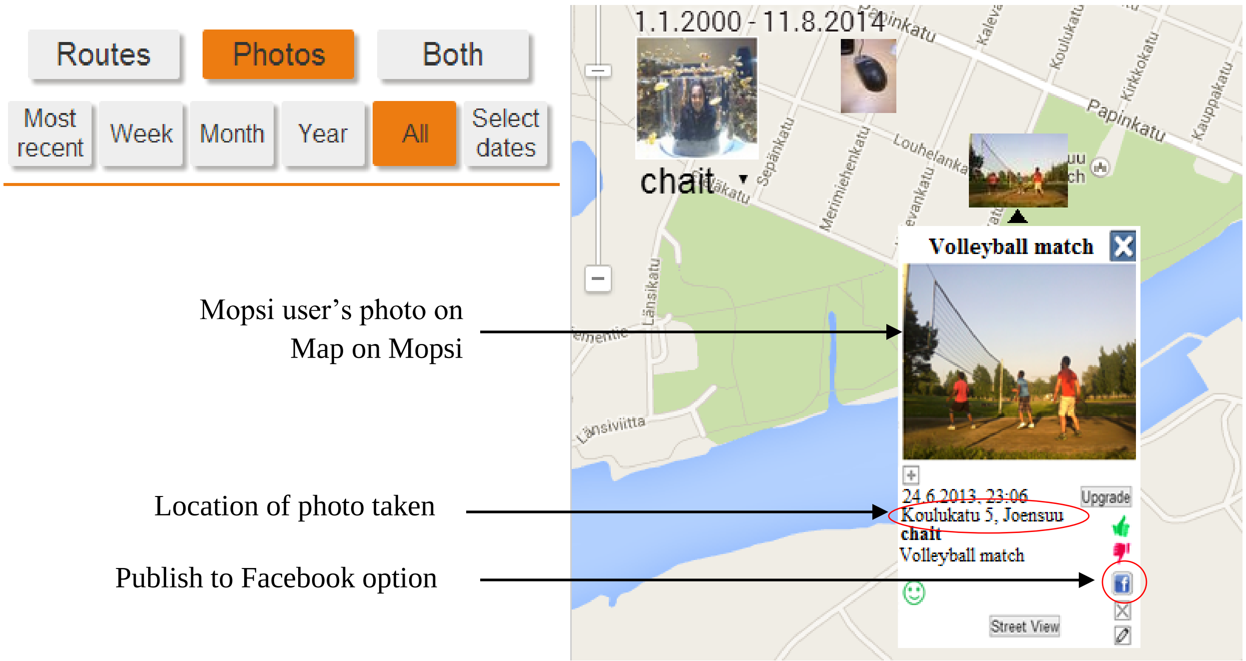
Figure 10: Welcome (left) and Settings (right) screens of Mopsi mobile

The shared photos get stored in Mopsi photo albums on Facebook. Before sharing a photo, Mopsi checks the existence of an album on Facebook using the recently stored album id. This results in two cases:

1. Album exists: An album can store 200 photos. Mopsi sends a request to Facebook to determine the number of photos in the current photo album. If it is 200, another request is sent to create a new photo album and add the new photo.

2. Album does not exist: Mopsi sends a request to Facebook to create a new photo album and adds the new photo to it.

An example of a Mopsi photo album is shown in Figure 12. The complete workflow for publishing a photo to Facebook is shown in Figure 13.



Volleyball match

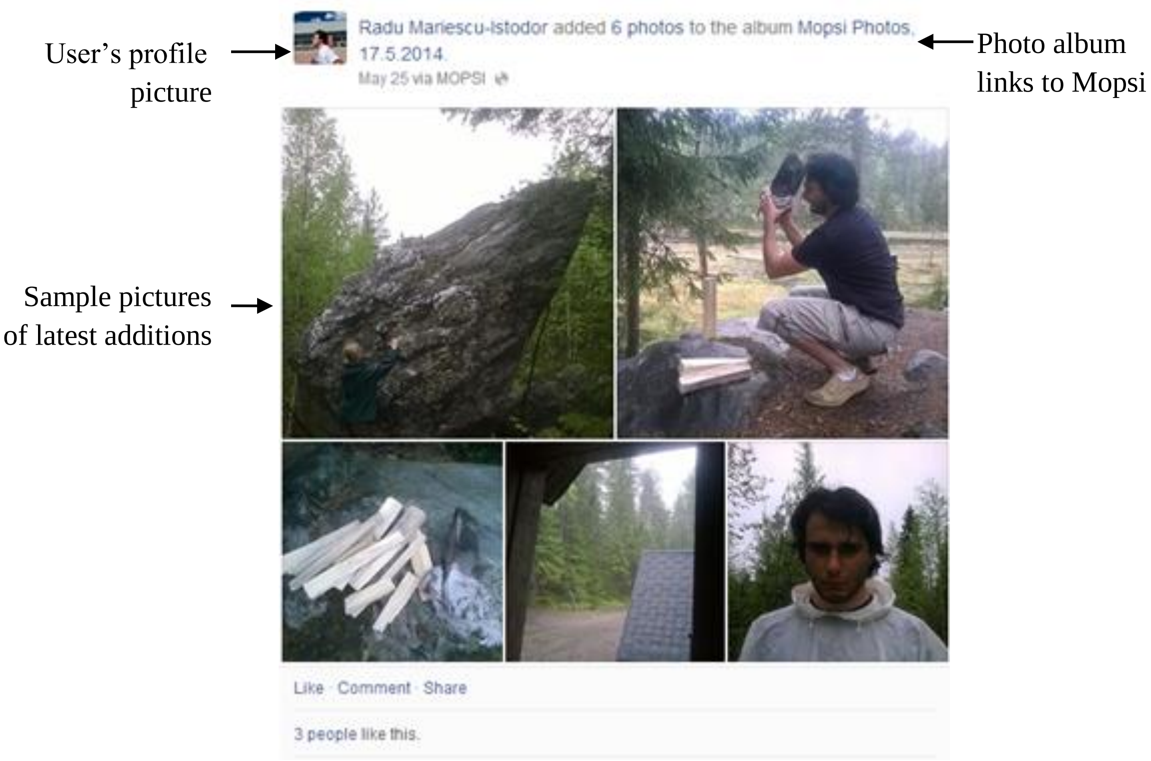
24.6.2013 23:06
Koulukatu 5, Joensuu
chait
Volleyball match

Mopsi user's photo on Map on Mopsi

Location of photo taken

Publish to Facebook option

Figure 11: Mopsi user's photo in Mopsi


Radu Mariescu-Istodor added 6 photos to the album Mopsi Photos, 17.5.2014.
May 25 via MOPSI

Like · Comment · Share

3 people like this.

User's profile picture

Photo album links to Mopsi

Sample pictures of latest additions

Figure 12: Mopsi user's photo album on Facebook

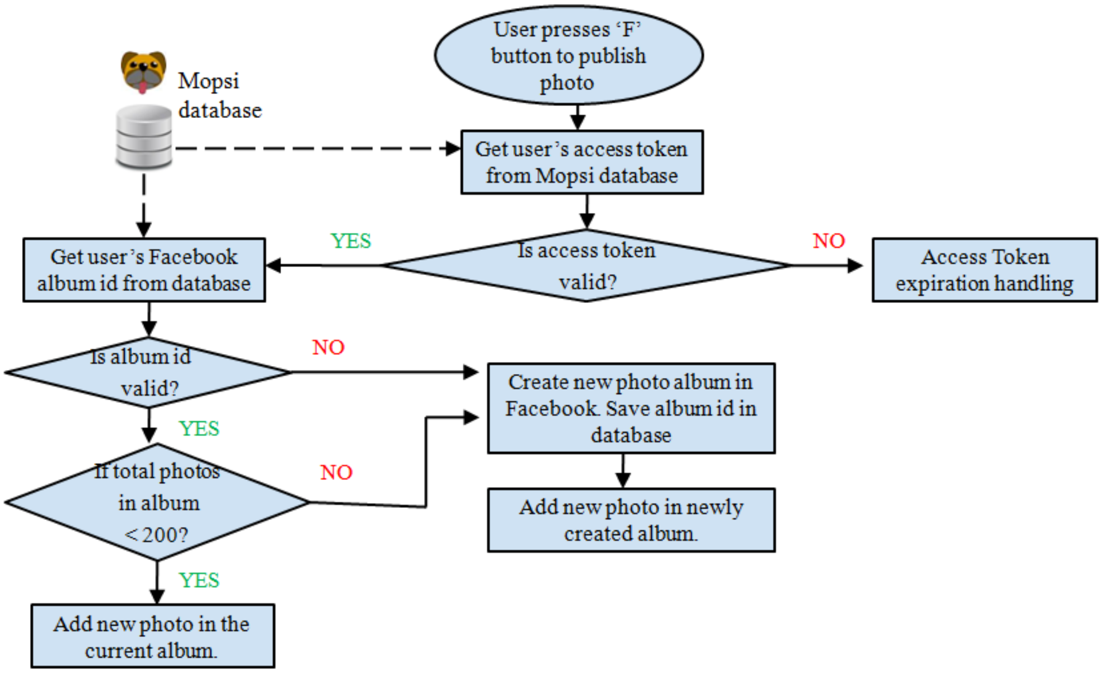

User presses 'F' button to publish photo
Mopsi database
Get user's access token from Mopsi database
YES
Is access token valid?
NO
Get user's Facebook album id from database
Access Token expiration handling
Is album id valid?
NO
Create new photo album in Facebook. Save album id in database
YES
If total photos in album < 200?
NO
Add new photo in newly created album.
YES
Add new photo in the current album.

Figure 13: Workflow for publishing a photo on Facebook

2.4 Route creation and sharing

Tracking a route means recording the route travelled by a user and saving it in Mopsi database. It can be done by Mopsi mobile application with any platform. To start tracking a route, user selects the *Tracking* option on the welcome screen, see Figure 8. When the tracking is complete, the route is published to Mopsi. It can be seen on the map in Mopsi web, see Figure 16.

Once complete, the route will automatically get shared on Facebook if the user has already selected the option *Share routes* on the settings, see Figure 10. If not, then it can also be shared from the Mopsi web, by pressing the Facebook button available, see Figure 16. Technical details of this route creation and sharing are given below.

While tracking, the Mopsi mobile application stores the latitude and longitude values of user's location at regular intervals in the mobile. The array of location points along with the user id, timestamp and flag are sent to the server in batches asynchronously if an Internet connection is available, see Figure 14. The location points remain saved in mobile if Internet connection is not available.

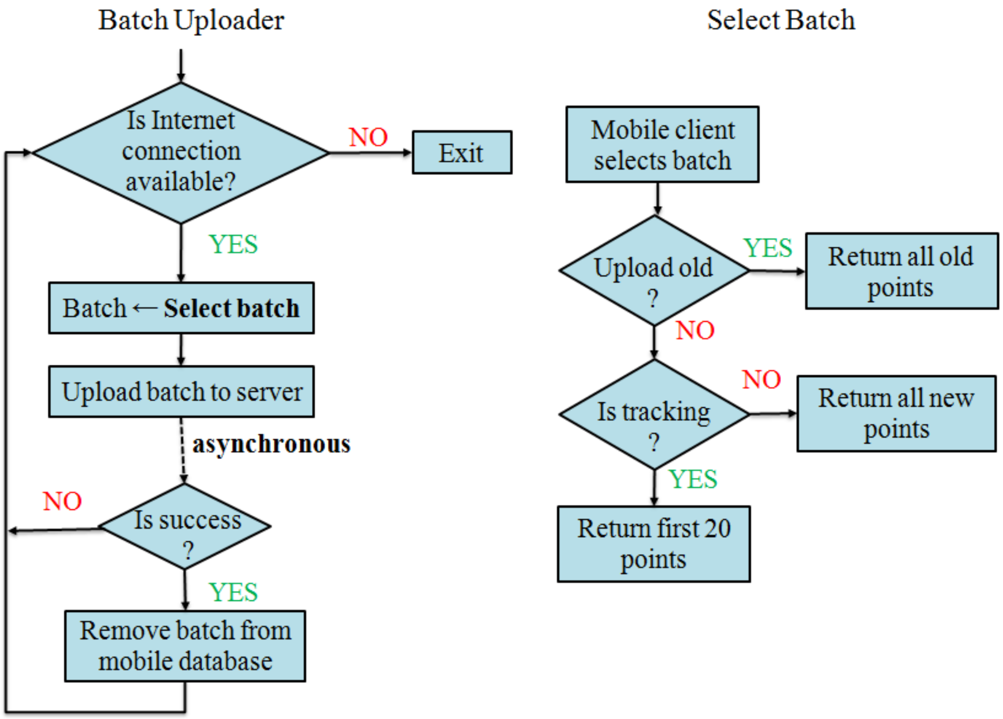


Figure 14: Workflow for the process of sending data from mobile to Mopsi server in batches

A stop flag is sent with the last batch of points. This invokes the server side process of publishing the route on Facebook. In this process, the route points are retrieved from the Mopsi database to calculate route distance and bounding box. The route distance is calculated as a sum of distances between every consecutive pair of location points. However, bounding box requires only minimum and maximum values of latitude and longitude. The process continues only if the route distance is more than 100 metres. Then, the street addresses of route's start and end are retrieved from the Mopsi database. The mode of transport is also determined using the route segmentation algorithm [3] implemented in Mopsi. After this, to create a route image, a request is sent to OSM API with the following data: bounding box values, location points as a route string, required width and height of the image. The API automatically generates a route image which is stored on the Mopsi server. This process is also shown in a flowchart in Figure 15.

The route image and the statistics, including duration, distance, speed, mode of transport, start and end locations, are shared on the Mopsi user's Facebook timeline, see Figure 17. The duration is calculated from start and end timestamps, and the speed is calculated by dividing distance by time.

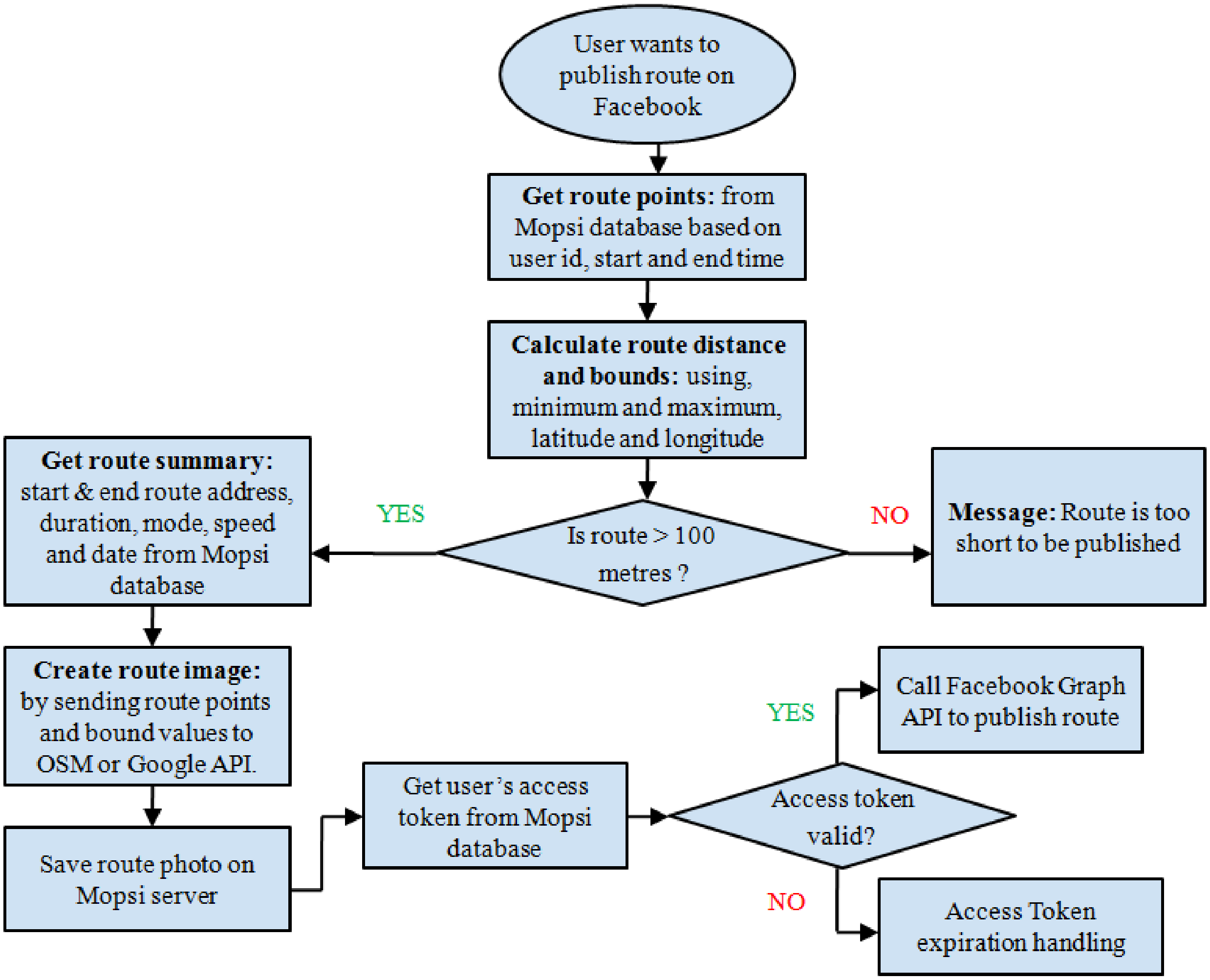
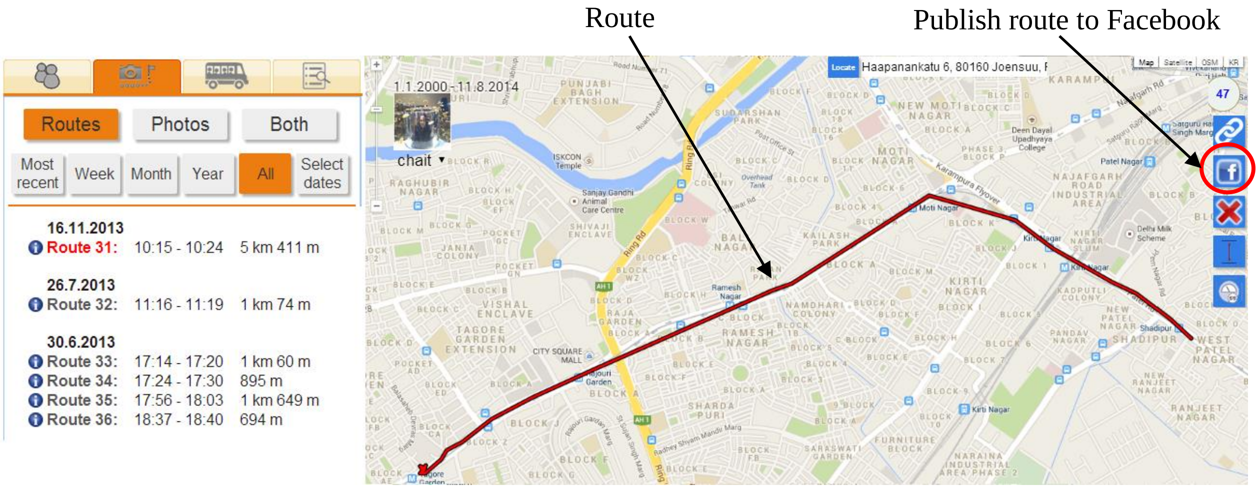

User wants to publish route on Facebook
Get route points: from Mopsi database based on user id, start and end time
Calculate route distance and bounds: using, minimum and maximum, latitude and longitude
Is route > 100 metres ?
YES
NO
Get route summary: start & end route address, duration, mode, speed and date from Mopsi database
Message: Route is too short to be published
Create route image: by sending route points and bound values to OSM or Google API.
Save route photo on Mopsi server
Get user's access token from Mopsi database
Access token valid?
YES
Call Facebook Graph API to publish route
NO
Access Token expiration handling

Figure 15: Flowchart for publishing the route on Facebook

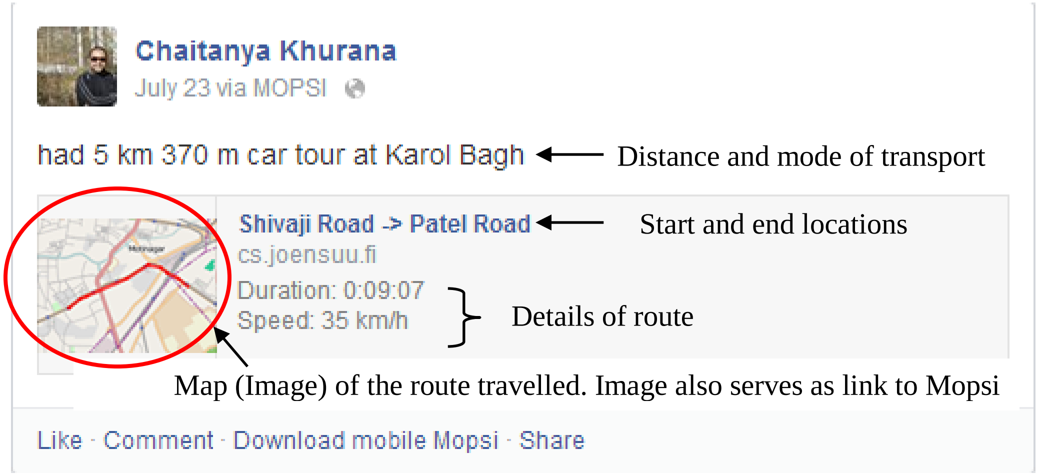


16.11.2013
Route 31: 10:15 - 10:24 5 km 411 m

26.7.2013
Route 32: 11:16 - 11:19 1 km 74 m

30.6.2013
Route 33: 17:14 - 17:20 1 km 60 m
Route 34: 17:24 - 17:30 895 m
Route 35: 17:56 - 18:03 1 km 649 m
Route 36: 18:37 - 18:40 694 m

Figure 16: A route on Mopsi



Chaitanya Khurana

July 23 via MOPSI

had 5 km 370 m car tour at Karol Bagh

Shivaji Road -> Patel Road

cs.joensuu.fi

Duration: 0:09:07

Speed: 35 km/h

Like · Comment · Download mobile Mopsi · Share

Distance and mode of transport

Start and end locations

Details of route

Map (Image) of the route travelled. Image also serves as link to Mopsi

Figure 17: Mopsi user’s route shared on Facebook

3 Identification of influential users

The concept of influence has drawn the interest of researchers in various fields such as communication, sociology, political science, and marketing [2,4]. Since 1960s, more than 70 marketing studies have focused on influence [5]. According to one of the theories of communication, proposed in 1962, influential people can easily persuade others to change their behavior [2]. In [4], it is discussed how influential people could reach out to a wider audience at a lower advertising cost. More recently, in 2004, the research on viral marketing [6] showed that consumers influence other consumers by analyzing pass along emails.

Researchers differ in their opinion on effect of influence on diffusion[6] of information among people. Some claim that diffusion is catalyzed by influence [7,8,9]. Example in a voting study [7], the researchers concluded that personal influence appears to be more effective than mass media in influencing voters. On the other hand, others suggest that diffusion is driven by susceptibility (getting influenced) [10-14]. An example of such claim is the threshold models of collective behavior [11]. This model emphasizes that an individual engages in an activity depending upon the proportion of the other individuals already engaged in the same activity. But, there is little empirical evidence to validate that either influence or susceptibility is responsible for diffusion of ideas or products.

In the research on online social networks, influence has been defined in different ways. In [15], *Social influence* is defined as the phenomenon in which the action of a user can induce his/her friends to behave in a similar way. In [16], *peer influence* is defined as how peer's behavior can change one's expected utility and thus change the likelihood that or extent to which one will engage in the behavior. In [17], the authors describe that user's influence on others can be determined using directed links in a social network. They consider user influence as an important concept for sociology and viral marketing. The authors suggest that studying user influence can help social scientists improve their understanding of the social behaviors such as how people vote [18] and how people adopt fashion [19]. Besides, it can help advertisers to plan more effective ad campaigns.

[6] diffusion is spreading of something more widely. Ref: http://www.oxforddictionaries.com/definition/english/diffusion

Different methods and approaches have been used to determine user influence based on the availability of data. In [5], influence was determined based on a user's activity level, as captured by site log-ins over time. In [17], the researchers analyzed three types of influence on Twitter based on indegree, retweets and mentions. The indegree, retweets and mentions refer to total followers, number of times a particular content is shared and number of times a user is mentioned by others in their post, respectively. In [20], the estimation of influence was based on adoption of an application by a user's friends after its adoption by the user. For this purpose a notification was sent to the randomly selected friends of the user. In [21], a user's influence on his friend was calculated using the number of retweets with respect to all tweets of the user. In [22], influence was determined using degree and centrality based-heuristics. In [23], a user's influence rank was estimated using several metrics including number of followers, ratio of affection, magnitude of influence and the influence rank itself. In [24], a user's influence on his friend is calculated using the proportion of investment done by the friend on him, relative to others in the friend's network.

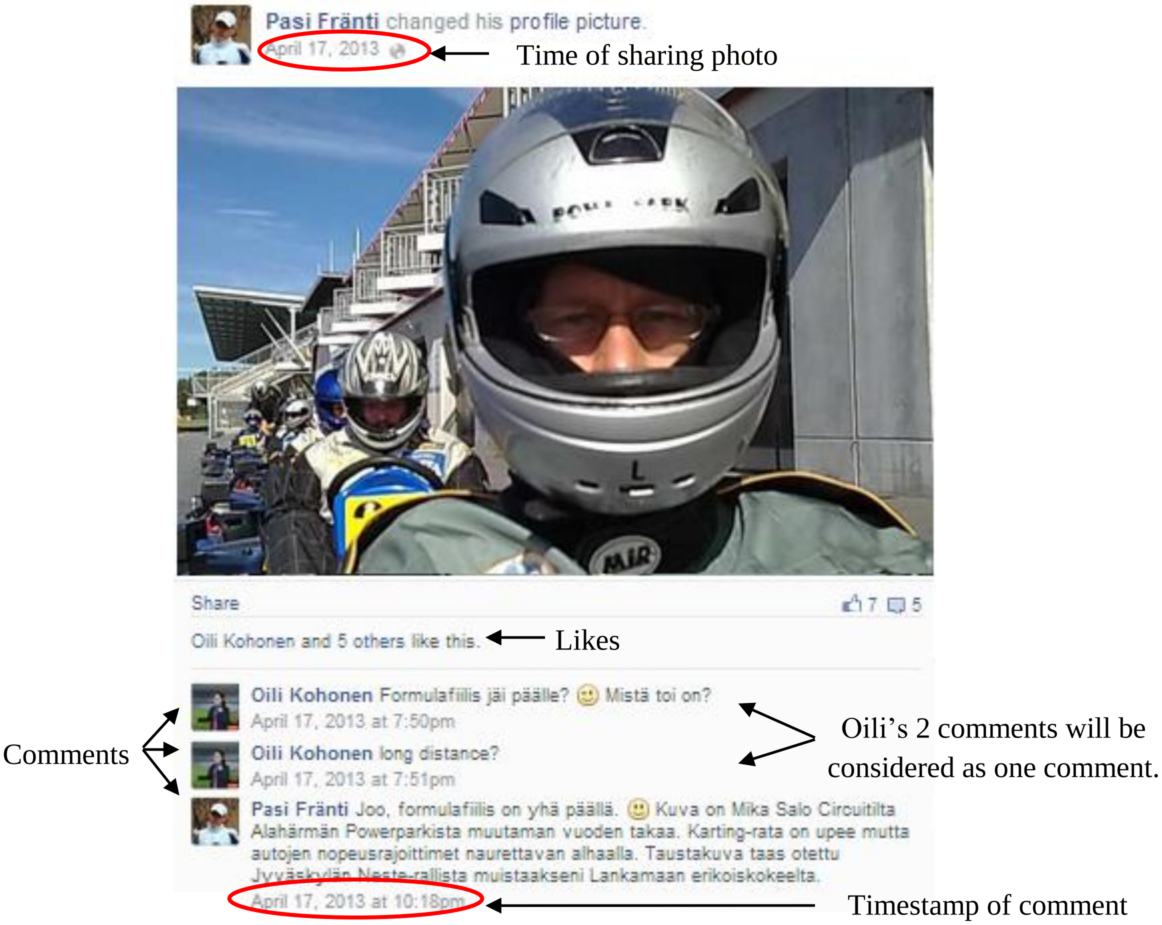
3.1 Influence

We define influence of a user based on the likes and comments made on the user's photos (see Figure 18) by his friends as shown in the following equation:

$$Influence(u_1, u_2) = \frac{l + 4 \cdot c}{p} \tag{1}$$

u_1, u_2 = influence of user u_1 to u_2
l = number of photos of u_1 liked by u_2
c = number of photos of u_1 commented by u_2
p = total photos of u_1

Figure 18 shows Pasi's photo with time of uploading photo, number of likes, comments with their timestamps. The photo has a total of 7 likes and 5 comments. Among these 5 comments, more than one comment by any user is considered as one comment. For example: Oili's 2 comments on Pasi's photo will be considered as 1 comment, see Figure 18. Also, Pasi's comments on his own photo are not included while calculating his influence.



Pasi Fränti changed his profile picture.
April 17, 2013

Share

7 5

Oili Kohonen and 5 others like this.

Oili Kohonen Formulafiilis jäi päälle? 🙂 Mistä toi on?
April 17, 2013 at 7:50pm

Oili Kohonen long distance?
April 17, 2013 at 7:51pm

Pasi Fränti Joo, formulafiilis on yhä päällä. 🙂 Kuva on Mika Salo Circuitilta Alahärmän Powerparkista muutaman vuoden takaa. Karting-rata on upee mutta autojen nopeusrajoittimet naurettavan alhaalla. Taustakuva taas otettu Jyväskylän Neste-rallista muistaakseni Lankamaan erikoiskokeelta.
April 17, 2013 at 10:18pm

Figure 18: Likes and comments on a photo, and the time of sharing a photo and making the comments

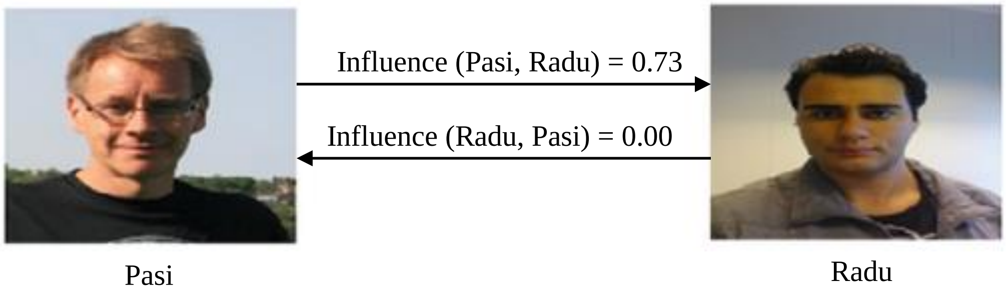


Figure 19: User influence on his friend

In Figure 19, we show influence values between Pasi and Radu. Table 1 shows only those users on whom Pasi and Radu exerts non-zero influence (influence > 0). Pasi influences 26 out of 76 friends (34%) whereas Radu influences 20 out of 274 friends (7%). Radu lies on 5th position in the list of friends influenced by Pasi. But, Pasi is not influenced by Radu. From this, we can conclude that Pasi influences higher percentage of users in his personal network than Radu.

Table 1: List of friends influenced by Pasi and Radu

Pasi				**Radu**			
Friend	Influence	Friend	Influence	Friend	Influence	Friend	Influence
Jussi	1.55	Pasi T.	0.36	Oili	3.33	Mohammad	0.33
Jukka	1.55	Esko	0.18	Katalin	2.33	Mikko	0.33
Oili	1.09	Merja	0.18	Karol	2.00	Arash	0.33
Sirpa	0.91	Jari	0.18	Chaitanya	2.00	Anton	0.33
Radu	**0.73**	Jarno	0.18	Ilea	2.00	Danut	0.33
Markku	0.64	Najlaa	0.18	Andrei	1.67	Iida	0.33
Staci	0.64	Margareta	0.18	Jukka	1.67	Sujan	0.33
Mohamed	0.55	Antero	0.18	Zhentian	1.67		
Jukka Vi.	0.45	Tero	0.09	Rudolf	1.33		
Zhentian	0.45	Keijo	0.09	Ville	0.67		
Tuomo	0.45	Kari	0.09	Bodea	0.33		
Tarja	0.45	Andrei	0.09	Najlaa	0.33		
Anne	0.36	Mikko	0.09	Paula	0.33		

User's total influence is calculated as the sum of a user's influence on all his friends:

$$Influence(u) = \sum_{i=1}^{k} Influence\ (u, f_i) \qquad (2)$$

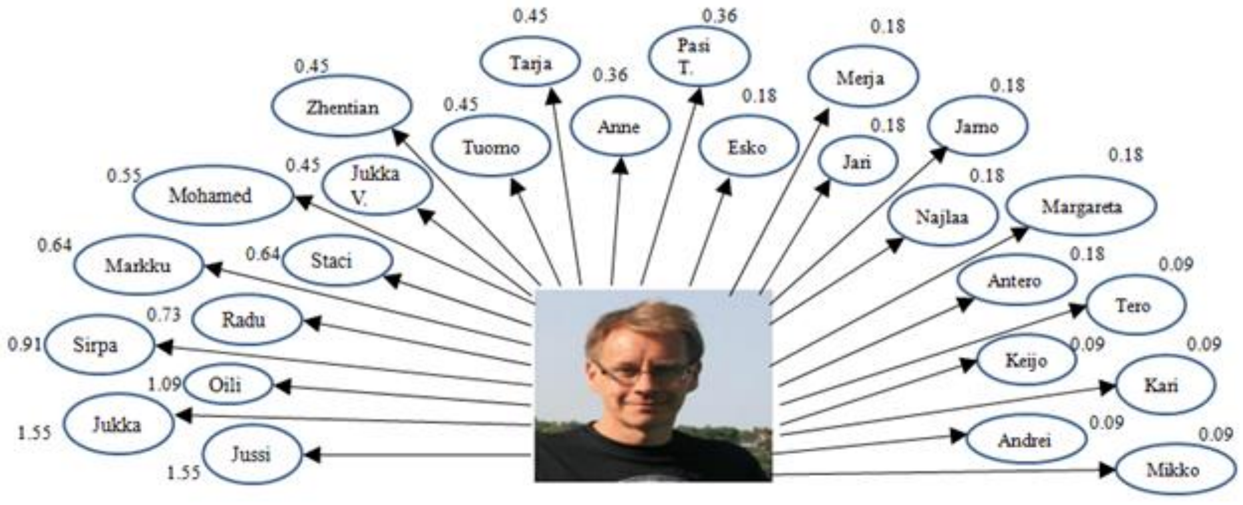
f_i = friend of user u

k = number of friends of user u

Figure 20 shows Pasi's influence to all his friends on whom he exerts non-zero influence. His total influence is 11.9. The maximum and minimum total influence values recorded in our sample of 98 users are 98 and 0.5 respectively. Pasi has 32nd rank among the 98 users.

The fact that the comment weighs 4 times than a like, is because writing a comment takes more effort than a like. The value 4 has been chosen based on the experiment conducted by Edge Rank Checker[7] which included random sampling of 5,500 Facebook pages and analysed more than 80,000 links posted. The links enabled calculation of number of clicks on the posts accurately. The number of clicks a link receives was analysed with respect to number of likes and comments on that particular post. These experiments show that one like and one comment resulted in 3 and 14 clicks respectively. From this, they concluded that a comment weighs 4 times a like.

[7] http://edgerankchecker.com/blog/2011/11/comments-4x-more-valuable-than-likes/



Pasi’s influence = **11.9**

Figure 20: Pasi’s friends on whom he has positive influence

Our own data also shows that likes are more common than comments as 50.6% photos have no comments at all. The data consists of 1,133 unique photos uploaded by 58 users on Facebook between May, 2013 and May, 2014.

3.2 Sample data

The sample data is collected using Mopsi. The data consists of the details of 58 Mopsi users (50% males, 50% females) who have 15,429 Facebook friends. It consists of users from 5 continents, see Table 2. Since European users form a significant subgroup in terms of size, we also conducted separate analysis on them.

Table 2: Percentage of users of different continents

Continent	% users in sample
Europe	69.0 %
Asia	13.8 %
North America	6.9 %
South America	6.9 %
Africa	3.4 %

The sample data includes the details of number of photos shared by user, likes and comments on photos, number of friends (male and female), user's gender, dominant gender in user's friend list and user's country. Table 3 shows the details of 20 Mopsi users which are a part of our sample.

Table 3: Details of 20 Mopsi users

User id	**Likes**	**Comments**	**Friends**	**Photos**	**Gender**	**Male Friends**	**Female Friends**	**Dominant Gender**	**Home Country**
10	89	3	436	10	Male	241	195	Male	Romania
85	39	5	112	7	Male	58	54	Male	Finland
119	5	6	27	3	Male	5	22	Female	Finland
406	192	22	566	6	Male	427	139	Male	India
1539	83	39	74	13	Female	23	51	Female	Germany
1552	130	34	315	8	Male	219	96	Male	Cambodia
1601	5	3	31	3	Female	18	13	Male	Germany
1613	12	7	36	8	Female	15	21	Female	Poland
1631	43	2	44	19	Female	16	28	Female	Bulgaria
1643	243	43	214	44	Male	95	119	Female	Romania
1653	50	10	529	28	Female	397	132	Male	Germany
1664	33	0	26	39	Male	22	4	Male	Mexico
1678	1	0	65	1	Female	10	55	Female	Slovakia
1683	0	4	30	8	Male	21	9	Male	Indonesia
1684	255	8	479	13	Female	237	242	Female	Slovakia
1692	62	3	105	17	Male	41	64	Female	Brazil
1693	11	11	53	10	Male	32	21	Male	Poland
1696	63	20	139	8	Male	99	40	Male	Croatia
1700	24	3	32	12	Male	19	13	Male	Germany
1728	363	78	278	70	Female	149	129	Male	Finland
Avg.	**85.1**	**15.0**	**179.5**	**16.3**	**60.0%** Male **40.0%** Female	**107.2**	**72.3**	**60.0%** Male **40.0%** Female	

User attributes: Number of photos, degree and gender

Number of photos

The user influence based on the number of photos uploaded during the last 1 year (red) period and that of all photos (blue) of all the users is shown in Figure 21.

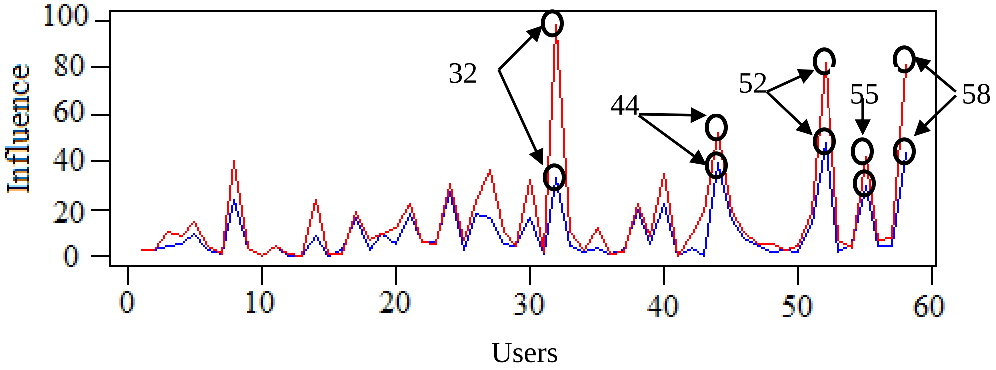


Figure 21: Comparison of influence based on photos uploaded during last year (red) and in total (blue)

We selected top 5 users out of 58 to observe how the influence values and influence rank varies when last 1 year photos were taken with respect to all photos, see Table 4. The result in the table shows that influence ranks of all users (except user 55) are different for both sets of photos.

User 32 has 4th rank when all photos are considered and 1st rank when last one year photos are considered. The improvement in user's influence rank shows that user is more active on social network during last one year. So, we can conclude that influence keeps on changing over time.

Besides, the variance in influence values across the 58 users using all photos is 127 and using last one year photos is 423. A high value of variance suggests that the data points are highly spread out from each other and around the mean, and thereby reduces the possibility of having a tie (data points with same value) among data points. Similarly, the high variance (423) in influence values for last one year photos reduces the possibility of a tie (users with same influence value) among different users. So, we have chosen the number of photos uploaded in last one year instead of all photos as we are interested in user's current influence value and reducing the possibility of a tie among users.

Table 4: Influence value and rank of top 5 users

	LAST YEAR		ALL	
User	**Influence**	**Rank**	**Influence**	**Rank**
Najlah (32)	33.1	4	98	1
Valdas.sketerskis (44)	39.4	3	52.3	4
Rusti.andrada (52)	47.5	1	82.3	2
Narcisa (55)	30.1	5	41.8	5
Bhalla.swati (58)	43.3	2	81	3

The distribution of the photos uploaded by all users in one year (see Figure 22) show a right skewed distribution. The skewness was also confirmed using normality test (see Appendix 1). So, we used median as the average for number of photos. The average is 10 for European users and 9 for all users.

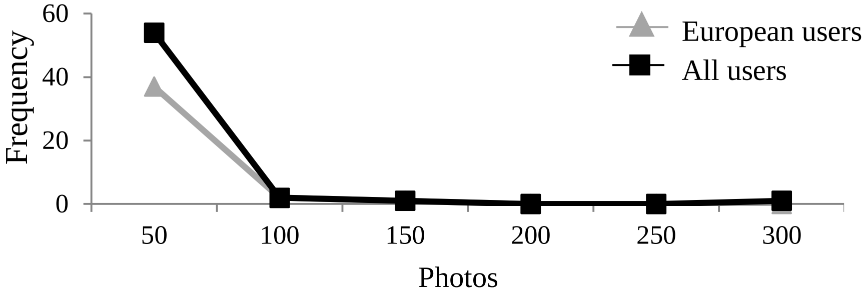


Figure 22: Frequency distribution for photos of European and all users.

Number of friends

The distribution of number of friends looks skewed (see Figure 23) and to confirm this, we used Anderson-Darling normality test (see Appendix 2).

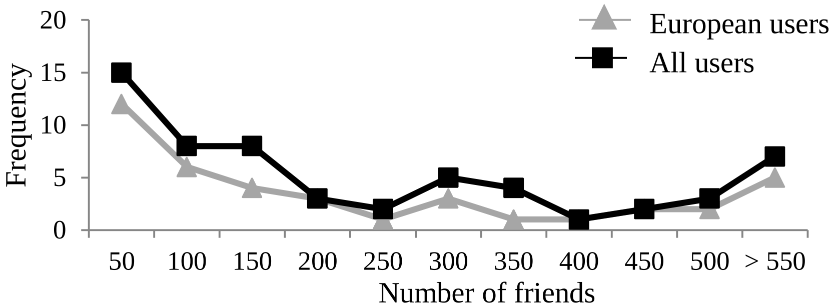


Figure 23: Frequency distribution for degree of European and all users

Since the distribution is skewed, we use median to measure the average number of friends of a user. For European users it is 129 and for all users it is 133. According to [25], the average number of Facebook friends of a user is 130 which is quite close to our average in both the sets. The frequency distribution of the number of friends of all users is quite similar to that of the European users, see Figure 23, as the user base has 68% European users.

Gender

We define dominant gender as the gender of majority (more than 50%) of user's friends. The results of our analysis show that most of the users are friends with same gender, see Table 5. This tendency is found to be 9.5% more among males than females when all users are considered. This finding might contribute in explaining the difference in user influence based on dominant gender.

Table 5: Dominant gender statistics of European and all users

Type	Users' gender	Male dominant	Female dominant
European	Male	78%	22%
	Female	27%	73%
All	Male	76%	24%
	Female	31%	69%

3.3 Relation between influence and user attributes

The influence value is calculated using equation 2. The median of the influence values was taken as the average since the distribution of the influence values is right skewed, see Figure 24. The average is 7.3 for all users and 7.9 for European users.

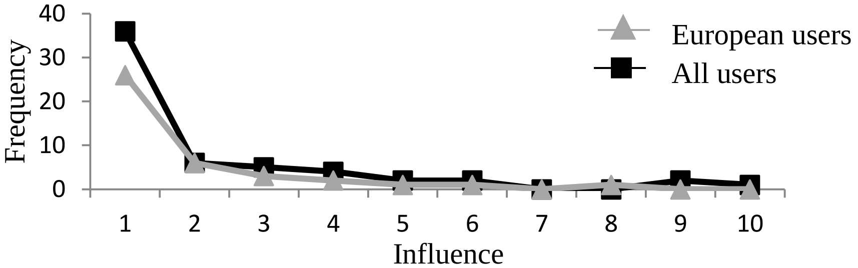


Figure 24: Frequency distribution for the influence values

We tested different hypotheses (discussed later) for the relation between average influence value and user attributes. The user attributes considered were gender, degree and number of photos. We used a nonparametric test for doing hypothesis testing as the frequency distribution of influence values and user attributes are non-normal, see Figure 22, 23 and 24. We chose Wilcoxon Rank-Sum Test [26] from a list of available nonparametric tests as it helps us to compare two different populations with respect to proposed hypothesis. For example, comparing males and females with respect to average influence value. All the statistical tests were performed using R[8] statistical software.

The further analysis considers all users as the sample set since the pattern for the distribution of influence values and user attributes for European users is similar to that for all users.

3.3.1 Gender of user and his friends

The spread of the influence values for both genders is quite similar except for a few outliers (super influential users), see Figure 25. Also, the average influence values for male and female users are 7.0 and 7.4, respectively.

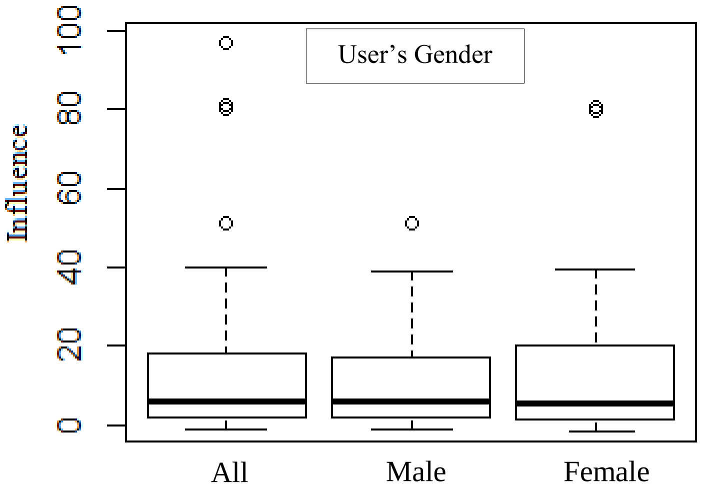


Figure 25: Box plot of distribution of influence values of users versus gender.

We tested the hypothesis that the average influence of a male user is equal to that of a female user. The test result shows that our hypothesis is true, see Appendix 3. Thus, gender is not a statistically significant parameter for a regression analysis of the influence values. We also analysed the relationship of likes and comments with gender. The average likes and comments for male, female and all users are shown in Table 6. Also, the distributions of likes and

[8] http://www.r-project.org/

comments based on gender are shown in Figure 26 and 27.

Table 6: Average likes and comments for male, female and all users

Data set	Average likes	Average comments
All users	44	9
Male users	33	6
Female users	61	9

Based on the average likes and comments values for both genders, we formulated two hypotheses: average likes of a female user is greater than that of a male user and average comments of a female user is greater than that of a male user. The test result shows that our first hypothesis is true (see Appendix 4). However, second hypothesis is false (see Appendix 5). Thus, we can conclude that average number of likes show a relation with user's gender i.e. a female user receives more likes (almost twice) than a male user, on average. In contrast, average comments do not show any significant difference.

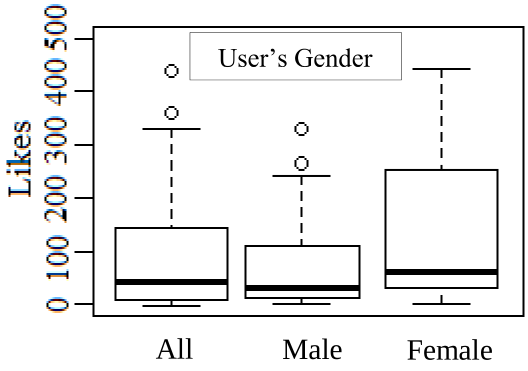


Figure 26: Distribution of Likes versus Gender

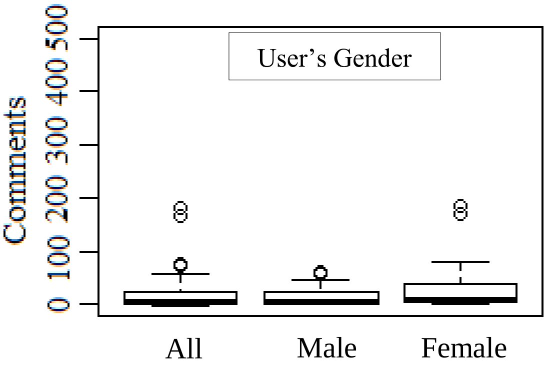


Figure 27: Distribution of Comments versus Gender

We also determined the relation between a user's influence and the dominant gender in his/her friend list. The box plot diagram (see Figure 28) shows that the spread of the influence of users with different dominant gender is different. The average influence values of the users with male and female as the dominant gender in their friend list are 10.0 and 6.5, respectively. So, we tested the hypothesis that the average influence of a user with male as the dominant gender in friend list is greater than that of a user with female as the dominant gender in friend list. The test result shows that the hypothesis is true (see Appendix 6), and the dominant gender in a user's friend list is a statistically significant parameter for regression analysis of the influence values. A possible reason of such phenomenon is that males are more active in making likes and comments on photos.

For further analysis of the relation between influence and dominant gender, the data is divided into two subsets based on a user's gender: male and female. The average influence values of the users of both subsets with respect to dominant gender are shown in Figure 28 and Table 7.

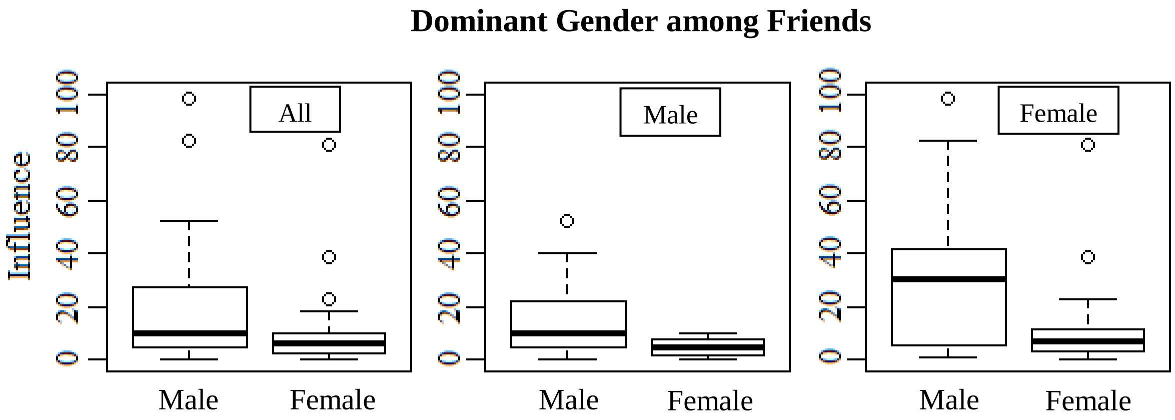


Figure 28: Box plot showing influence versus dominant gender in a user's friends list

We also studied the relationship of likes and comments with dominant gender for male, female and all users. The statistics of average likes and comments with dominant gender are shown in Table 7 and the distributions of these are shown in Figure 29.

Table 7: Average likes and comments with respect to dominant gender

Data set	Dominant gender	Average influence	Average Likes	Average comments
All users	Male	10.0	50	11
	Female	6.5	39	8
Male users	Male	10.0	42	10
	Female	4.3	10	3
Female users	Male	30.6	156	19
	Female	7.0	52	8

Based on the results shown in Table 7 and Figure 29, we can conclude that a user with male dominant friend list gets more likes and comments, on average. A female user with male dominant friend list seems to be best combination for attracting likes and comments. However, a female user likes and makes comments more on females' content.

Additionally, our analysis of the sample shows that for 75% of the users, the likes and comments made by the dominant gender are more than that made by the other gender. For example, if a user has a male dominant friend list, then the probability of having more than 50% of total likes and comments from male friends is 0.75.

The fact that average likes for a female user is almost twice than that of a male user provide a strong reason to perform separate regression analysis for influence based on a user's gender. The analysis will be helpful in targeting gender specific products to the influential user. This kind of analysis should be performed separately for different locations of interest as gender based interaction is different in different parts of the world.

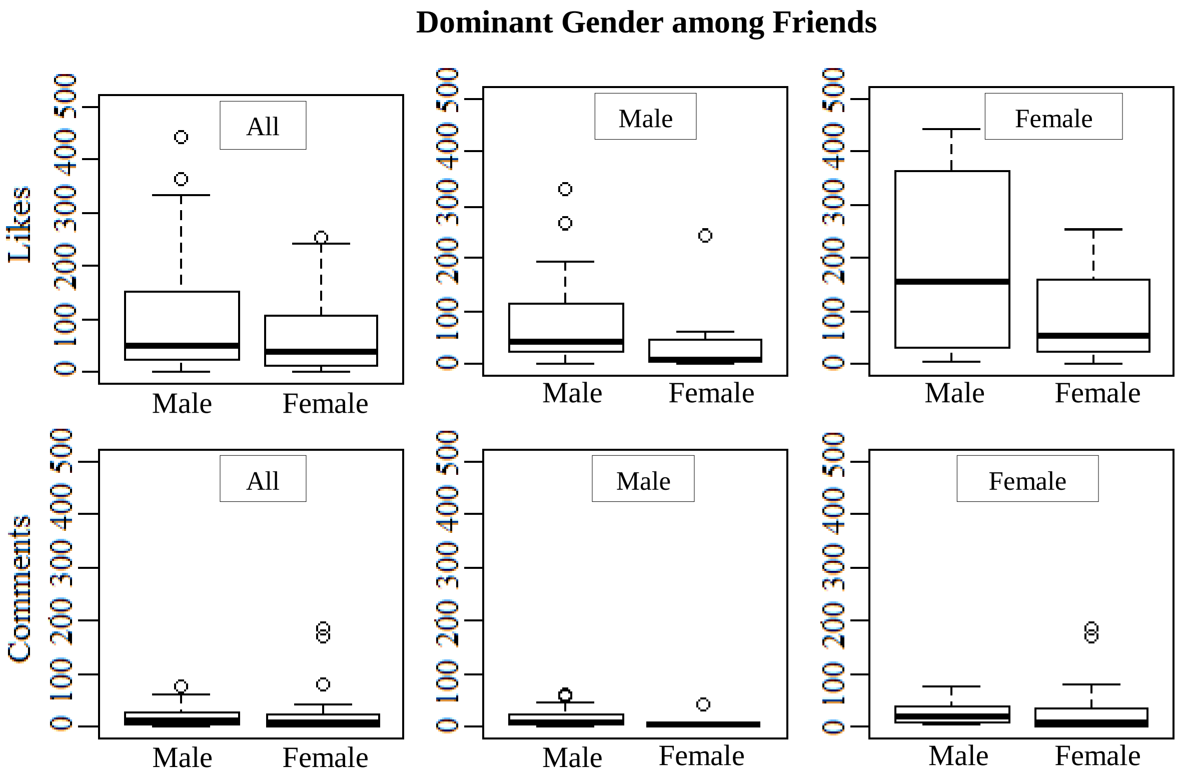


Figure 29: Distribution of likes and comments with respect to dominant gender of friends for male, female and all users.

3.3.2 Number of friends

For this, we determined the average influence of users with respect to their number of friends. We created two subsets based on the number of friends. One subset has users whose number of friends are less than or equal to 133 (average degree calculated using median). The other subset includes the users with friends more than 133.

The spread of the influence values for both subsets is shown in Figure 30. The average influence of the users who have less friends (friends≤133) and more friends (friends>133) are 4.3 and 18.6, respectively. A similar trend is observed for male and female data sets when analysed separately. Based on the results shown in Table 8, we tested the hypothesis that whether the influence of users with friends more than 133 is greater than that of users with friends less than or equal to 133. The test result (see Appendix 7) shows that our hypothesis is true and that number of friends is a statistically significant parameter for regression analysis of influence. So, we can conclude that users with friends more than 133 are more influential than users with friends less than 133, on average.

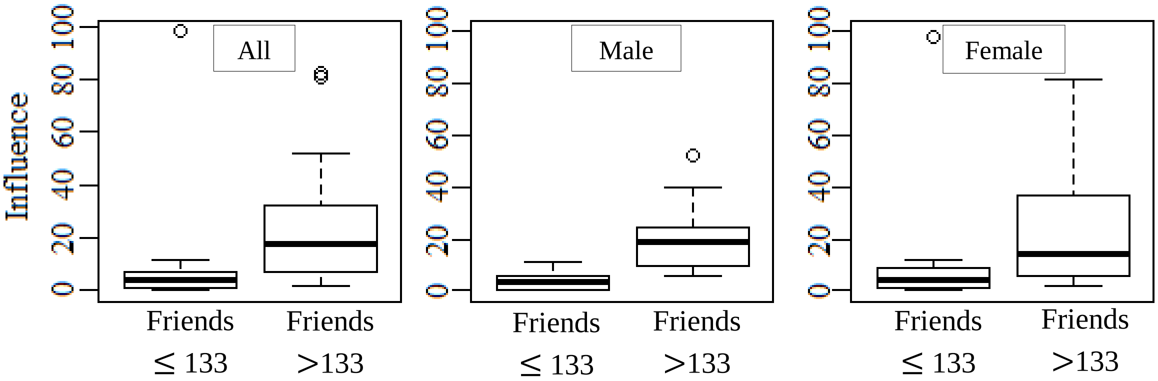


Figure 30: Box plot for influence of the users with number of friends less or more than 133

Table 8: Average influence of different users with respect to number of friends

Data set	Number of friends	Average influence
All users	Friends ≤ 133	4.3
	Friends > 133	18.6
Male users	Friends ≤ 133	3.6
	Friends > 133	19.0
Female users	Friends ≤ 133	4.8
	Friends > 133	14.4

Further, to investigate the relationship between influence and number of friends based on gender, we checked correlations between influence and number of friends using scatter plot, see Figure 31. The plot shows that correlation is stronger in case of males than females. If we remove three super influential users from the set of female users, then the correlation rises to 0.61 (from 0.37), which is still weaker than male users (0.78).

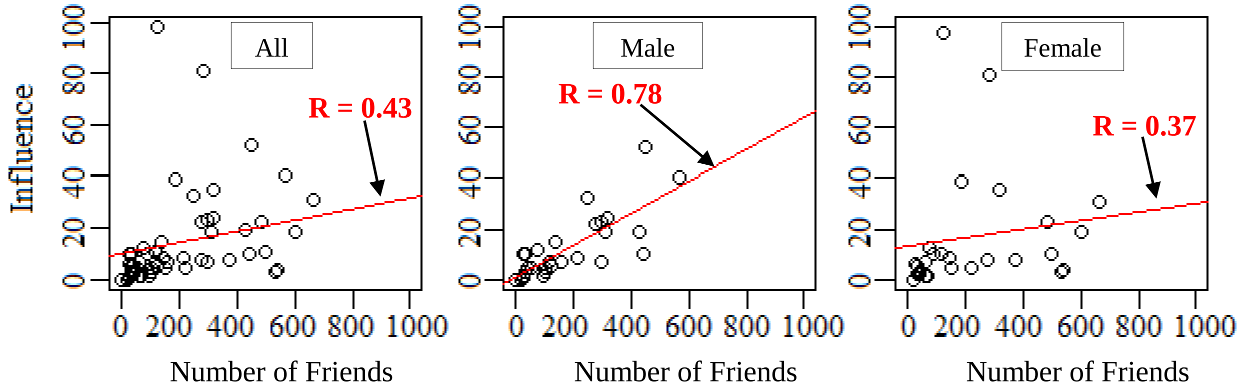
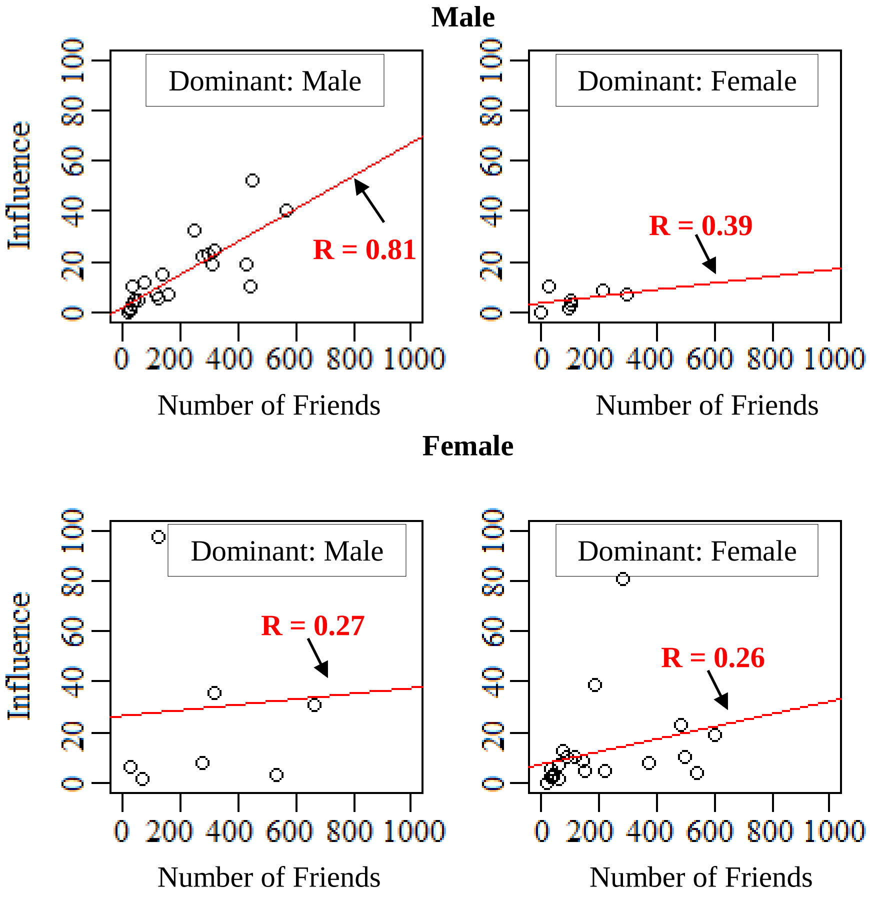


Figure 31: Correlation between influence and number of friends with respect to gender


Male

Female

Figure 32: Correlation between influence and number of friends with respect to dominant gender

In case of male users, the correlation between influence and number of friends is stronger with male dominant friend list than female, see Figure 32. Correlation values are nearly same in case of female users. However, if we remove three super influential users from the female set, new correlation values are 0.68 (male dominant) and 0.36 (female dominant).

So, we can conclude that a user with more male friends is expected to draw more likes or comments than that of a user with more female friends. Also, female users depend on the friendship (link) strength while making like or comment on friends' content. This provides a strong basis for performing separate regression analysis for influence based on user's gender.

3.3.3 Number of photos uploaded

For this, similar to the above mentioned analysis for relation with number of friends, we created two subsets of our sample set. One subset is of the users with less than or equal to 9 photos (average number of photos calculated using median) and the other subset of users with more than 9 photos.

The box plot diagram shows the spread of influence for both subsets in all the three data sets: all, male and female users, see Figure 33. Table 9 shows the average influence value for different data sets with respect to number of photos. The values in the table show that the average influence of a user with less than the average number of photos is lower than that of a user with more than average number of photos. This is true for the set of all users and for the set of female users. For male users, the reverse is true. Further, we tested the hypothesis that whether the average influence of a user with less photos (photos≤9) is equal to that of a user with more photos (photos > 9). The result of the hypothesis testing (see Appendix 8) shows that the average influence is same for both subsets and hence, it is a statistically insignificant parameter for regression analysis of the influence values.

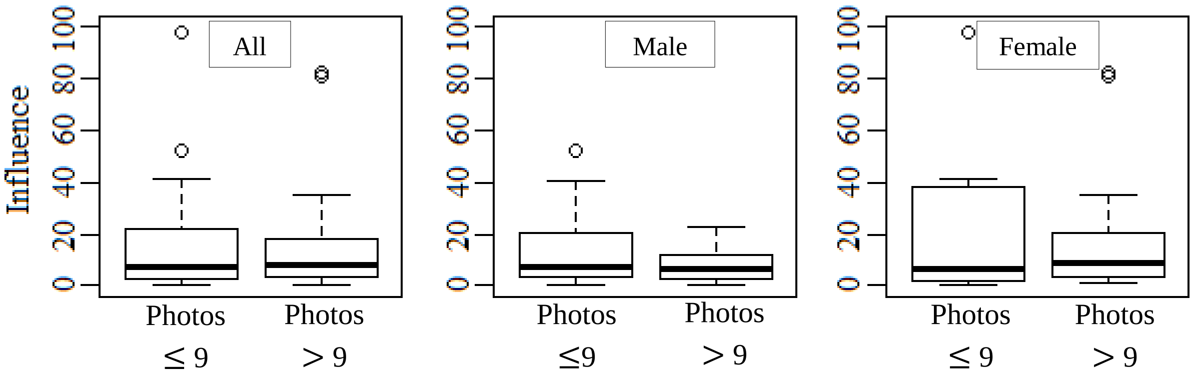


Figure 33: Box plot for the influence of users with less and more photos

Table 9: Average influence of different users with respect to the number of photos

Data set	Number of photos	Average influence
All users	Photos ≤ 9	7.0
	Photos > 9	8.4
Male users	Photos ≤ 9	7.0
	Photos > 9	6.5
Female users	Photos ≤ 9	6.4
	Photos > 9	8.6

4 Prediction of influence

The influence equation discussed in last chapter uses three parameters: likes, comments on photos and number of photos. To retrieve likes and comments from Facebook, we have to make one Graph API call per photo to Facebook and save the response to Mopsi database. The total time consumed in retrieving likes and comments and requirement of free database space depends on the number of photos. To minimize database space requirement, time consumption and number of Graph API calls, we propose an influence predictive model later in this chapter. The detailed description of advantages of using an influence predictive model is as follows:

Database space: The influence prediction model will require less input parameters than the original influence equation 1. We can therefore store less data to the database. For example, we took a sample of 100 users to compare database space requirement by both methods. The influence calculation using equation 1 requires 3.4 MB whereas influence predictive model requires just 15 KB of the database space.

Time consumption: The time required to retrieve the data of likes and comments on all photos of a user is directly proportional to the number of photos. For example, retrieving likes and comments of 200 photos might take even 200 s. However, the parameters required in the prediction model can be retrieved in just about 1 s.

Graph API calls: Facebook maintains the statistics for requests sent by each application, including number of calls made, CPU time spent and memory used by each application. When an application uses more resources than allowed, an error is generated. The influence calculation for 200 photos using equation 1 requires 200 API calls whereas influence prediction model requires only 1 API call.

The inputs considered for the influence model are number of friends and dominant gender as these have been proved to be statistically significant user attributes. We have developed separate models for male and female users because of two observations:

- Average likes of a female user is greater than that of a male user
- Correlation between influence and number of friends is stronger for males than females.

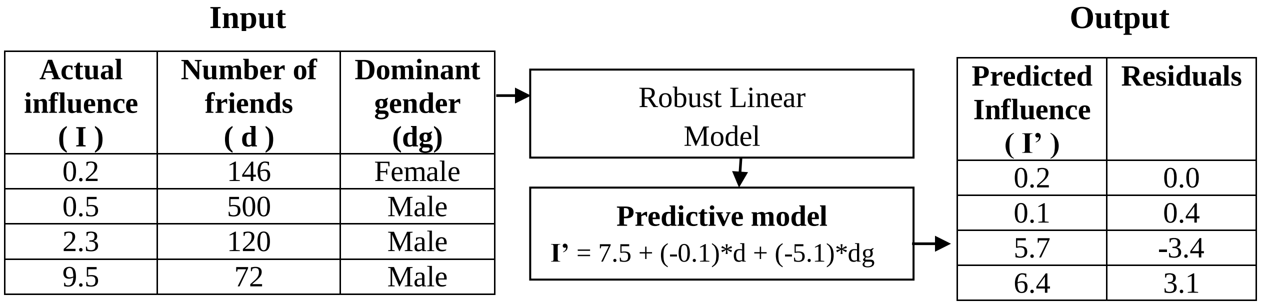
4.1 Modelling techniques

A predictive model can be used for predicting the unknown values of the response variable given the values of explanatory variable. In our influence predictive model, user influence is the response variable and user attributes (number of friends and dominant gender) are the explanatory variables. Our goal is to predict the value of user influence given the values of user attributes: number of friends and dominant gender. Linear regression techniques are the common modelling techniques used in predictive modelling. We evaluated two such techniques (non-weighted and weighted) to build influence predictive model. These were least squares and robust linear model. The weight used in the weighted version of the modelling techniques is:

$$w = \frac{1}{residual^2} = \frac{1}{(I - I')^2} \tag{3}$$

residual = difference between actual value and predicted value of influence. The actual influence value (I) is given as an input to the modelling technique for building the influence predictive model which results in predicted influence value (I'). Figure 34 shows an example of predicted influence values generated using actual influence values given in the input data. The residual values for each input data point are shown in the output table. The values of weights in this example are:

$$w_1 = \frac{1}{0.0^2} = N.A, \qquad w_2 = \frac{1}{0.4^2} = 6.25, \qquad w_3 = \frac{1}{-3.4^2} = 0.08, \qquad w_4 = \frac{1}{3.1^2} = 0.10$$



Input

Actual influence (I)	Number of friends (d)	Dominant gender (dg)
0.2	146	Female
0.5	500	Male
2.3	120	Male
9.5	72	Male

Robust Linear Model

Predictive model

I' = 7.5 + (-0.1)*d + (-5.1)*dg

Output

Predicted Influence (I')	Residuals
0.2	0.0
0.1	0.4
5.7	-3.4
6.4	3.1

Figure 34: Workflow for generating predicted values using actual values

The input data points which have lower residual values are assigned with higher weights. The weighted modelling technique uses the weight values as an additional input corresponding to each input data point.

In general, the choice of a linear modelling technique is based on the following assumptions [27] about the data:

1. The linearity of the relationship between response and explanatory variables
2. The variance of the residuals is constant i.e. homoskedastic
3. The residuals are uncorrelated
4. The residuals are normally distributed

Our sample data, in both subsets, violates some of the above stated assumptions (see Appendix 9). We used ordinary least square (OLS) to develop a predictive model and verified the above mentioned assumptions for the residuals. We found that the variance of residuals is heteroskedastic (see Figure 35), the residuals are correlated and their distribution is not normal (see Figure 36). In case of such violations the techniques such as ordinary least square (OLS) and weighted least square are expected to perform poorly and therefore, we do not use them.

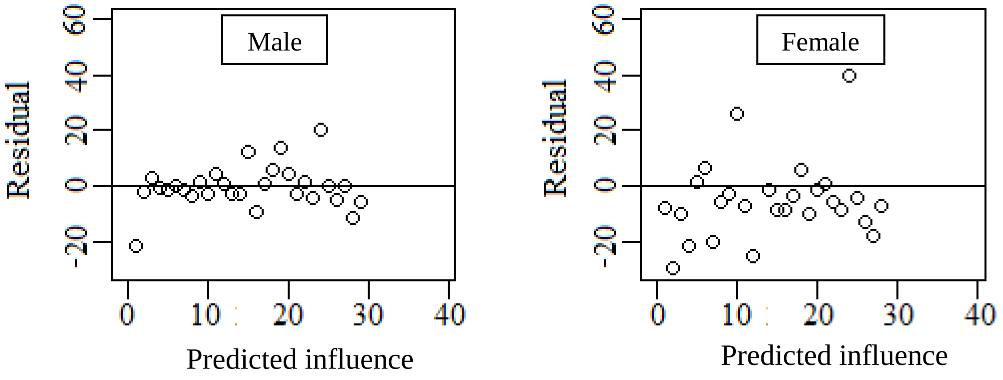


Figure 35: Residual plots showing heteroskedasticity of residuals

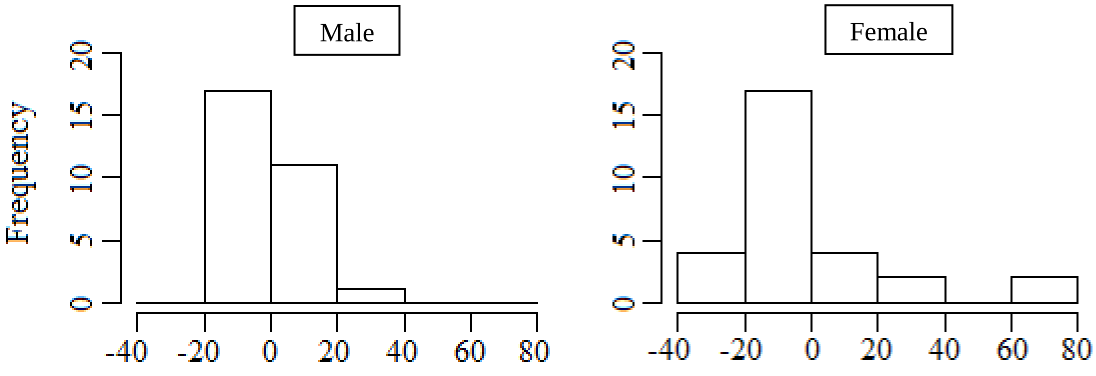


Figure 36: Residual distribution

In view of the non-normal distribution of residuals, as indicated by the application of the OLS technique, we have used the robust linear model [28] for our sample data. Also, a weighted version of this model was used to account for heteroskedasticity of the residuals. For this, we used weights ($1/\text{residual}^2$) [29] in the modelling equation. The results shown in Table 10 indicate that the robust linear model (weighted version) has a lower residual standard error and AIC[9][30] than that in the unweighted robust linear model. The residual standard error value of a model indicates the goodness of its fit. The lower the value of residual standard error the better is the fit. Hence, we considered robust linear model (weighted version) as a better fit for our data.

Table 10 shows the summary of the goodness of fit values (AIC and residual standard error) for all models. The results indicate that these values decrease when weights are applied in case of both least squares and robust linear models. So, we can conclude that application of weights improve the accuracy of models.

Table 10: Summary of the goodness of fit values for male and female data subsets

Male		
	AIC	**Residual standard error**
Ordinary least square	206	7.8
Weighted least square	157	0.9
Robust linear model	506	4.4
Robust linear model (weighted)	157	1.4
Female		
	AIC	**Residual standard error**
Ordinary least square	272	24.4
Weighted least square	211	1.0
Robust linear model	275	8.4
Robust linear model (weighted)	211	1.3

We evaluated the performance of the models by the strength of the association (correlation) between ranks assigned to the users based on the actual and predicted values of influence. The ranks were assigned using the actual influence (Equation 2) and by the prediction model (Equation 5). The most influential user was assigned the 1st rank and increasing rank values indicated reducing influence. Ties in influence values of users were resolved by assigning an average of their ranks to each user in the tie [31].

For the correlation analysis, we used the Spearman's rank correlation coefficient ρ which is

[9] AIC (Akaike Information Criterion) is a widely accepted criterion for a comparison of models. The lower the value of AIC, the better is the model.

determined using Equation 4. In the equation, d is the difference between the ranks assigned by the actual influence (Equation 2) and the prediction model (Equation 5), and n is the number of observations.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{4}$$

4.2 Influence predictive model

The influence predictive model is used to determine user's influence based on only two inputs about a user's friends: their number and the dominant gender, see Figure 37.

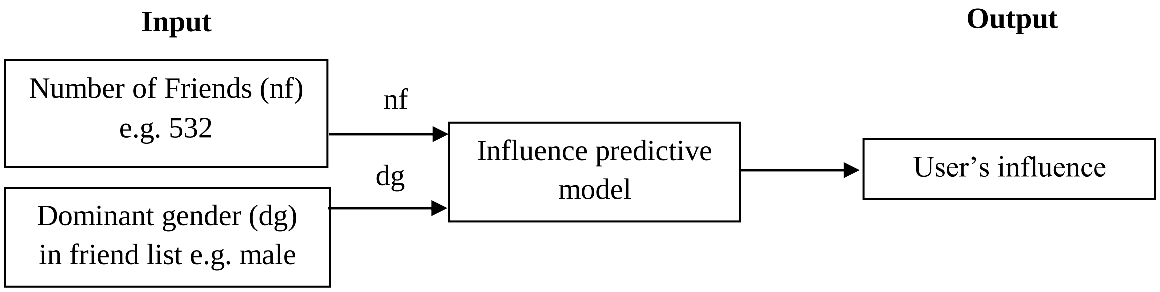

Input
Output
Number of Friends (nf) e.g. 532
nf
dg
Dominant gender (dg) in friend list e.g. male
Influence predictive model
User's influence

Figure 37: Influence predictive model workflow

The influence predictive model equation for a user is given below.

$$Influence(nf, dg) = c_{nf} \cdot nf + c_{dg} \cdot dg + c_i \tag{5}$$

where nf is number of friends and dg is dominant gender (0 for female and 1 for male).

The model has same inputs (nf and dg) for male and female users but the coefficients (c_{nf}, c_{dg} and c_i) are different, see Table 11. The values of coefficients are learnt by performing regression analysis between dependent variables (influence) and independent variables (nf and dg).

Table 11: Coefficients of influence equation for male and female users

Coefficients	Male	Female
c_{nf}	0.06	0.01
c_{dg}	5.04	12.66
c_i	-3.28	8.82

The AIC value is used to determine the quality of a model. The lower the AIC value the better is the model. It is used to select best model from a set of candidate models. It has no fixed range. AIC values of the models used for male and female are 157 and 211, see Table 10. These values can be used to compare current model (Equation 5) with possible future models even if future models will have additional parameters.

The residual plots for the models for males and females (Equation 5) are shown in Figure 38. These plots are different from the residual plots based on applying ordinary least squares, see Figure 35. We can observe reduction in the vertical spread of residuals in the residual plot and the clustering of residuals near black horizontal line which is essentially a zero residual line. From this, we can conclude that the residual plot confirms that the selection of weighted robust linear model as a modelling technique is a better choice. For detailed statistics, see Appendix 10 and 11.

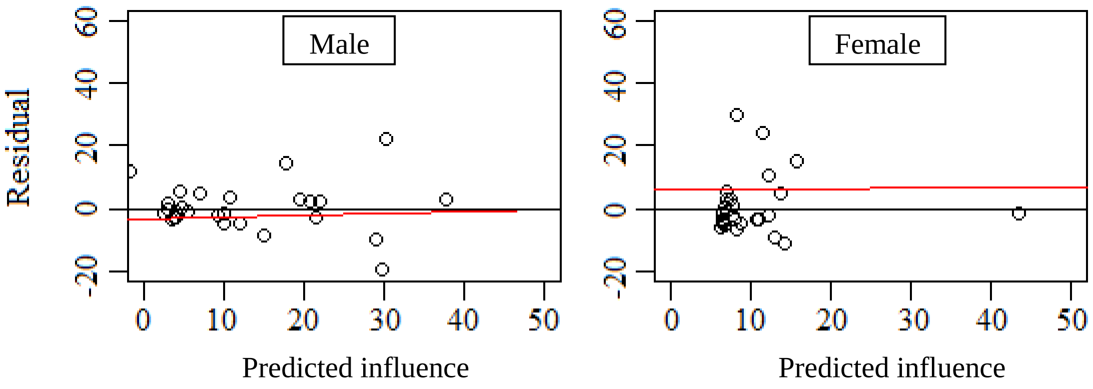


Figure 38: Residual plots resulted after using weighted robust linear model

The Spearman's rank correlation coefficients between influence ranks calculated using Equation 1 and Equation 5 for males and females are 0.85 and 0.53, respectively. This shows that the influence predictive model can fit the data better for male users than for female users. This result is justified by comparing correlation of actual influence (input to modelling technique) with number of friends for male and female users. The correlation values are 0.78 for males and 0.37 for females, see Figure 31. Higher correlation in case of male users resulted in better influence predictive model and hence, better prediction of influence values.

For male users, it is useful to develop separate versions of the influence predictive model based on dominant gender. The reason for this is the 70% difference in correlation coefficients between influence and number of friends based on dominant gender, see Figure 32. We can take advantage of this difference and improve the performance of influence prediction by building separate models based on dominant gender. For female users, there is only 3% difference in correlation between influence and number of friends based on dominant gender, see Figure 32. Hence, there is little benefit of developing separate versions of the model for female users, unless more data points widen the difference between the correlation coefficients. As part of future work, some other variables can be identified to improve the goodness of fit of influence predictive model.

5 Temporal analysis of user response

Temporal analysis is an analysis of any phenomenon with respect to time. In [32], researchers did temporal analysis of resharing of two photos, using histogram, which went viral on a social network. The photos were Obama Victory Photos (OVP) and Million Like Meme (MLM). They found that majority of reshares were done in the first 24 hours after the photos were posted: 96% for MLM and 90% for OVP. We also did a temporal analysis of the comments on a photo with respect to the time of uploading. This helps us identify the percentage of comments received in first 24 hours of uploading a photo. Besides, it will help us understand the relationship between the influence on user and the duration before he makes a comment.

The sample data includes 549 photos that had at least 1 comment and they were uploaded by 60 different users. Temporal analysis is conducted on 30 most popular and 30 randomly selected photos. The popularity of a photo, in this analysis, is based only on the number of comments since the timestamp of comments is only available from the Facebook database.

The analysis on the popular photos shows that most (80%) of these received more than 80% of total comments in first 24 hours, see Figure 39. A similar trend can be observed in randomly selected photos, see Figure 39. About 70% of these photos also received more than 80% comments in first 24 hours. The noticeable difference in Figure 39 is that 10% photos in random selection received 0% comments in first 24 hours. The total number of comments received by these photos lie between 1 and 3 and hence, they are relatively unpopular photos. We can conclude that most of the popular photos receive more than 80% of comments in first 24 hours.

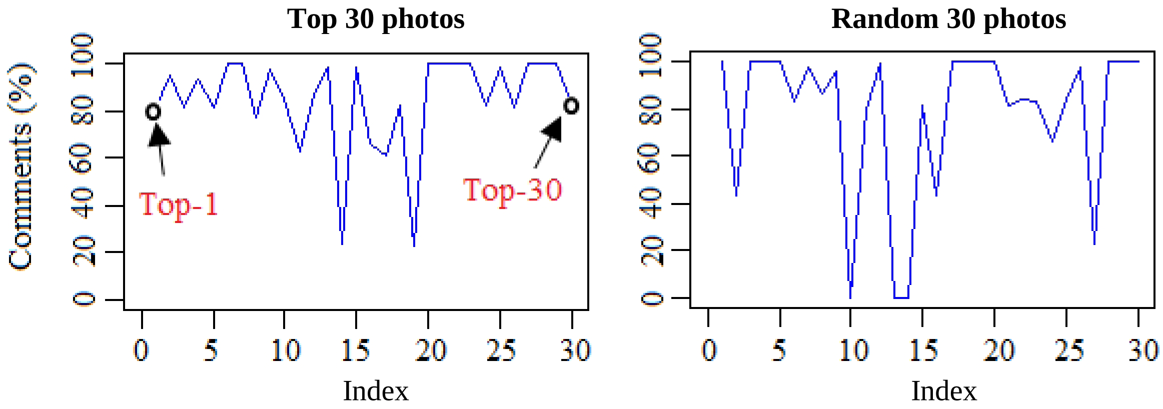


Figure 39: Number of comments received in first 24 hours of uploading

The distribution of comments received by the top 3 popular photos with respect to time (hours) is shown in Figure 40. The top 3 photos have got 24, 23 and 17 unique comments respectively. About 83%, 96% and 82% comments were received in first 24 hours respectively. The most popular photo is shown in Figure 41 which has total 40 comments with 24 unique users. We consider total comments on a photo as number of unique users commented on it.

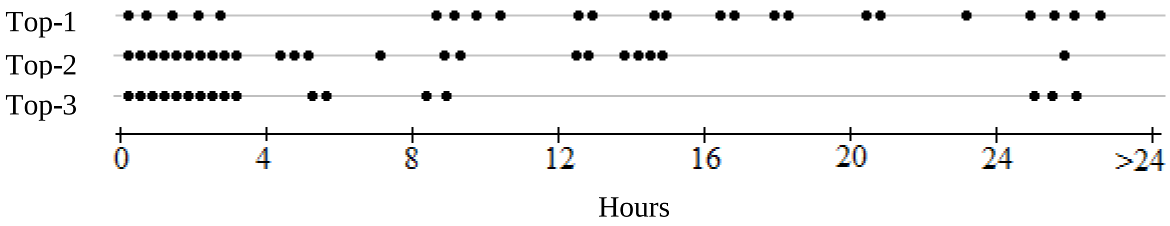
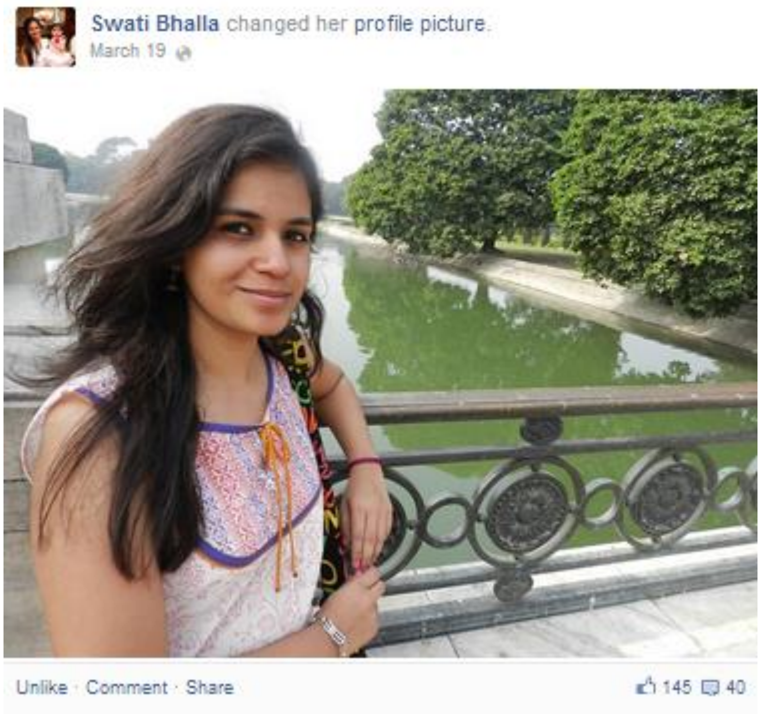


Figure 40: Number of comments received by the top 3 popular photos with respect to time


Swati Bhalla changed her profile picture.

March 19

Unlike · Comment · Share

145 40

Figure 41: Most popular photo

Users vary in their willingness to make comments on their friends' photos. To analyse this, we compared the influence on users with respect to time of making comment. We classified the users into two subsets: early comment makers (who comments in first hour) and late comment makers (who comments after 24 hours).

We selected a user *Swati* who has the most popular photo. The values of Swati's influence to all her 218 friends (calculated using Equation 1) are shown in Figure 42. Swati's 24 friends (early: 5 and late: 4) made comments on the most popular photo. The values show that Swati's influence on early comment makers is more than on the late ones, see Figure 43. To confirm this further, we took Swati's 6 photos (out of top 30) which had 14 early and 12 late comment makers and compared her influence on these comment makers, see Figure 44. We get a similar observation in both Figures 43 and 44 that the early comment makers are relatively more influenced than the late comment makers.

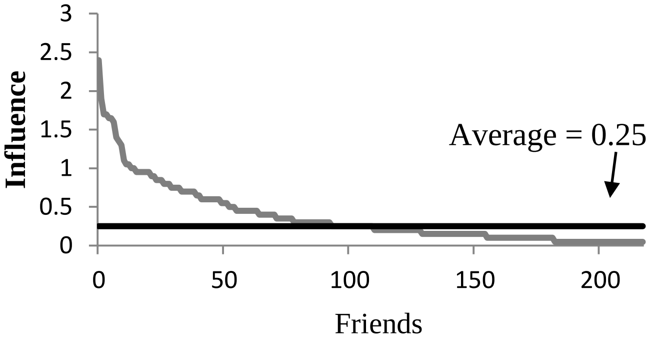


Figure 42: Values of influence of Swati to her friends

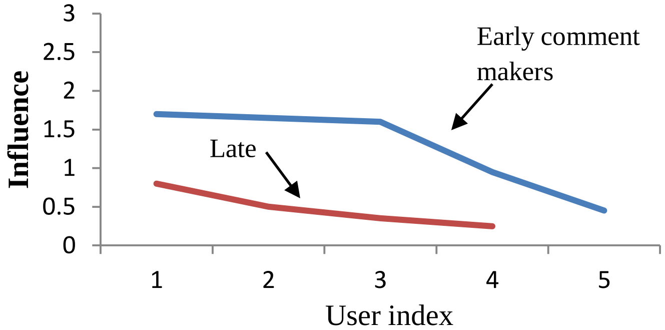


Figure 43: Influence on early (blue) and late (red) comment makers of most popular photo

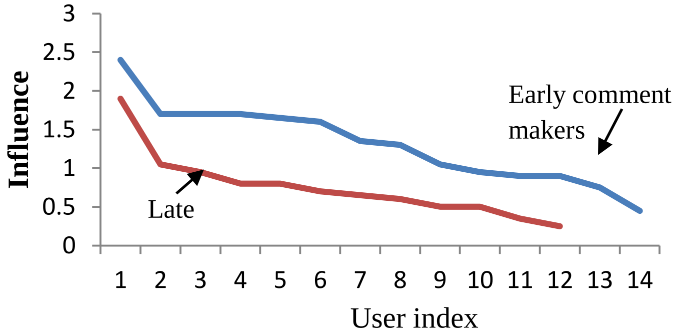


Figure 44: Influence on early (blue) and late (red) comment makers of 6 photos

We checked whether similar trend, as per the observations for Swati, is valid for other users who have popular photos as well. For this, we took a sample of 10 popular photos which had both early and late comment makers. The result shows that average values of influence are 1.25 to 6 times higher for early comment makers than that for late comment makers in 80% photos, see Figure 45. Late comment maker may also be related to the concept of collective behaviour. In this concept, an individual's involvement in any activity depends upon the proportion of other individuals already involved in the same activity [11].

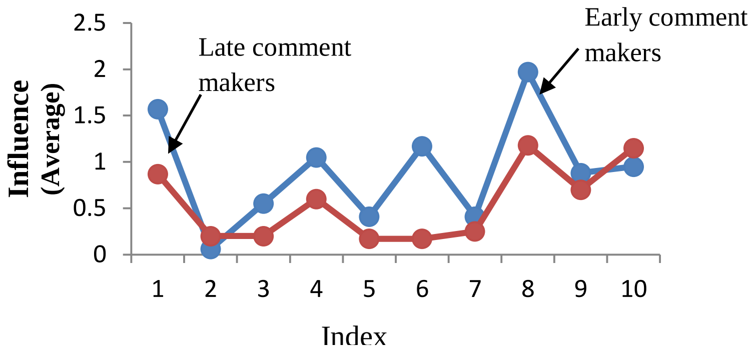


Figure 45: Average values of influence on early (blue) and late (red) comment makers

We can conclude that there is a statistically significant relationship between the influence on a user and the duration before he makes a comment. The early comment makers are expected to be the early product adopters in a user's network. So, the characteristics (e.g. location, age, gender) of such comment makers will be of interest while selecting the actual product for targeting in a user's network.

6 Popularity of a photo

Popularity of a photo uploaded by a Facebook user is calculated as follows:

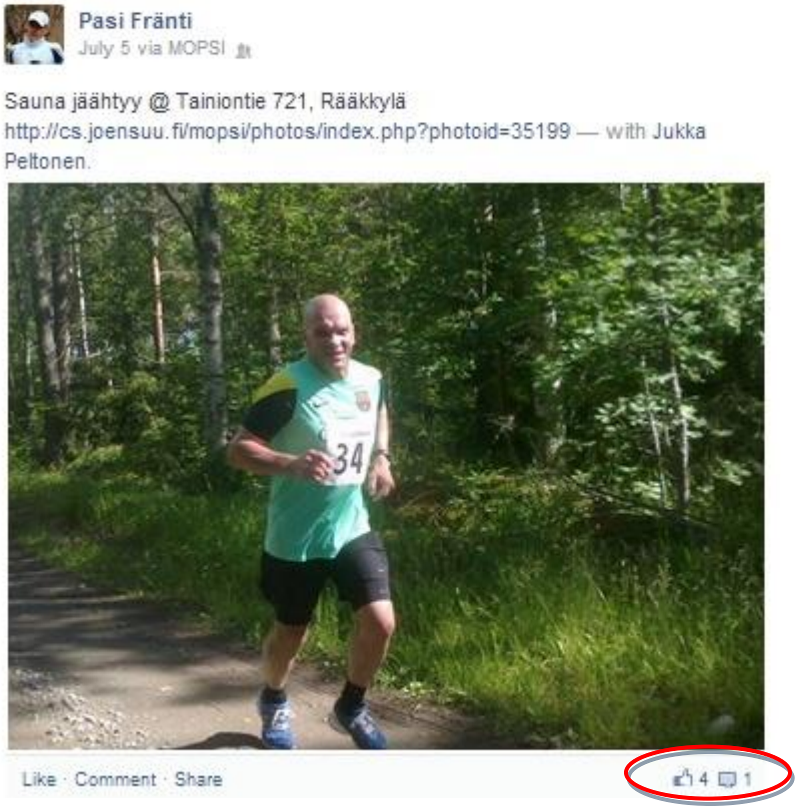
$$Popularity\ (p) = l + 4 \cdot\ c \qquad (6)$$

p = photo

l = number of likes on photo ‘p’

c = number of unique users who made comments on photo ‘p’

For example, the popularity of a photo in Figure 46 is 8 (4+4·1)



Pasi Fränti
July 5 via MOPSI

Sauna jäähtyy @ Tainiontie 721, Rääkkylä
http://cs.joensuu.fi/mopsi/photos/index.php?photoid=35199 — with Jukka Peltonen.

Figure 46: Photo with likes and comments

We analysed popularity of photos with respect to parameters such as day and time of uploading photo, degree, gender and dominant gender in the friends list of the user who uploaded the photo. The sample used for the analysis had 1,133 photos uploaded on Facebook by 63 users. The distribution of the popularity values of these photos is right skewed. We have used median as a parameter to calculate average popularity of a photo. The sample data set is divided further into two subsets: less popular (popularity ≤ 9) and more popular (popularity > 9) photos.

6.1 Day of uploading

In this, we analyse the relationship between the popularity and the day of uploading of a photo. The number of photos uploaded with respect to the days of a week is shown in Figure 47. The uploading activity is less frequent (15%), on Fridays and Saturdays than on other days of a week, see Figure 47. A possible reason could be that on these days users spend relatively more time on other activities (visiting clubs or meeting friends). The data on popularity shows that Monday and Sunday are better than other days as more than 50% of the photos uploaded on these days became popular (popularity > 9). Saturday is the worst performing day as 66% of the uploaded photos remain less popular. Overall, we can observe a declining trend in the percentage of more popular photos uploaded on weekdays (Monday to Thursday). Friday seems to deviate from this trend. Saturday and Sunday (weekend) show an upward trend.

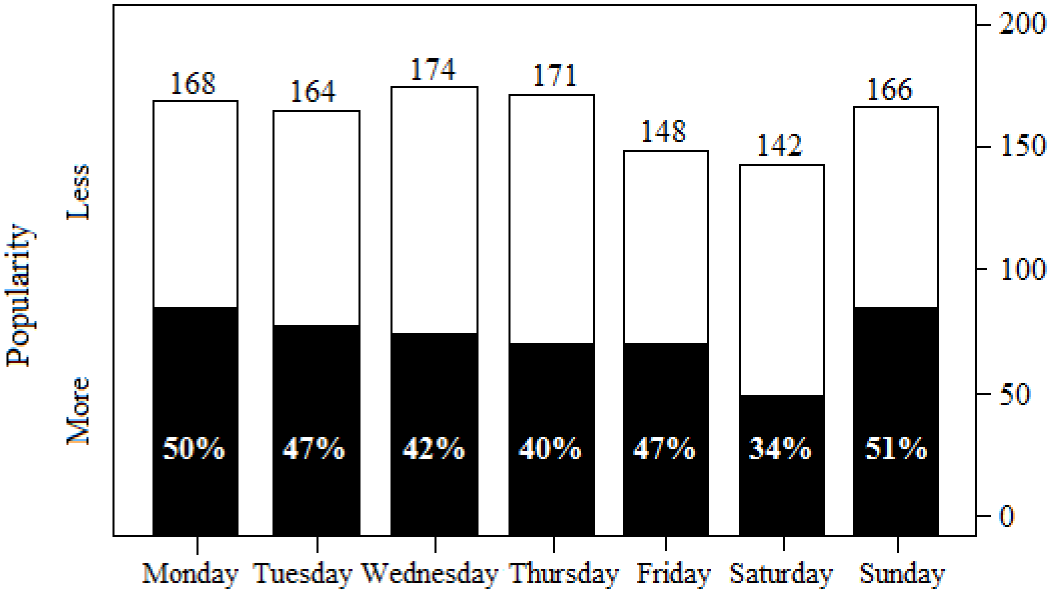


Figure 47: Popularity versus the day of uploading

The average popularity values of a photo uploaded from Monday to Sunday are 10.5, 10.0, 7.0, 7.0, 9.0, 6.5 and 11.5. We compare the average popularity of a photo uploaded on Sunday with that of all other days statistically. The results show that:

- Average popularity of a photo uploaded on Sunday is equal to that of a photo uploaded on Monday (see Appendix 12), Tuesday (see Appendix 13) and Friday (see Appendix 16).
- Average popularity of a photo uploaded on Sunday is higher than that of a photo uploaded on Wednesday (see Appendix 14), Thursday (see Appendix 15) and Saturday (see Appendix 17).

The statistical results described above show that popularity of a photo correlates on the day it is uploaded. Sunday is best for uploading photos as 51% of uploaded ones become popular. A possible reason could be that people relax by staying at their homes during Sunday and hence spend more time on social networks. So, we can conclude that the day of a week is an important parameter for predicting popularity of a photo.

6.2 Time of uploading

In this section, we analyse the relationship between a photo's popularity and time of uploading it. The number of photos uploaded during different times of day: morning (6 am-12 pm), afternoon (12 pm-6 pm), evening (6 pm-12 am) and night (12 am-6 am) is shown in Figure 48. The total number of photos uploaded increase from morning to evening, see Figure 48. However, the percentage of more popular photos decreases from morning to evening. The reason could be that as the number of photos available to a user increases, he becomes more selective while making a like or comment. This resulted in a situation that only 42% photos uploaded during evening became more popular. However, more than 47% photos uploaded during morning and night became more popular. Since less photos are available to users during these times, the situation enables users to see and engage with a higher proportion of uploaded photos.

The average popularity values of a photo uploaded during morning to night are 9, 8, 8 and 10. We compare the average popularity of a photo uploaded during morning with all other times of the day. The results show that:

- Average popularity of a photo uploaded during morning is higher than that of a photo uploaded during afternoon (see Appendix 18) and evening (see Appendix 19).

- Average popularity of a photo uploaded during morning is equal to that of a photo uploaded during night (see Appendix 20).

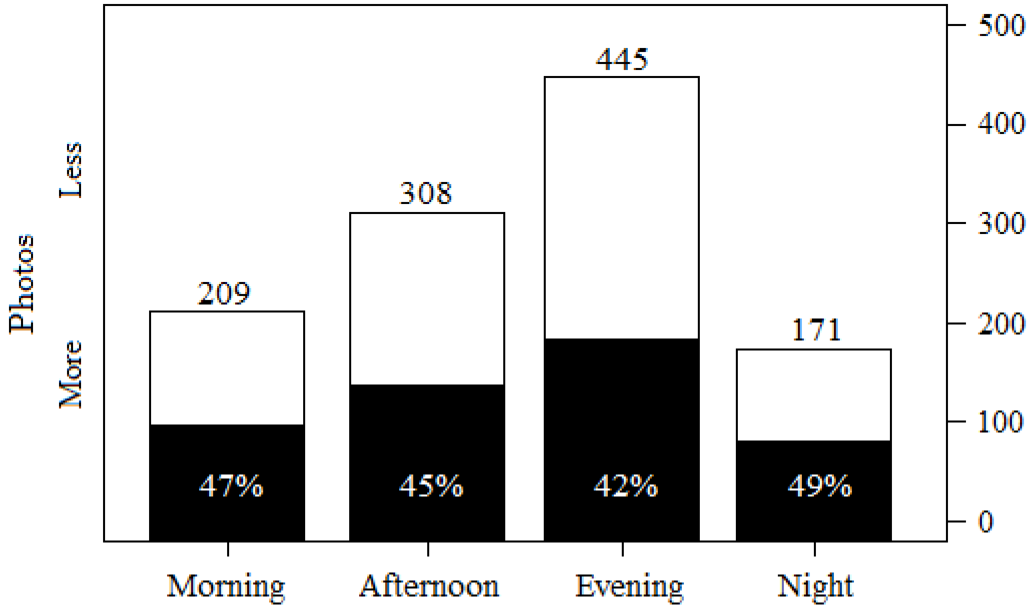


Figure 48: Popularity versus the time of uploading of a photo

We can conclude that a photo uploaded in morning will have higher popularity than uploaded during afternoon or evening, on average. Similar results are highlighted in the study conducted for more than 1,500 brands by Virtue[10]. The study shows that morning posts are more effective in terms of user engagement than those published in afternoon. However, it considered only comments only and not likes and shares. From the above stated observations, we can conclude that time of the day is an important attribute for predicting popularity of a photo.

6.3 Number of friends of user

In this section, we analyse the relationship of a photo's popularity with a user's friends. The analysis shows that users who have more friends have photos with higher popularity than other users, see Figure 49. The average popularity values of a photo uploaded by users who have more friends (friends > 133) and less friends (friends ≤133) are 11 and 5, respectively. Further, we tested the hypothesis that the average popularity value of a photo uploaded by a user with more

[10] http://mashable.com/2010/10/28/facebook-activity-study/

friends is higher than that uploaded by a user with less friends. The results show that our hypothesis is true, see Appendix 21.

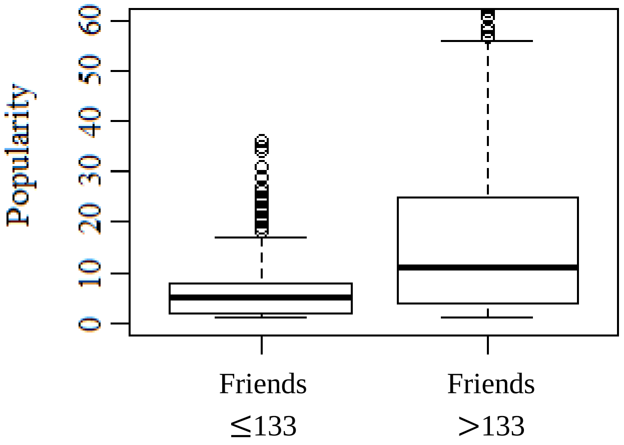


Figure 49: Box plot for the relationship between popularity value of a photo and number of friends

6.4 Gender of user and his friends

For this, we consider a random subset from the sample in which the number of photos uploaded by both male and female users was same. The relation between a user's gender and the popularity of photo he/she uploaded (see Figure 50) shows that average popularity of a photo uploaded by both male and female users is 8.0.

We also studied the relationship of popularity of photos with the dominant gender of friends for male, female and all users. The statistics of average popularity of a photo with dominant gender are shown in Table 12 and the distributions of these are shown in Figure 51. The average popularity of a photo uploaded by a user with male dominant friend list is higher than that of a user with female dominant friend list for two data sets (all users and male users), see Table 12. However, this is not valid for the data set of female users. To verify these observations statistically, we did hypothesis testing for all the three data sets.

For all users data set, the hypothesis that the average popularity value of a photo uploaded by a user with male as dominant gender is higher than that for a user with female as dominant gender is true, see Appendix 22. The same hypothesis is true for the male users' data set, see Appendix 23. However, for the female users we had a different hypothesis. We checked whether the average popularity value of a photo uploaded by a female user with female as dominant gender is higher than that with male as dominant gender, see Appendix 24. The result of the hypothesis is false and shows that dominant gender has no relation with popularity value for female users. Overall, we conclude that dominant gender is an important parameter for predicting popularity of a photo as it shows relation with photo's popularity for all and male users' data set. Besides, the hypothesis test shows that gender is also an important parameter as relation between dominant gender and photo's popularity is different for both male and female users.

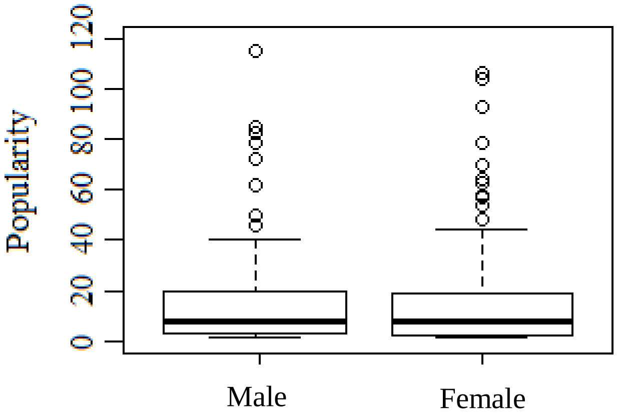


Figure 50: Box plot for the popularity of photos with respect to a user's gender

Table 12: Average popularity of a photo with respect to dominant gender

Data set	All users		Male users		Female users	
Dominant gender	Male	Female	Male	Female	Male	Female
Average popularity	9	7	9	4	6	8

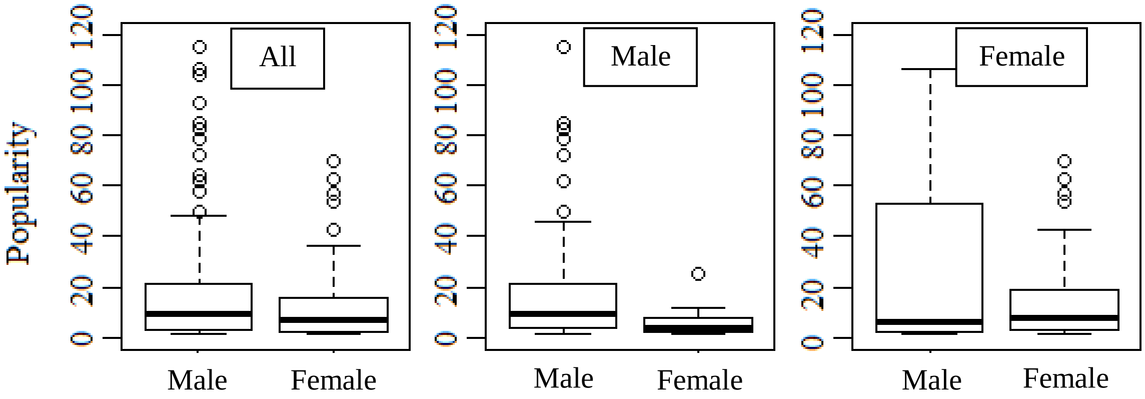


Figure 51: Popularity of photos uploaded by users with respect to the dominant gender

6.5 Popularity predictive model

We developed a predictive model to predict popularity of a newly uploaded photo. The advantage of using such model is that it suggests the time and day for uploading photo so as to maximize popularity based on user's previous data. The predictive model is user specific such that the set of input parameters are same for all users but with user specific coefficients of the predictive model. The values of coefficients depend upon the user's data of the previous photos. The model is developed by application of a suitable modelling technique (discussed in the next paragraph). The predictive model is used to generate popularity matrix which selects a particular time and day when a photo is expected to have maximum popularity. The resultant time and day can be used by a user to upload photo, see Figure 52.

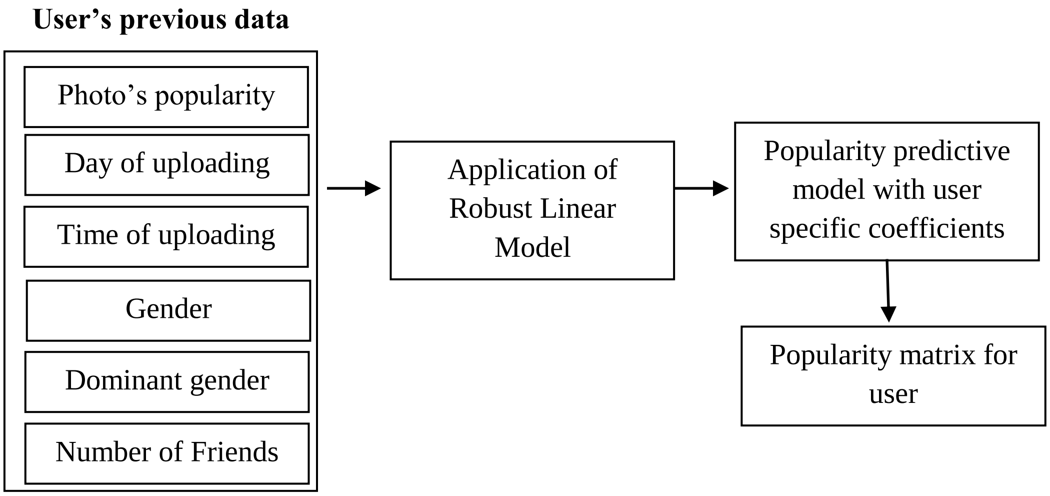
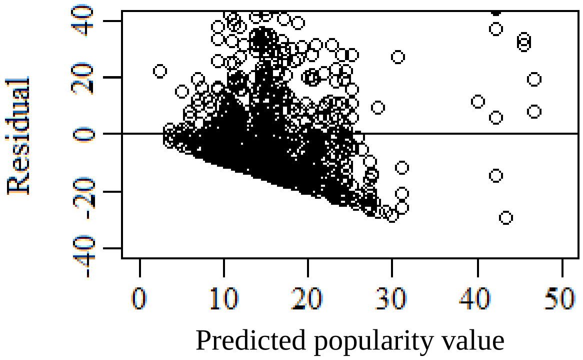

User's previous data
Photo's popularity
Day of uploading
Time of uploading
Gender
Dominant gender
Number of Friends
Application of Robust Linear Model
Popularity predictive model with user specific coefficients
Popularity matrix for user

Figure 52: Block diagram of popularity predictive model generation

Selection of modelling technique

We took a sample of 1133 photos to verify if the data satisfy the assumptions for application of linear model, see Appendix 9. We found that data violates some of the assumptions. Specifically, the variance of residuals is heteroskedastic (see Figure 53), the residuals are correlated and their distribution is not normal (see Figure 54). We therefore use the robust linear model (RLM) [28], a weighted version of this model was used to account for heteroskedasticity of the residuals. For this, we used weights (1/residual^2) [29] in the modelling equation.



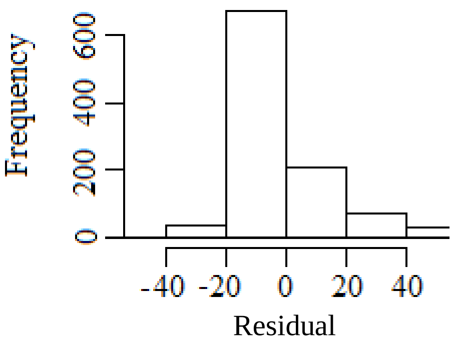


Figure 53: Residuals and predicted values of photo popularity

Figure 54: Distribution of residuals

The popularity predictive model has five input parameters. The equation of the model is described below:

$$Popularity = k_1 \cdot d + k_2 \cdot t + k_3 \cdot g + k_4 \cdot dg + k_5 \cdot f + k_6 \qquad (7)$$

d = day of uploading photo

t = time of uploading photo

g = gender

dg = dominant gender

f = total number of friends

The coefficients k_1 to k_6 have generic values and user specific values. The generic coefficient values are generated using data of more than one user. The user specific coefficient values are generated using the data for each user. The generic coefficient values can be used when user specific coefficient values cannot be obtained because of unavailability of user’s previous data. For example, if we wish to predict popularity of a photo for a user who didn’t upload any photo in past one year or more, we can use generic coefficient values. The generic coefficient values are shown in Table 13. The AIC value for the model (Equation 7) is 6974. For detailed summary statistics see Appendix 25. The different methods are compared in Table 14.

Table 13: Generic coefficient values

$\mathbf{k}_1$		$\mathbf{k}_2$		$\mathbf{k}_4$	
Monday	-1.69	Morning	-0.44	Male	1.55
Tuesday	-0.67	Afternoon	0.00	Female	0.00
Wednesday	-2.75	Evening	-2.87	$\mathbf{k}_5$	
Thursday	-2.43	Night	-0.53	0.01	
Friday	0.00	$\mathbf{k}_3$		$\mathbf{k}_6$	
Saturday	-2.71	Male	-0.24	9.80	
Sunday	0.35	Female	0.00		

Table 14: Performance of different models for predicting popularity of photo

Modeling techniques	**R-squared value**	**AIC**	**Residual standard error**
Ordinary least square	0.08	9460	22.89
Weighted least square	0.99	7372	0.99
Robust linear model	-	9519	11.96
Robust linear model (wt)	-	6974	1.47

For example, we apply robust linear model (rlm) on Pasi's (user) data to find the best day and time to upload the photo. The sample of Pasi's data shows that three data columns (Friends, gender and dominant gender) have singular values and hence rlm cannot be applied on these columns, see Table 15. In general, all the data columns are subject to change except gender.

Table 15: Sample of Pasi's previous data

Popularity	***d***	***T***	***f***	***G***	***Dg***
5	Tue	Evening	79	Male	Male
22	Wed	Afternoon	79	Male	Male
35	Tue	Morning	79	Male	Male
24	Wed	Evening	79	Male	Male

We obtained the user specific coefficient values of predictive model after applying RLM, see Table 16. The predictive model is used to generate popularity matrix for Pasi, see Table 17. Based on the values of matrix, the model predicts Wednesday morning is the best time for Pasi to upload photo.

Table 16: User specific coefficient values for Pasi

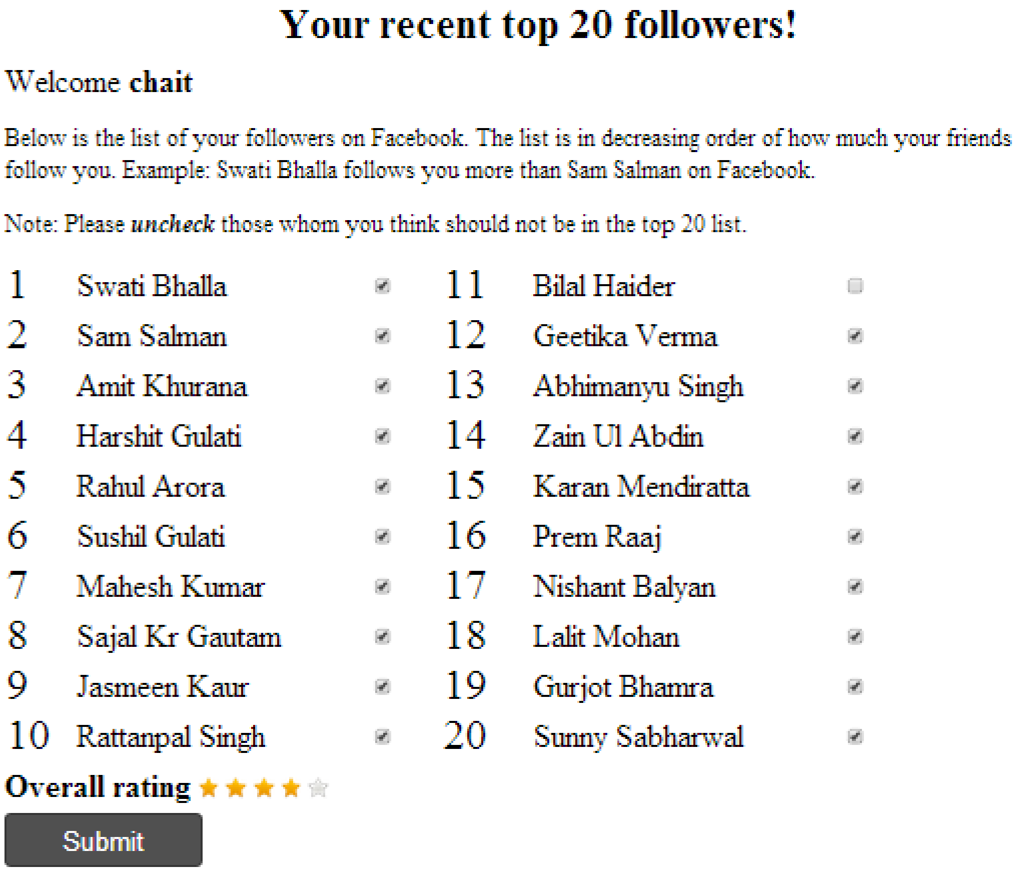
k_1		**k_2**	
Monday	-4.00	Morning	15.50
Tuesday	0.00	Afternoon	0.00
Wednesday	19.00	Evening	2.00
Thursday	-10.00	Night	0.00
Friday	0.00	**k_6**	
Saturday	-11.50	3.00	
Sunday	0.00		

Table 17: Popularity matrix for Pasi

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Morning	14.5	18.5	**37.5**	8.5	18.5	7.0	18.5
Afternoon	N.A	3.0	22.0	N.A	3.0	N.A	3.0
Evening	1.0	5.0	24.0	N.A	5.0	N.A	5.0
Night	N.A	3.0	22.0	N.A	3.0	N.A	3.0

7 Validation of results by online survey

The influence values were validated by an online survey for a sample of users. The survey shows a list of user's friends (followers) to each user, in decreasing order of being influenced, see Figure 55. A user can tick the checkbox if he believes that his friend follows him on Facebook. A user can also provide feedback on the overall ranking of followers on a scale of 1 (poor) to 5 (excellent). The feedback result of five Mopsi users is shown in Figures 56 and 57.



Your recent top 20 followers!

Welcome **chait**

Below is the list of your followers on Facebook. The list is in decreasing order of how much your friends follow you. Example: Swati Bhalla follows you more than Sam Salman on Facebook.

Note: Please ***uncheck*** those whom you think should not be in the top 20 list.

1	Swati Bhalla	☑	11	Bilal Haider	□
2	Sam Salman	☑	12	Geetika Verma	☑
3	Amit Khurana	☑	13	Abhimanyu Singh	☑
4	Harshit Gulati	☑	14	Zain Ul Abdin	☑
5	Rahul Arora	☑	15	Karan Mendiratta	☑
6	Sushil Gulati	☑	16	Prem Raaj	☑
7	Mahesh Kumar	☑	17	Nishant Balyan	☑
8	Sajal Kr Gautam	☑	18	Lalit Mohan	☑
9	Jasmeen Kaur	☑	19	Gurjot Bhamra	☑
10	Rattanpal Singh	☑	20	Sunny Sabharwal	☑

Overall rating ★★★★☆

Figure 55: Chait's top 20 followers

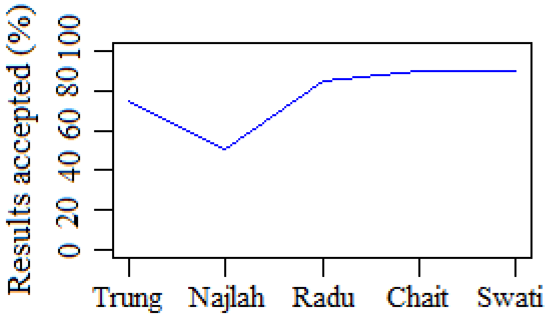


Figure 56: Percentage of results (followers) accepted

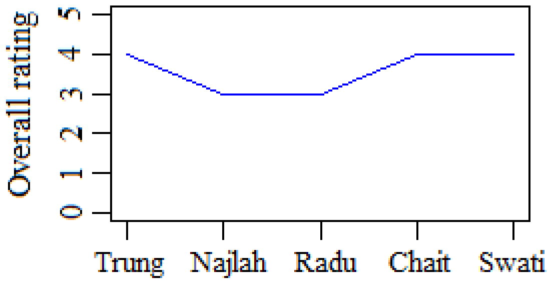


Figure 57: Overall rating

Figure 56 shows that 80% users agree with more than 75% followers mentioned in their respective list of top followers. The feedback on the ordering of top followers shows that 60% users rate it as 4 whereas others rate as 3, see Figure 57. The users' feedback seems promising and encourages us to perform this experiment on a larger scale.

8 Conclusions

A variety of user attributes were considered to analyse their relationship with user influence. The attributes include number of photos, gender, dominant gender in friend list and degree. Influence is determined using likes and comments.

Average influence of a male user is equal to that of a female user. We found that average likes of a female user is greater than that of a male user. However, comments do not show any statistical significant difference. Average influence, likes and comments of a user with male dominant friend list is greater than that of a user with female dominant friend list. A female user with male dominant friend list seems to be best combination for attracting likes and comments. At the same time, a female user likes and comments more on females' content. This shows that gender of user's friends have a positive correlation with user influence. Additionally, our analysis shows that for 75% of the users, the likes and comments made by the dominant gender are more than that made by the other gender.

User's total friends show positive correlation with user influence. Average influence of a user with more friends (friends>133) is greater than that of a user with less friends (friends≤133). The correlation between user influence and total friends is stronger for male than female users. We also found that the correlation values for a user is different when compared based on dominant gender. The correlation was much stronger for a user who has a male dominant friend list than female. This shows that a user with more male friends is expected to draw more likes or comments than that of a user with more female friends. Also, female users depend on the friendship strength while making like or comment on friends' content.

The number of photos did not show any correlation with user influence. Hence, the user average influence of a user with more photos is equal to that of a user with fewer photos.

Our statistical analysis shows that weighted robust linear model can be used as a modelling technique for developing influence predictive model. The model takes two inputs: number of friends and dominant gender. The same model with different coefficients is developed for male and female users. The predicted influence values were compared with actual influence values (calculated using likes and comments) using residual standard error and spearman's rank correlation. The residual standard errors for the predicted male influence and female influence values are 1.4 and 1.3. Also, the spearman's rank correlation between influence rank derived using original influence equation and predicted influence equation is 0.85 for males and 0.53 for females. This shows that predictive model fits the data well for male than female users. As a part of future work, some other attributes can be identified to further improve the predictive model.

Our analysis of comments with time shows that 80% of top 30 and 70% of 30 random photos received more than 80% comments in first 24 hours. We also analysed the relationship between influence on a user and time of making comment. The results show that early comment makers are relatively more influenced than late comment makers. These observations will be helpful in setting up a policy for advertising model. For example: a user who shares photo (advertisement) cannot delete or hide the photo from the social network before a fix number of hours after sharing. Any user violating such policy can lose the right for future discounts.

We analysed the relationship of photo's popularity with day and time of uploading, user's gender, total friends and dominant gender of friend list. We found that average popularity of a photo uploaded on Sunday is greater than that of a photo uploaded on Wednesday, Thursday or Saturday. We also found that average popularity of a photo uploaded during morning is greater than uploaded during afternoon or evening. Average popularity of a photo uploaded by a male user with male dominant friend list is greater than with female dominant friend list. However, such result was not found for female users. Average popularity of a photo uploaded by a user with more friends is greater than with less friends. We found all the user attributes are important for developing photo's popularity predictive model. The model predicts popularity value for each day and time which results in popularity matrix where row represents time and column represents day. This matrix predicts best day and time for uploading photo for a user so as to maximize popularity. Additional user attributes can be identified as a part of future research to further improve the goodness of fit of the predictive model.

REFERENCES

[1] M. Jamali, H. Abolhassani, “Different Aspects of Social Network Analysis”, *IEEE International Conference on Web Intelligence*, pp. 66-72, 2006

[2] E. M. Rogers, *Diffusion of Innovations*, The Free Press, 1962.

[3] K. Waga, A. Tabarcea, M. Chen and P. Fränti, "Detecting Movement Type by Route Segmentation and Classification", *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, Pittsburgh, USA, 2012

[4] E. Katz, P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*, The Free Press, 1955.

[5] M. Trusov, A.V. Bodapati, R. E. Bucklin, “Determining Influential Users in Internet Social Networks”, *Journal of Marketing Research*, pp. 643-658, 2010

[6] J. E. Phelps, R. Lewis, L. Mobilio, D. Perry, N. Raman, “Viral Marketing or Electronic Word-of-Mouth Advertising: Examining Consumer Response to Pass Along Email,” *Journal of Advertising Research*, vol. 44, pp 333-348, 2004

[7] P.F. Lazarsfeld, B. Berelson, and H. Gaudet, “The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign”, *New York: Columbia University Press*, pp. xxxiii, 178, 1948.

[8] E. Katz, “The Two-Step Flow of Communication: An Up-to-Date Report on an Hypothesis”, *Public Opinion Quarterly*, vol. 21, pp. 61-78, 1957

[9] T.W. Valentene, “Network Models of the Diffusion of Innovations”, *Cresskil, NJ: Hampton Press*, 1995

[10] D. Kempe, J. Kleinberg, E. Tardos,” Maximizing the Spread of Influence Through a Social Network”, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 137-146, 2003

[11] M. Granovetter, “Threshold Models of Collective Behavior”, *The American Journal of Sociology*, vol. 83, pp. 1420-1443, 1978.

[12] D.J.Watts, P.S. Dodds, “Influentials, Networks, and Public Opinion Formation”, *Journal of Consumer Research*, vol. 34, pp. 441-458, 2007

[13] P.S.Dodds, D.J.Watts, “Universal Behavior in a Generalized Model of Contagion”, *Physical Review Letters*, vol. 92, 2004

[14] D. Centola, M. Macy,”Complex Contagions and the Weakness of Long Ties”, *American Journal of Sociology*, vol. 113, pp. 702-734, 2007

[15] A. Anagnostopoulos, R. Kumar, M. Mahdian, “Influence and Correlation in Social Networks”, *ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 7-15, 2008

[16] S. Aral, “Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion”, *INFORMS Marketing Science Conference,* 2010

[17] M. Cha, H. Haddadi, F. Benevenuto, K. P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy”, *International AAAI Conference on Weblogs and Social Media*, Washington, DC, USA, 2010.

[18] J. Berry, E. Keller, *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy,* Free Press, 2003.

[19] M. Gladwell, “The Tipping Point: How Little Things Can Make a Big Difference”, *Back Bay Books,* 2002.

[20] S. Aral, D. Walker. "Identifying Influential and Susceptible Members of Social Networks" *SCIENCE*, pp. 337-341, 2012.

[21] D. M. Romero, W. Galuba, S. Asur, B. A. Huberman, “Influence and Passivity in Social Media”, *International Conference Companion on World Wide Web*, pp. 113-114, 2011

[22] S. Wasserman, K. Faust, *Social Network Analysis*, Cambridge University Press, 1994

[23] B. Hajian, T. White, “Modelling Influence in a Social Network”, *IEEE International Conference on Social Computing*, pp 497-500, 2011

[24] S. Hangal, D. MacLean, M.S. Lam, J. Heer, “All Friends are Not Equal: Using Weights in Social Graphs to Improve Search”, *SNA-KDD Workshop*, Washington, USA, 2010.

[25] T. Steiner, R. Verborgh, J. G. Vallés, R. V. Walle, “Adding Meaning to Facebook Microposts via a Mash-Up API and Tracking its Data Provenance”, *International Conference on*

Next Generation Web Services Practices (NWeSP), pp. 342-345, 2011

[26] M. J. Crawley, “Chapter 8: Classical Tests”, *The R Book*, A John Wiley & Sons Publication, 2013.

[27] P. Bühlmann, M. Mächler, “Chapter 1: Multiple Linear Regression”, *Computational Statistics*, pp.1-15, 2014

[28] R. Bellio, L. Ventura, “Chapter 3: Estimation in Scale and Regression Models”, *An Introduction to Robust Estimation with R Functions*, pp. 17-32, 2005

[29] R. J. Freund, R. C. Littell, *SAS System for Regression 3rd Edition*, A John Wiley & Sons Publication, pp. 84, 2000

[30] H. Akaike, “Information Theory and an Extension of the Maximum Likelihood Principle”, *Breakthroughs in Statistics*, pp. 610-624, 1992

[31] W. Buck, “Tests of Significance for Point-Biserial Rank Correlation Coefficients in the Presence of Ties”, *Biometrical Journal,* vol. 22, pp. 153-158, 2007

[32] P. A. Dow, L. A. Adamic, A.Friggeri, “The Anatomy of Large Facebook Cascades”, *International AAAI Conference on Weblogs and Social Media*, Boston, USA, 2013.

Appendices: Statistics tests conducted.

All the following tests were conducted using R statistical software.

Appendix 1: Normality test on distribution of photos

Test: Anderson-Darling normality test

Null Hypothesis (H_o): Distribution of photos is normal.

Alternate Hypothesis (H_a): Distribution of photos is not normal (skewed)

Result: A = 6.6037, p = 2.2e-16 < 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 2: Normality test on distribution of number of friends

Test: Anderson-Darling normality test

Null Hypothesis (H_o): Distribution of number of friends is normal

Alternate Hypothesis (H_a): Distribution of number of friends is not normal.

Result: A = 3.0501, p = 9.334e-08 < 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 3: Hypothesis test on average influence of a male and female user

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average influence of male and female users is equal.

Alternate Hypothesis (H_a): Average influence of male and female users is not equal.

Result: W = 457.5, p = 0.565 > 0.05 ⇒ Null hypothesis is **ACCEPTED**.

Appendix 4: Hypothesis test on average likes of a male and female user

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average likes of female and male users is equal.

Alternate Hypothesis (H_a): Average likes of female is greater than male user.

Result: W = 527.5, p = 0.04878 < 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 5: Hypothesis test on average comments of a male and female user

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average comments of female and male users is equal.

Alternate Hypothesis (H_a): Average comments of female is greater than male user.

Result: W = 496.5, p = 0.1198 > 0.05 ⇒ Null hypothesis is **ACCEPTED**.

Appendix 6: Hypothesis test on average influence of a user with respect to dominant gender

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average influence of users with female and male dominant gender is equal.

Alternate Hypothesis (H_a): Average influence of users with female dominant gender is less than users with male dominant gender.

Result: W = 301.5, $p = 0.03406 < 0.05$ ⇒ Null hypothesis is **REJECTED**.

Appendix 7: Hypothesis test on average influence of a user with respect to number of friends

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average influence of users with more and less friends is equal.

Alternate Hypothesis (H_a): Average influence of users with more is greater than users with less friends.

Result: W = 639, $p = 2.468\text{e-}05 < 0.05$ ⇒ Null hypothesis is **REJECTED**.

Appendix 8: Hypothesis test on average influence of a user with respect to photos

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average influence of users with less and more photos is equal.

Alternate Hypothesis (H_a): Average influence of users with less and more photos is not equal.

Result: W = 355.5, $p = 0.5708 > 0.05$ ⇒ Null hypothesis is **ACCEPTED**.

Appendix 9: Assumptions for application of a linear model

The following assumptions [27] should be valid for application of a linear model:

1. The relationship between response and each of the explanatory variables should be linear (see Figure 58). If this assumption is violated, we should use models other than linear models.

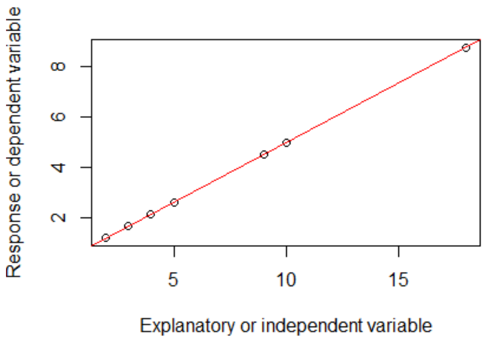


Figure 58: Linear relationship between response and explanatory variables

2. The variance of the residuals with respect to predicted values should be constant i.e. homoskedastic. Homoskedastic residuals can be seen in the left plot of Figure 59. Center and right plots of shows the residuals with non-constant variance and hence called as heteroskedastic residuals.

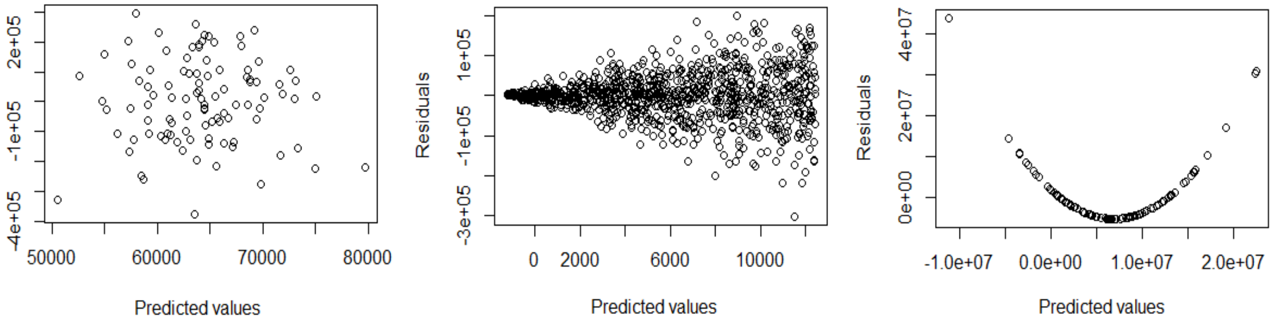


Figure 59: Left plot shows constant variance and other plots show non-constant variance.

3. The residuals are uncorrelated i.e. ∑ $(X_i .e)_i = 0$ and ∑ $(Y_i .e)_i = 0$ and where X_i, Y_i and e_i are independent (explanatory) variable, dependent (response) variable and residuals.

4. The residuals should be normally distributed. We have shown the example of normal distribution of residuals in Figure 60.

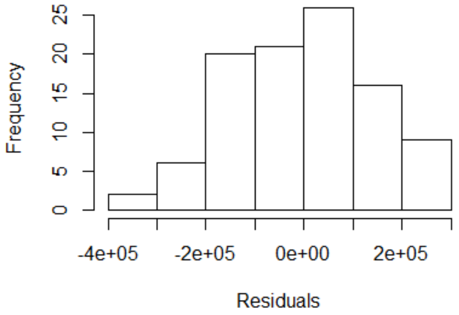


Figure 60: Histogram showing distribution of residuals is normal

Appendix 10: Summary statistics of influence predictive model for male

The summary statistics of the prediction model for influence of male users is given below.

```
rlm(formula=male$Influence~ male$Friends + male$Dominant.gender, weights = (1/square_error_male))
Residuals:
Min             1Q              Median          3Q              Max
-1.2353         -0.9449         -0.5859         0.9417          1.2113
Coefficients:
                                Value           Std. Error      t value
(Intercept)                     -3.2870         0.6314          -5.2061
male$Friends                    0.0652          0.0036          17.9500
male$Dominant.genderm           5.0415          0.6012          8.3855
Residual standard error: 1.401 on 26 degrees of freedom
```

Appendix 11: Summary statistics of influence predictive model for female

The summary statistics of the prediction model for influence of female users is given below.

```
rlm(formula=female$Influence~    female$Friends    +    female$Dominant.gender,    weights    =
(1/square_error_female))
Residuals:
Min             1Q              Median          3Q              Max
-1.2154         -0.8906         -0.8059         0.6023          1.7945
Coefficients:
                                Value           Std. Error      t value
(Intercept)                     8.8244          0.8608          10.2517
male$Friends                    0.0149          0.0025          5.8445
male$Dominant.genderm           12.6625         2.8584          4.4299
Residual standard error: 1.32 on 26 degrees of freedom
```

Appendix 12: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Monday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Monday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than Monday.

Result: W = 12829, p = 0.1029 > 0.05 ⇒ Null hypothesis is **ACCEPTED**.

Appendix 13: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Tuesday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Tuesday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than Tuesday.

Result: W = 14560, $p = 0.1365 > 0.05 \Rightarrow$ Null hypothesis is **ACCEPTED**.

Appendix 14: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Wednesday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Wednesday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than Wednesday.

Result: W = 16560, $p = 0.009602 < 0.05 \Rightarrow$ Null hypothesis is **REJECTED**.

Appendix 15: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Thursday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Thursday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than Thursday.

Result: W = 16143, $p = 0.01446 < 0.05 \Rightarrow$ Null hypothesis is **REJECTED**.

Appendix 16: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Friday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Friday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than

Friday.

Result: W = 11303, $p = 0.1106 > 0.05$ ⇒ Null hypothesis is **ACCEPTED**.

Appendix 17: Hypothesis test on average popularity of a photo uploaded on Sunday with respect to Saturday

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded on Sunday and Saturday is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded on Sunday is greater than Saturday.

Result: W = 9806, $p = 0.005471 < 0.05$ ⇒ Null hypothesis is **REJECTED**.

Appendix 18: Hypothesis test on average popularity of a photo uploaded during morning with respect to afternoon

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded during morning and afternoon is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded during morning is greater than afternoon.

Result: W = 29672, $p =0.06552 = 0.05$ ⇒ Null hypothesis is **REJECTED**.

Appendix 19: Hypothesis test on average popularity of a photo uploaded during morning with respect to evening

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded during morning and evening is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded during morning is greater than evening.

Result: W = 40676, $p =0.004793 < 0.05$ ⇒ Null hypothesis is **REJECTED**.

Appendix 20: Hypothesis test on average popularity of a photo uploaded during morning with respect to night

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded during morning and night is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded during morning is greater than night.

Result: W = 19135, p =0.117 > 0.05 ⇒ Null hypothesis is **ACCEPTED**.

Appendix 21: Hypothesis test on average popularity of a photo uploaded with respect to number of friends

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded by user with more friends and user with less friends is equal.

Alternate Hypothesis (H_a): Average popularity of a photo uploaded by user with more friends is greater than user with less friends.

Result: W = 130922, p =2.2e-16 < 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 22: Hypothesis test on average popularity of a photo uploaded with respect to dominant gender

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded by user with male as dominant gender and with female as dominant gender is equal

Alternate Hypothesis (H_a): Average popularity of a photo uploaded by user with male as dominant gender is greater than with female as dominant gender

Result: W = 9484, p =0.05323 = 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 23: Hypothesis test on average popularity of a photo uploaded by male user with respect to dominant gender

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded by male user with female as dominant gender and with male as dominant gender is equal

Alternate Hypothesis (H_a): Average popularity of a photo uploaded by male user with female as dominant gender is less than with male as dominant gender

Result: W = 698.5, p =0.0005736 < 0.05 ⇒ Null hypothesis is **REJECTED**.

Appendix 24: Hypothesis test on average popularity of a photo uploaded by female user with respect to dominant gender

Test: Wilcoxon rank sum test

Null Hypothesis (H_o): Average popularity of a photo uploaded by female user with female as dominant gender and with male as dominant gender is equal

Alternate Hypothesis (H_a): Average popularity of a photo uploaded by female user with female as dominant gender is greater than with male as dominant gender

Result: W = 1584, *p* =0.5453 > 0.05 ⇒ Null hypothesis is **ACCEPTED**.

Appendix 25: Summary statistics of popularity predictive model using robust linear model.

```
rlm(formula=photo_data$Popularity.value~photo_data$Friends+photo_data$Gender+
photo_data$Dominant.gender+photo_data$Uploading.day+photo_data$Uploading.time,
weights= (1/(rlm_model$residuals^2)))
```

```
Residuals:
Min          1Q           Median       3Q           Max
-1.2666      -0.9925      -0.9398      1.0018       1.7809
Coefficients:
                                      Value      Std. Error   t value
(Intercept)                           9.8046     0.1148       85.3775
photo_data$Friends                    0.0130     0.0001       92.5166
photo_data$Gendermale                 -0.2475    0.1687        -1.4674
photo_data$Dominant.gendermale        1.5506     0.1704         9.1010
photo_data$Uploading.dayMonday        -1.6994    0.0622       -27.3077
photo_data$Uploading.daySaturday      -2.7144    0.1095       -24.7802
photo_data$Uploading.daySunday         0.3534    0.0777         4.5517
photo_data$Uploading.dayThursday      -2.4311    0.1617       -15.0299
photo_data$Uploading.dayTuesday       -0.6717    0.0822        -8.1677
photo_data$Uploading.dayWednesday     -2.7527    0.0796       -34.5625
photo_data$Uploading.timeevening       -2.8783   0.0640       -44.9578
photo_data$Uploading.timemorning       -0.4456   0.1421        -3.1358
photo_data$Uploading.timenight         -0.5378   0.0697        -7.7147

Residual standard error: 1.479 on 1025 degrees of freedom
```