Fenil S. Tailor (N18730085)
Deepti Vaddireddy (N15361877)

Final Project Report

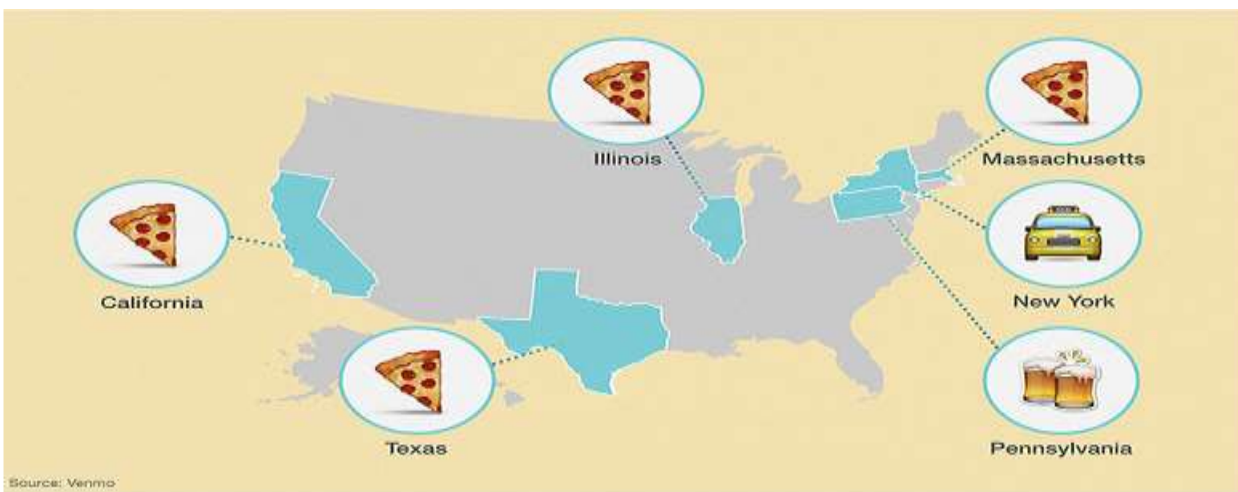
1) What did you propose to do? What is the motivation/background?

A: If information from all the transactions made on Venmo is added up, patterns emerge showing who we interact with, when we interact with them, and most importantly the context of interaction, which infers a lot about one's life.

Using the Venmo API to access the public data for people's transaction, our project primary goals were:

- To analyze the items or events for which there are significant monetary transactions between individuals or a group.
- To assess payment types and any relationship to purchasing patterns.
- To determine the social network sizes
- To try and determine the relationship of the people between whom the transactions are taking place.

The Motivation/Background:



What millennials spend more on is shown by the data Venmo pulled for CNN Money- the world's largest business website- (http://money.cnn.com/2015/10/31/news/economy/venmo-millennials-spend-big-on-this/)

According to this data, what certain demographics spend on is frequently represented by the above picture. They are:

Taxi/Uber in New York

Pizza/Food in California, Texas, Illinois, Massachusetts

Beer in Pennsylvania

2) Explain the data you used and model in detail.?

A:We have used data from the Venmo API which is publicly available and the link where we get the data from is https://venmo.com/api/v5/public.This link has the Live feed for the data(Recent transaction is uploaded onto this link)

1) Each link file is a json document which contains exactly 20 transactions which we can find under "data" document.
2) At the top of the each json document we have two links named "next" and "previous" from which we can go back and forth to the Venmo data
3) Each transaction can be mainly divided into four sub json documents
   1)Payment information in the first sub document
   2)Actor(source) information in the actor subdocument
   3)Transaction + Target information in the transactions subdocument
   4)Likes of the transaction in the likes sub document

Data has the following features for every transaction:

As we can see, most of the variables are of the type String. The only numerical data type is of the count(likes count)variable. We are not dealing with any ordinal type of variables here. All the variables originating with the name 'actor' contain the user's information(source) and the variables with the name 'target' contain information of the person the transaction is occurring with. Other variables dealing with the Date Time or Date IST format are not of much importance but just serve as a reference to see when the transaction took place. Out of all these, the actor.username and the target.username are the most important in further steps of network visualizations.

Variable	Type	Variable	Type
Target.username	String	Message	String
Payment_id	String	Type	String
Target.is_business	Boolean	Count	Integer
Target.Name	String	actor.id	String
Target.Firstname	String	actor.username	String
Target.Lastname	String	actor.firstname	String
Target.cancelled	Boolean	actor.lastname	String
Target.datecreated	Date IST format	actor.external_id	String
Target.external_id	String	Created_time	DateTime
Target.target_id	String	transaction.date_created	DateTime

We have collected the full data of every transactions into dataframe named "Total". Below is the screenshot for one transaction in our dataframe.

	Updated_ID	username	picture	is_business	name	firstname	lastname	ca
1	1	Eric-Wilhelmy-1	https://graph.facebook.com/v2.0/1282614578420...	FALSE	Eric Wilhelmy	Eric	Wilhelmy	FAL
2	2	Mark-Mutton	https://venmopics.appspot.com/u/v1/m/2c4cc5e0...	FALSE	Mark Mutton	Mark	Mutton	FAL
3	3	Alyssa-Stone-3	https://venmopics.appspot.com/u/v4/n/6db05866-...	FALSE	Alyssa Stone	Alyssa	Stone	FAL
4	4	sdjohnson25	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Scott Johnson	Scott	Johnson	FAL
5	5	hollyfullmer13	https://venmopics.appspot.com/u/v2/n/dd2da623-...	FALSE	Holly Fullmer	Holly	Fullmer	FAL
6	6	Tal-Lederman	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Tal Lederman	Tal	Lederman	FAL
7	7	Mackenzie-Barmen	https://venmopics.appspot.com/u/v1/m/8844f556-...	FALSE	Mackenzie Barmen	Mackenzie	Barmen	FAL
8	8	Hiram-Alvarez	https://venmopics.appspot.com/u/v4/m/922205d2...	FALSE	Hiram Alvarez	Hiram	Alvarez	FAL
9	9	Jennifer-Bennett-5	https://venmopics.appspot.com/u/v1/m/095b6780...	FALSE	Jennifer Bennett	Jennifer	Bennett	FAL
10	10	Cameron-Penta	https://venmopics.appspot.com/u/v2/m/26a71264...	FALSE	Cameron Penta	Cameron	Penta	FAL
11	11	Hannah-Maclellan	https://venmopics.appspot.com/u/v1/n/4ca2c6ef-...	FALSE	Hannah Maclellan	Hannah	Maclellan	FAL
12	12	noelravitz	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Noel Ravitz	Noel	Ravitz	FAL
13	13	Alex-Leicht	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Alex Leicht	Alex	Leicht	FAL
14	14	Daisy-Kang	https://venmopics.appspot.com/u/v1/m/888b5d9d...	FALSE	Daisy Kang	Daisy	Kang	FAL
15	15	Keelin-OConnell	https://venmopics.appspot.com/u/v3/n/0dbd8321-...	FALSE	Keelin O'Connell	Keelin	O'Connell	FAL
16	16	Hannah-Hughes-1	https://venmopics.appspot.com/u/v3/n/9de37521-...	FALSE	Hannah Hughes	Hannah	Hughes	FAL
17	17	Ryan-Kendro	https://venmopics.appspot.com/u/v1/m/b638abc0...	FALSE	Ryan Kendro	Ryan	Kendro	FAL
18	18	Sean-Maguire41	https://venmopics.appspot.com/u/v1/m/f6d24a6b-...	FALSE	Sean Maguire	Sean	Maguire	FAL
19	19	Yining-Lu	https://venmopics.appspot.com/u/v1/n/08f24f58-...	FALSE	Yining Lu	Yining	Lu	FAL
20	20	David-Sollenberger	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	David Sollenberger	David	Sollenberger	FAL
21	1	Corey-Raney	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Corey Raney	Corey	Raney	FAL
22	2	Jared-Nevens	https://graph.facebook.com/v2.0/8845224/picture...	FALSE	Jared Nevens	Jared	Nevens	FAL
23	3	Ruhan	https://graph.facebook.com/v2.0/1000023808450...	FALSE	Ruhan Wang	Ruhan	Wang	FAL
24	4	JonCarlo-Dominguez	https://venmopics.appspot.com/u/v1/n/b00af806-...	FALSE	Jon Carlo Dominguez	Jon Carlo	Dominguez	FAL
25	5	raychillo	https://venmopics.appspot.com/u/v1/m/2f7379d0-...	FALSE	Rachel Ortega	Rachel	Ortega	FAL
26	6	Ava-Cunningham	https://s3.amazonaws.com/venmo/no-image.gif	FALSE	Ava Cunningham	Ava	Cunningham	FAL
27	7	Collin-Kluchman	https://venmopics.appspot.com/u/v2/n/f4d857fd-a...	FALSE	Collin Kluchman	Collin	Kluchman	FAL
28	8	Jacqueline-Williams	https://venmopics.appspot.com/u/v1/m/6216a4be...	FALSE	Jacqueline Williams	Jacqueline	Williams	FAL
29	9	itsjaneinthemembrane	https://venmopics.appspot.com/u/v1/m/07dcdecc...	FALSE	Jane Henderson	Jane	Henderson	FAL

We have attached the csv file which contains all the data which we got by running our code:

https://drive.google.com/file/d/0ByCWO2MHeUtEdWU3VXNCb1FJOXc/view?usp=sharing

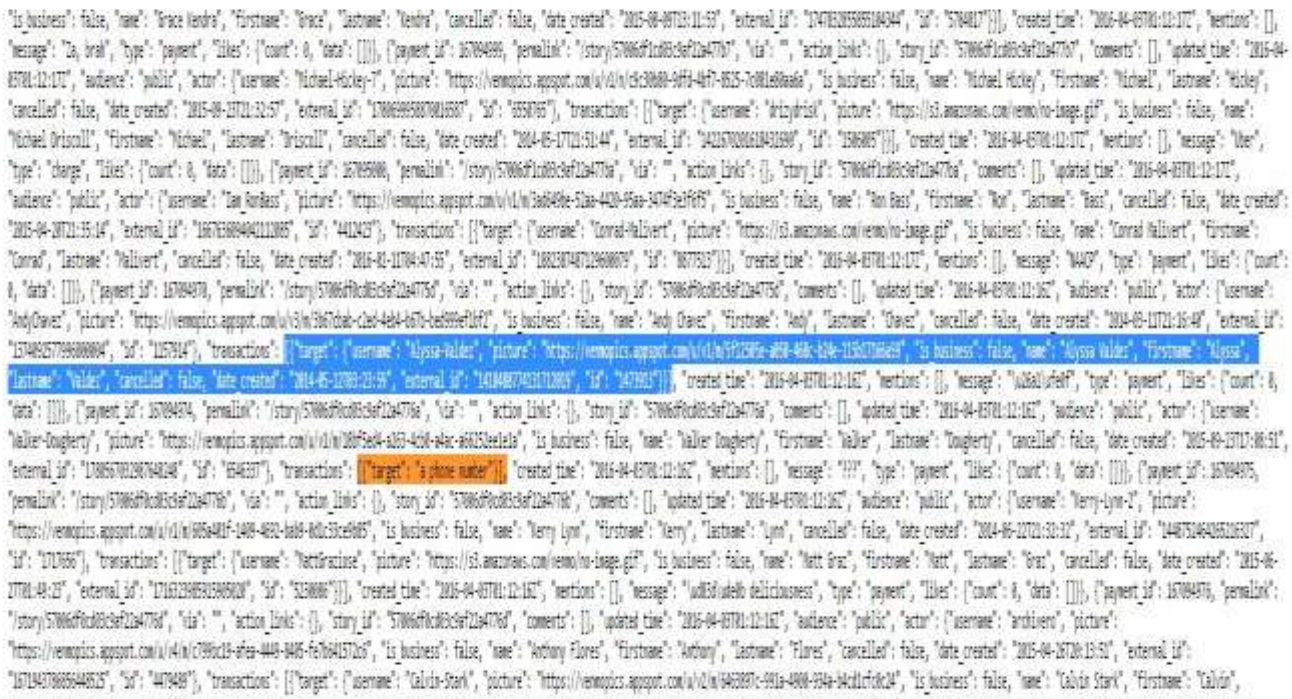
Model

The initial steps that we followed are as follows:

1)**Data gathering from the Venmo API**-We have gathered data from the json document from the Venmo API into a well-structured dataframe which contains all the information about all transactions.

2)**Data cleaning-** We have done the data cleaning part during our implementation only. As there are many transactions where the transactions subdocument in the "data" json Document has just only one attribute which is as follows

"transactions":[{target:"a phone number"}]



As we are not sure if it is a fraud transaction or the further data is not available because of some technical problems we are not taking that transactions as a suitable record in our further analysis

3)**Data Transformation-**As we have combined three data frames (subdocument) to get the full data for a particular transaction, we have created a common column “ID” based on which we are joining the data frame

Data frame that we are combining is:

1)final_df<-contains the payment info data for particular transaction

2)transactiondf<-contains the transaction info data for the particular transaction

3)likes<-contains the likes and comment info data for the particular transaction

After merging these 3 dataframes, we are merging the full data into our master dataframe named “total”

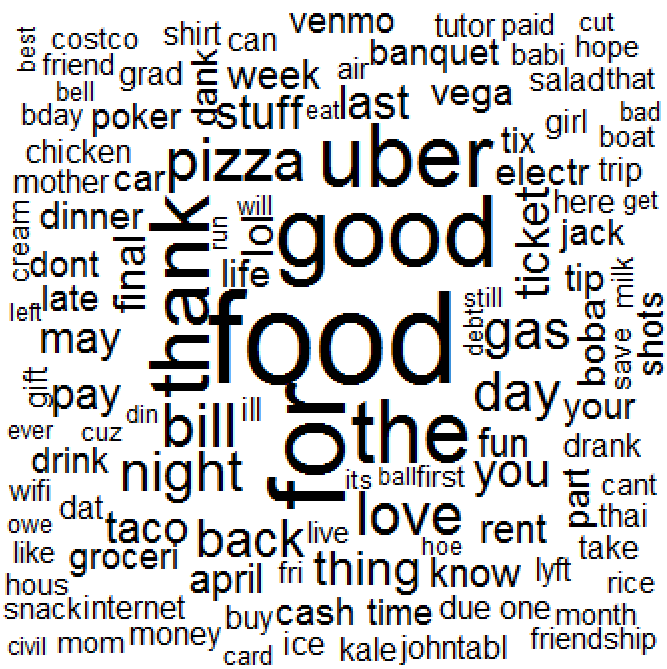
As such for data analysis of the Venmo Api, we haven’t use any specific model. But to visualize our output we have generated a Wordcloud which internally uses K-means clustering algorithm (To group the likelihood word into one cluster) and it also uses the correlation between frequency of a word and its corresponding count.

3) What did you end up doing?

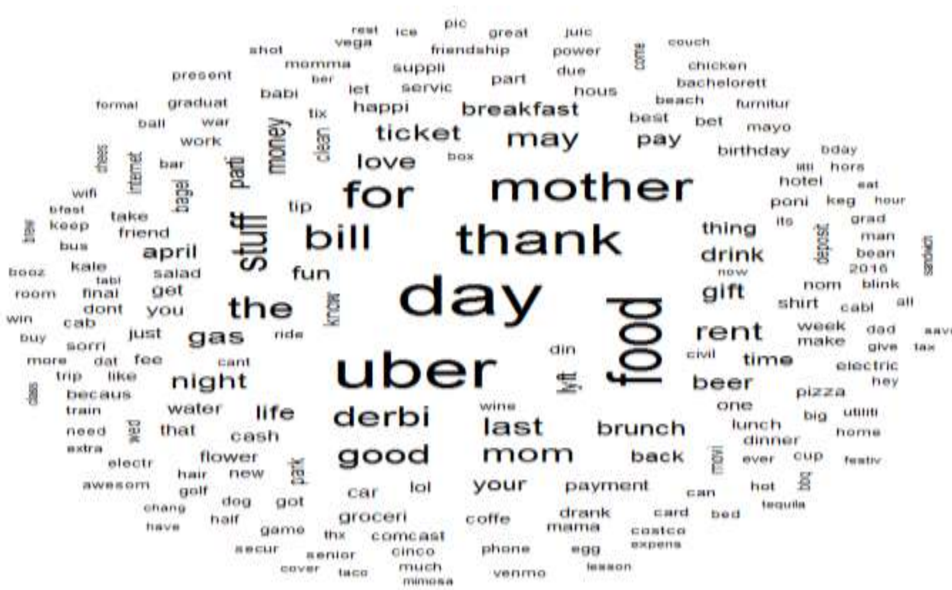
A: Generating a WordCloud- We are converting the dataframe columns from class “factor” to class “character” so that we can compute some String operation for our analysis**.** After converting the dataframe to a class character, we are encoding **message** column of the data frame **total1** to ASCII just to get the english readable characters (ignoring all the emojis and random alphanumeric characters that do not make sense) to form a proper word cloud.

We are generating word cloud using the frequency count of each word and by using three libraries tm, SnowballC, wordcloud

The Wordcloud that we usually get and which corresponds to CNN's findings:



Wordcloud that we got around the time period of May6, May 7 and May 8: (We assume that this is due to the Mother's day event on May 8th as words like Mother, day, mom, etc are also popular rather than the conventional findings of food,rent,uber,etc).



Word frequency in a dataframe that we got:-

	frequency
day	260
uber	232
food	214
thank	184
mother	178
for	163
bill	143
stuff	128
the	127
derbi	113
good	104
mom	96
last	95
rent	93
may	91
gas	77
ticket	77
night	76
love	74
gift	68
brunch	66
beer	65

B: Got the frequency of each word in our cloud-We have created **"documentterm_matrix"** to get the frequency of each word appearing in our dataset after cleaning and after removing English language stop words, punctuation marks.

For that we are summing up the count column where we find same word in our dataframe

C: Generating a network for some particular user-We have generated a graph taking source node as the username which end user has put in the response of readusername() function and target as the username with which the source has done transactions.We are plotting that by using simpleNetwork() function

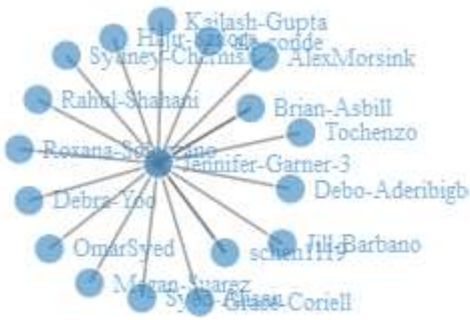
At the time of analysis I have put username="Jiapei-Chen" in the response of readusername() function. Below is her transaction summary with the target username with whom she had transactions

	rn	source	target	message
1	1	Jiapei-Chen	Lantian-Chen	for att april
2	10	Jiapei-Chen	Judy-Chang-1	for Ã¼ber to piphi
3	11	Jiapei-Chen	Judy-Chang-1	for Ã¼ber back to 91 sidney
4	2	Jiapei-Chen	Yijun-Jiang	for att march
5	3	Jiapei-Chen	Yibo-Gao	for att april
6	4	Jiapei-Chen	Yibo-Gao	for att march
7	5	Jiapei-Chen	Jiayi-Liu	for att april
8	6	Jiapei-Chen	Jiayi-Liu	for att march
9	7	Jiapei-Chen	Sonia-Zhang	for Ã¼ber to senior ball
10	8	Jiapei-Chen	becha2	for Ã¼ber to piphi
11	9	Jiapei-Chen	becha2	for Ã¼ber to senior ball

The Network that we got is as follows:

As we can see from the data frame that **Jiapei-Chen** has two transactions with **Yibo-Gao, Jiayi-Liu, becha2** and **Judy-chang1,** the edge connecting with the source as **Jiapei-Chen and** target as **Yibo-Gao,Jiayi-Liu,becha2** and **Judy-chang1** is slighter darker than all other transaction edges

Another example for network visualization for Jeniffer-Garner-3



4) What if anything did you change about your approach and why?

A: Yes, we have changed the way we want to make the network. Initially we wanted to visualize the network by taking all the user data of Venmo that we currently had. But after looking at the data transactions we realized that for visualizing the network for all users, we need some powerful computers. So after that we decided to visualize network only for the user for which end user has put the username in response to the readusername() function.

Also, at first we were considering the transactions with only one attribute (as the phone number, instead of a contact in the user's friend list) as a Fraud transaction. But after getting the response from the Venmo team, we decided to remove that transaction from our analysis.

Venmo Team Response: **"We cannot be sure that this transaction is a fraud transaction just by looking at the json doc. We need some other analysis also to conclude that as a fraud transaction. Also, it might be the case that because of some technical difficulties the data is not uploaded on the server"**

5) What visualization(s) have you included? Explain what is conveyed in the visualization and why?

A: For our visualizations we have used:

1)WordCloud to see what Millennials are frequently and mostly spending on. This constantly changes according to the trends or any important events. For example, the wordcloud displayed above is the most recent one that has day, mother and mom apart from the usual uber, food, rent, etc as the predominant ones. We infer that this is because of Mother's day event celebration that took place on May 8th.

2)SimpleNetwork() for the graph (mapping the source username, target username). Since the size of the graph and mapping all the transactions done on Venmo was a huge concern for us, we decided to use a user query that inputs only a particular user's name and maps/shows the visualizations for them to see who they have transactions with and use this this information for future analysis regarding social network sizes, relationship between the user and target, purchasing patterns, etc.

6) What evaluation method did you propose?

A: Firstly, we can evaluate our work by looking at the results of the WordCloud itself which correlates with the CNN findings.

Secondly, we could not plot all the Venmo users in our data and show their transactions. Instead, we wrote a user query that takes into account the particular user name mentioned and plots only their transactions to answer questions about social network sizes, relationship between the user and target, purchasing patterns, etc.

7) How did your model perform according to this evaluation?

A: From the data we have analyzed, we did find that food, uber, rent, etc are some of the things on what people usually spend on the most.

Also, by inputting the user's name, we can see the list of their transactions and their message (i.e, for what they are paying/receiving the money).The more the number of transactions the user has with somebody (the target), the darker are the edges connecting the nodes. This answers who the user is regularly involved with and his/her social network size (the number of people he has transactions with) and on what is the user spending on (based on the messages) and also the relationship between them and the target could be analyzed by a detailed transaction history between the user and a target.

Lastly, we can cross check these transactions on www.Venmo.com to see if our analysis for the particular user is right or not.

8) Based on your results what conclusions do you draw?

A: The main conclusions to be drawn are:

- From the data we have analyzed, we saw that food, uber, rent , etc are some of the things on what people usually spend on the most.
- Secondly, based on the transaction analysis of the user "**Jiapei-Chen"**, we can see that she has most of her transactions with the following 7 users on Venmo: Lantian-Chen, Yijun-Jiang, Yibo-Gao, Jiayi-Liu,Sonia-Zhang, becha2 and Judy-Chang.
- We can safely conclude that we can talk about the Jiapei-Chen's social network and its size based on the above information.
- When we got to www.Venmo.com, and search for the user ""**Jiapei-Chen"**, we can also see that most of her transactions with the 7 girls mentioned above are for :

Jiapei Chen charged Lantian Chen
for att march
Like · Saturday at 12:41 PM

:D Leave a comment...

Jiapei Chen charged Lantian Chen
for att april
Like · Saturday at 12:41 PM

:D Leave a comment...

Jiapei Chen charged Yibo Gao
for att april
Like · Saturday at 12:20 PM

:D Leave a comment...

Jiapei Chen charged Yibo Gao
for att march
Like · Saturday at 12:20 PM

:D Leave a comment...

Jiapei Chen charged Yijun Jiang
for att march
Like · Saturday at 12:27 PM

:D Leave a comment...

Jiapei Chen charged Yijun Jiang
for att april
Like · Saturday at 12:27 PM

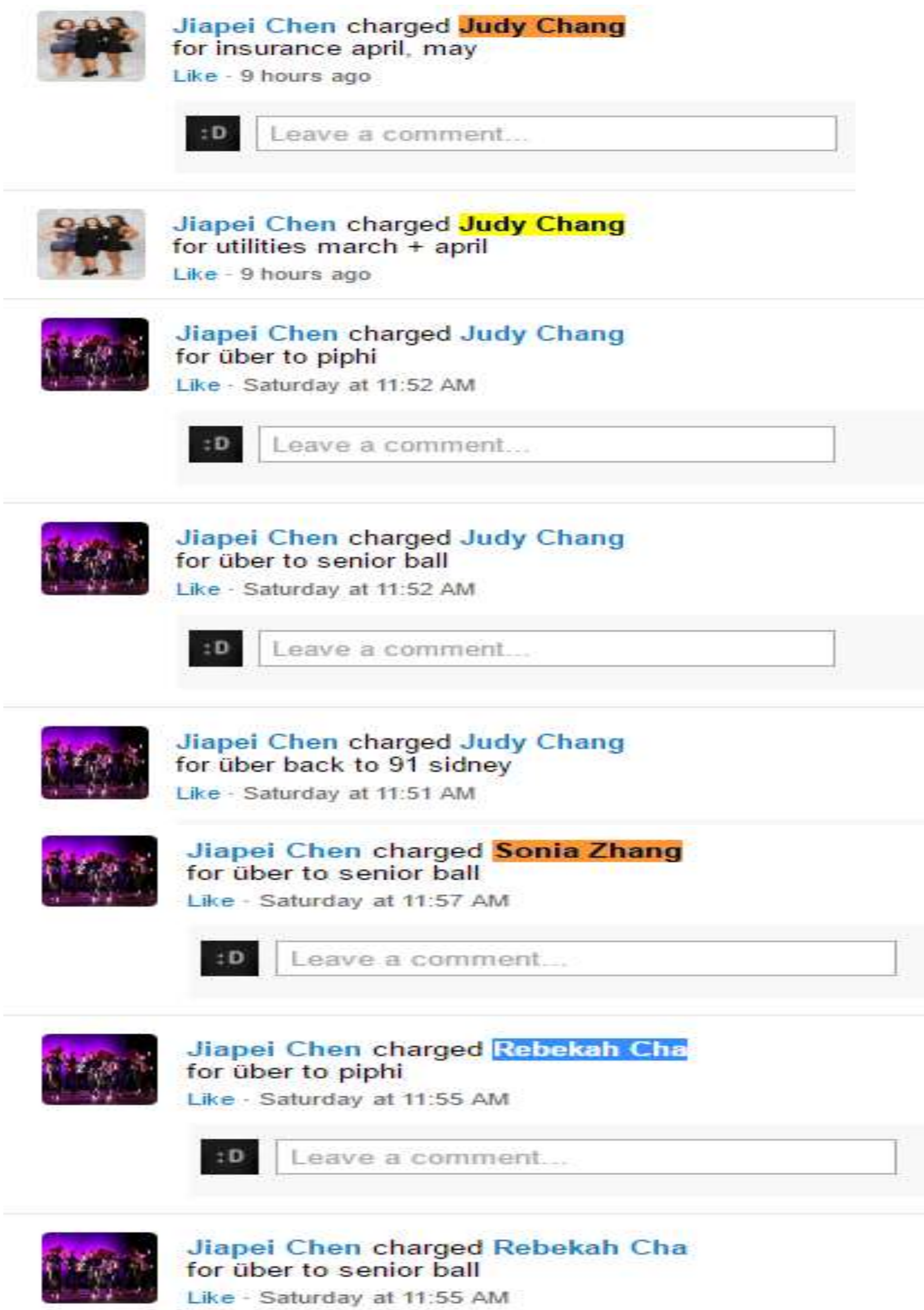
:D Leave a comment...

Jiapei Chen charged Jiayi Liu
for att april
Like · Saturday at 12:03 PM

Jiapei Chen charged Rebekah Cha
for insurance april, may
Like · Saturday at 6:22 PM

:D Leave a comment...

Jiapei Chen charged Rebekah Cha
for utilities march + april
Like · Saturday at 6:22 PM



- We can see that Jiapei Chen has transactions with: Lantian-Chen, Yijun-Jiang, Yibo-Gao and Jiayi-Liu, regarding the bill att March and April. This could mean that all these 5 girls may be under the same family plan in AT&T and Jiapei has paid the total bill for these 2 months and split the expenses.
- Also, Jiapei has transactions with Rebekah Cha and Judy Chang for insurance in April and May and also for utilities in April and May.

- Lastly, we can see that Jaipei shared an Uber ride to her senior ball with Rebekah Cha, Sonia Zhang and Judy Chang. We can also see that she had shared another Uber ride to “piphi” with Rebekah and Judy.
- Based on all the above points, we could probably guess that these 3 girls (Rebekah, Jaipei and Judy) are roommates or they might at least live in the same building/neighborhood.

9) Based on your results what further studies would you do or are warranted?

A: Based on what we have done so far; we would like to further do the following if possible:

- Most people sign up from Facebook to use Venmo. So, we would have liked to inculcate the facebook api into our project so that we could get the images of all our users making the transactions and their targets so that our visualizations could have been better.
- Also, to do a more thorough analysis of purchasing patterns according to demographics, like CNN, getting any geographical details of the transactions (i.e, from where they originated/being sent to) would have definitely helped.
- To see the time trends of the most popular/frequent things that people use Venmo for, we would have liked to do a Time series Analysis (2016 vs 2015 vs 2014 and so on). But, as there are around a billion of Venmo transactions taking place daily, getting the data to do a thorough time analysis (say atleast for 2-3 years) seemed very tough.