Wharton 2018 GroupMe

MTKG776: Applied Probability Models in Marketing

2017-02-23

Contents

1 Executive Summary **1**

2 Data **1**

2.1 GroupMe Platform 1

2.2 Wharton 2018 GroupMe 3

2.3 Count Datasets 4

3 NBD Model **5**

3.1 Posts 5

3.2 Mentions 7

3.3 Likes 10

3.4 Compare Parameter Estimates 12

3.5 Lorenz Curve 12

3.6 Gender 12

4 Limitations **12**

1 Executive Summary

There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class.There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class. There are approximately 850 members of the Wharton 2018 MBA class.

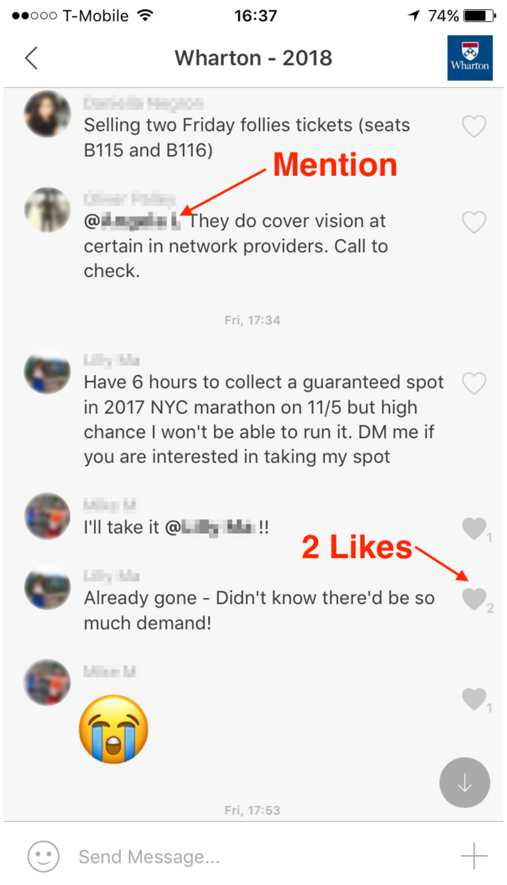
2 Data

2.1 GroupMe Platform

Figure 1: GroupMe Logo

You have just decided to attend Wharton business school. After paying your deposit and joining the Facebook group, the next thing you do is join the class GroupMe. GroupMe is messaging service conceived and built in 2010 and was later acquired by Skype (and thus a Microsoft holding). Unlike Whatsapp or iMessage, GroupMe is designed for *group messaging* rather than one-on-one conversations. As such it's become the message platform *du jour* for university students as it supports groups with hundreds of users. Below is a sample message within a thread. You can see 3 primary actions:

1. Posts - messages sent by users
2. Mentions - "@"ing another user, which alerts them
3. Likes - heart-ing a post to show you like it



[redacted]
Selling two Friday follies tickets (seats B115 and B116)

[redacted]
@[redacted] They do cover vision at certain in network providers. Call to check.

Fri, 17:34

[redacted]
Have 6 hours to collect a guaranteed spot in 2017 NYC marathon on 11/5 but high chance I won't be able to run it. DM me if you are interested in taking my spot

[redacted]
I'll take it @[redacted] !!

[redacted]
Already gone - Didn't know there'd be so much demand!

[redacted]
😭

Fri, 17:53

Figure 2: Screenshot showing posts, mentions, and likes

The data in this analysis is from the "Wharton - 2018" GroupMe group (often just referred to as the Wharton 2018 GroupMe). GroupMe has an API that allows developers to access groups and messages. After creating an access token, we built a pipeline to acquire and process the users and messages from GroupMe for this group (see this data processing documentation for details). After parsing the json's and cleaning the data, we created a dataset of tables illustrated in the diagram below:

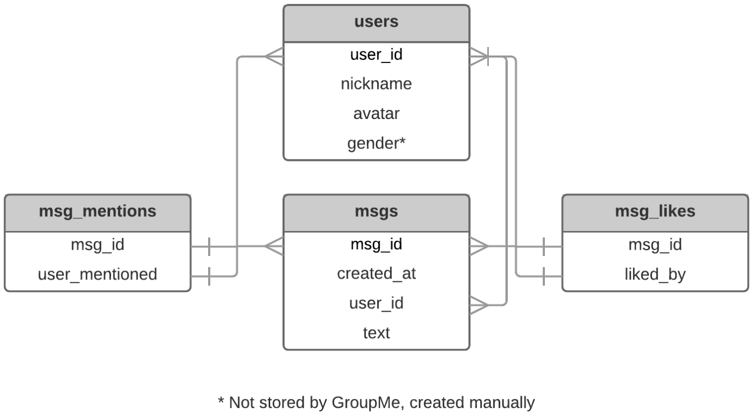
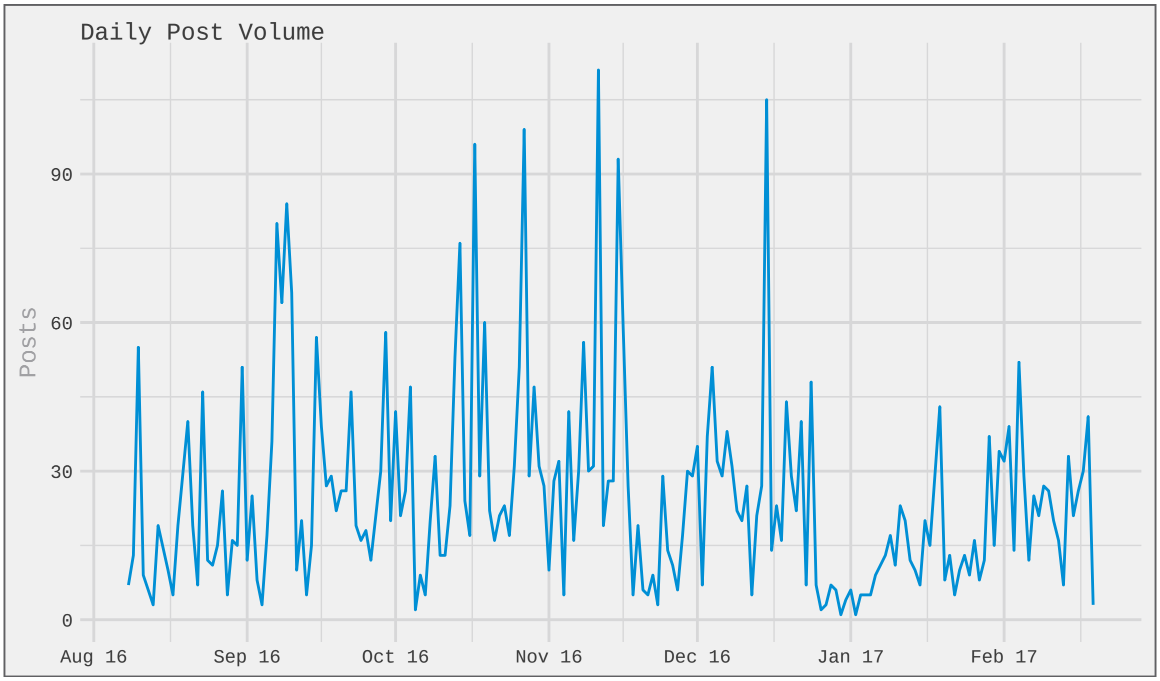


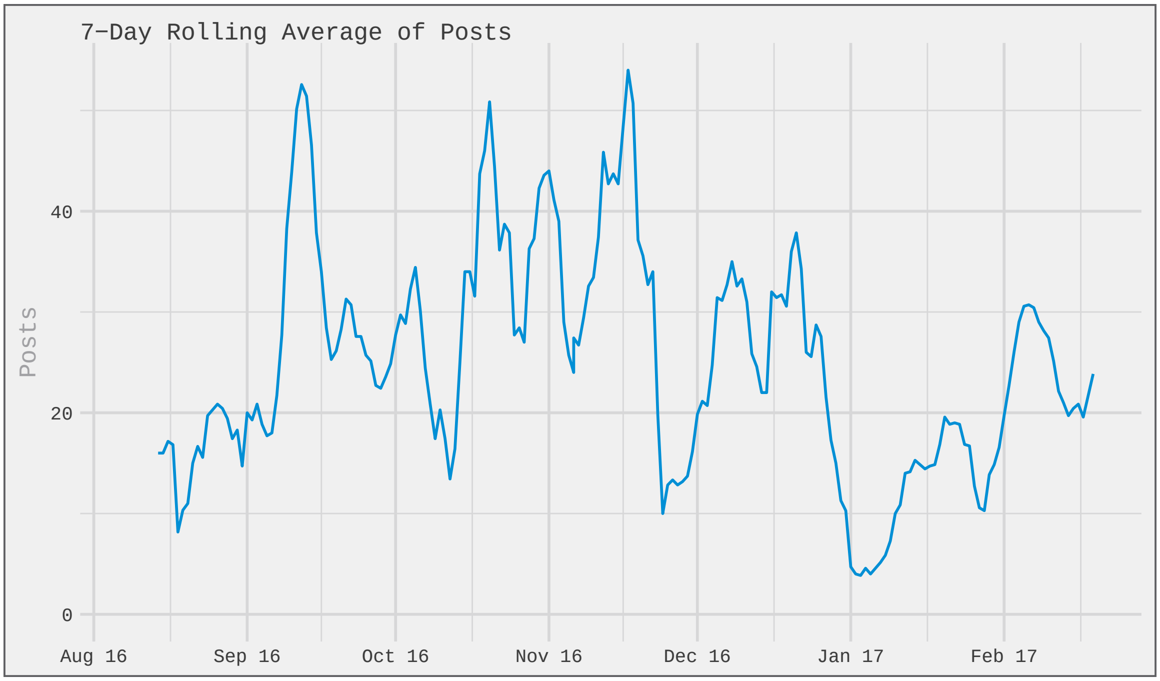
Figure 3: GroupMe data organization

2.2 Wharton 2018 GroupMe

There are **812** users in the Wharton 2018 GroupMe. There are approximately 850 members of the Wharton 2018 MBA class. Though the group was created in January 2016, we trimmed the dataset to August 8, 2016 (first day of pre-term) to provide an accurate window in which to observe the actions of the users. In other words, all users have the same observation period. We removed users from the dataset that have left the group and discuss the possibility of late joiners in the final Limitations section. The last post in our dataset is **2017-02-19 02:11:26**. There have been **4,847** posts by **563** distinct users. Below is a time series of the posts:



From the plot above we see a great deal of daily volatility. Below is a plot of a 7-day rolling average that helps smooth out spikes and exhibit the trend.



2.3 Count Datasets

2.3.1 Three Events

The three actions that we will investigate (posts, mentions, and likes) each arise from count processes and thus deserve a count model (i.e. NBD).

Table 1: Count datasets arising from the Wharton 2018 GroupMe

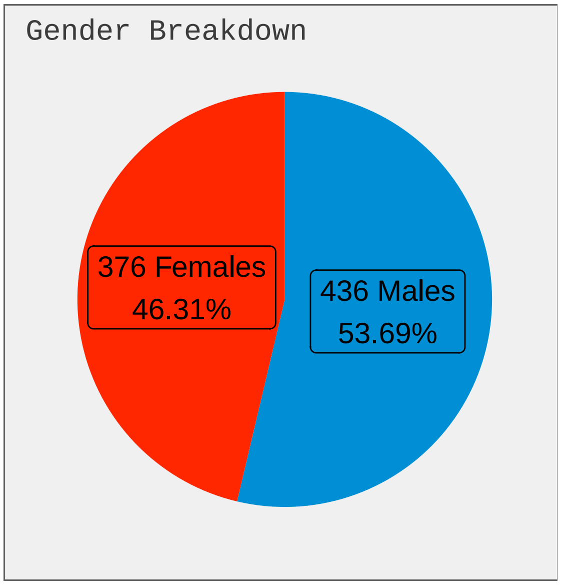
Event	Individual-level Story	Source of Heterogeneity
Posts	Users in the Wharton 2018 GroupMe can post as many times as they would like - there is no upper bound. Thus we can think of each user as having a **post rate**, λ, in the observed time window.	Users interact with GroupMe differently. Some post a lot, some have never posted. However, all users have the same opportunity to post.
Mentions	Users in the Wharton 2018 GroupMe can be mentioned an infinite number of times - there is no upper bound. Other users can create a new post and mention them. Unlike the posts event, the act of being mentioned is not in the agency of individual. Nevertheless, we can think of each user has having a **mention rate**, λ, during the observed time window that determines how many times they will be mentioned	Popularity. In all seriousness, some users of the group will be mentioned more than others. Some will not be mentioned at all. Heterogeneity arises from the social construct.
Likes	The number of posts a user has liked is a choice dataset, as there is a finite number of opportunities to like a post (i.e. the number of posts). However, given the high upper bound, we can reasonably view this dataset as a count process. As such, each user has some **like rate**, λ, during the observed time window that determines how many posts they like. An individual can be someone that likes every post or has never liked a post.	Users have different levels of engagement on the Wharton 2018 GroupMe. Thus, it follows there will be variation in like rates within the user population.

Event	Individual-level Story	Source of Heterogeneity

We expect to observe difference in heterogeneity of each of the three events. For example, we would presume that there is more heterogeneity in **like rate** than in **post rate** as liking is less visible and risky (to one's reputation) than posting in a group of 812.

2.3.2 Gender

In addition to three behaviors that are the primary interest of this analysis, we included an attribute of the user: gender. We will use this to identify if there are differences in posting, being mentioned, or liking between male and female Wharton students.



3 NBD Model

3.1 Posts

In the plot below we show the distribution of posts per user. The distribution is positively skewed with a long right tail. There are a few users that have posted more than 50 times, but the majority are less active. The median number of post per user is **2** posts though the mean posts per user is **5.97** posts.

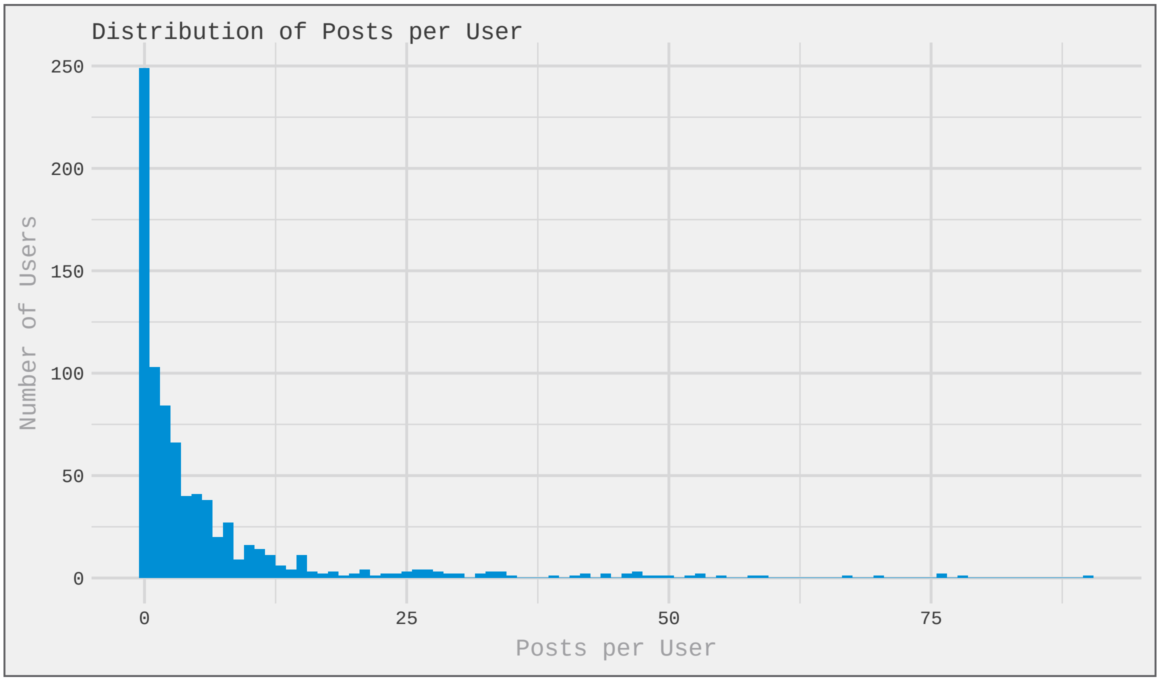


Table 2: Number of users for count of posts in period (bottom 10)

posts	users
0	249
1	103
2	84
3	66
4	40
5	41
6	38
7	20
8	27
9	9

We fit an NBD model, including a zero-inflated NBD given the noteable spike at 0, using MLE, method of moments, and means and zero to estimate parameters. We find through MLE that a zero-inflated model does not help describe the data as $\pi = 0$.

Table 3: NBD parameters estimates for different methods

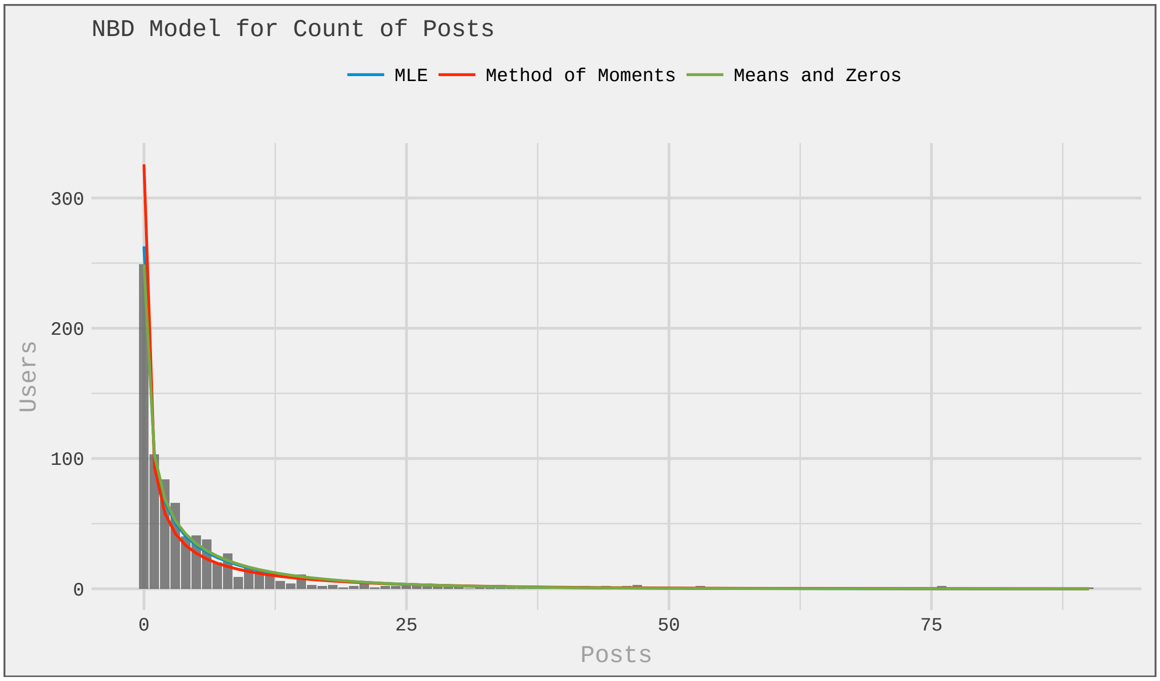
model	r	alpha	pi
MLE	0.4113	0.0689	
MLE (Zero-Inflated)	0.4113	0.0689	0
Method of Moments	0.3006	0.0504	
Means and Zeros	0.4420	0.0740	

We note the divergence between the method of moments and MLE/means and zeros parameter estimates. The large standard deviation, **11.16**, shrinks the esimate of alpha as $\hat{\alpha} = \frac{\bar{x}}{s^2 - \bar{x}}$.

Below is a table that shows the estimated number of users for posts counts less than five by the three parameter estimation techniques. A plot showing all post counts follows. We see that the methods are not that different, but method of moments certainly performs the worst.

Table 4: Estimated number of users for posts (<= 5) by different estimation methods

posts	Actual	MLE	Method of Moments	Means and Zeros
0	249	263	326	249
1	103	101	93	102
2	84	67	58	69
3	66	50	42	52
4	40	40	33	42
5	41	33	27	35



In order to perform the $\chi\hat{}2$ goodness of fit test for the NBD model, we need rollup the tail so that 80% of the expected counts have more than 5 counts. We create a 25+ bucket and calculate the χ^2 test statistic and p-value for each paremeter estimation method using 25 - 2 - 1 = 22 degrees of freedom. Based on the p-values shown below, we have no evidence that the data came from the NBD model. Nevertheless, the plot above shows a relatively good fit, at least for the estimates from MLE and means and zeros.

Table 5: Goodness of Fit Test

model	chisq	p.value
MLE	52.48	0.000268
Method of Moments	94.40	0.000000
Means and Zeros	50.70	0.000471

3.2 Mentions

Like posts we start by looking at the distribution of the number of times a user has been mentioned both in graphic form and the the table below. Like posts, mentions are postive skewed with a long right tail - one user has 40 mentions. The median number of mentions for a user is **0** mentions though the mean is **1.74** mentions.

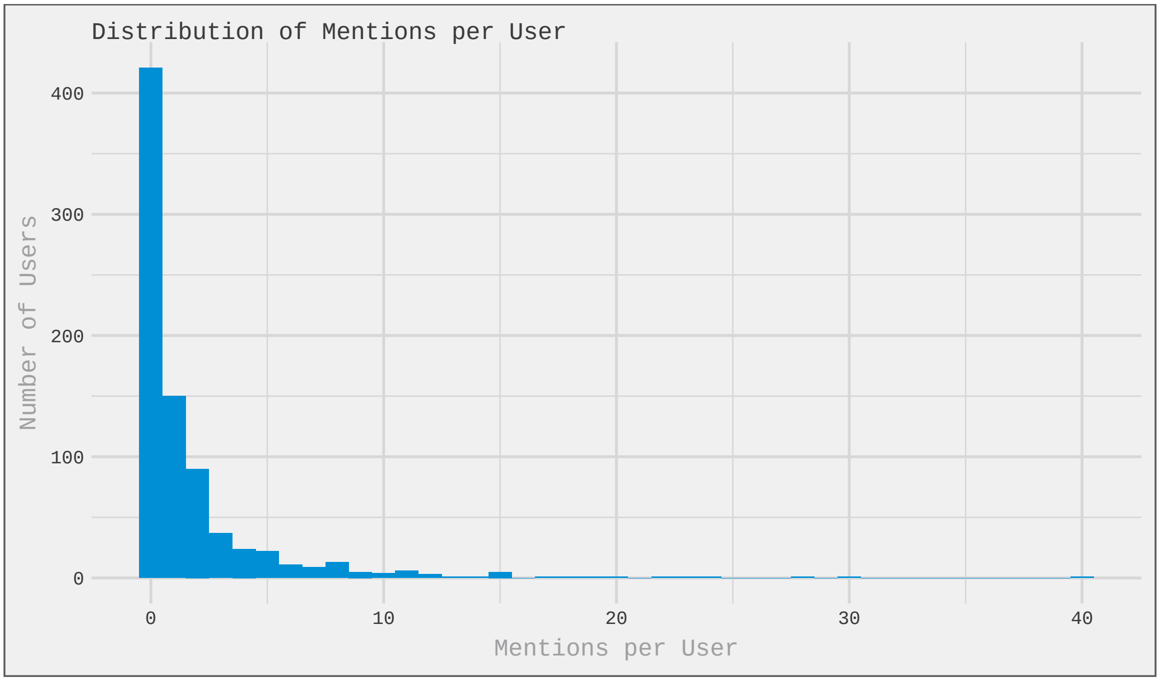


Table 6: Number of users for count of mentions in period (bottom 10)

mentions	users
0	421
1	150
2	90
3	37
4	24
5	22
6	11
7	9
8	13
9	5

We perform the parameter estimation using the same techniques and find that zero-inflated model does not fit the data. Like the method of moments estimates for posts, the method of moments estimates for mentions are quite different from the estimates by MLE and means and zeros.

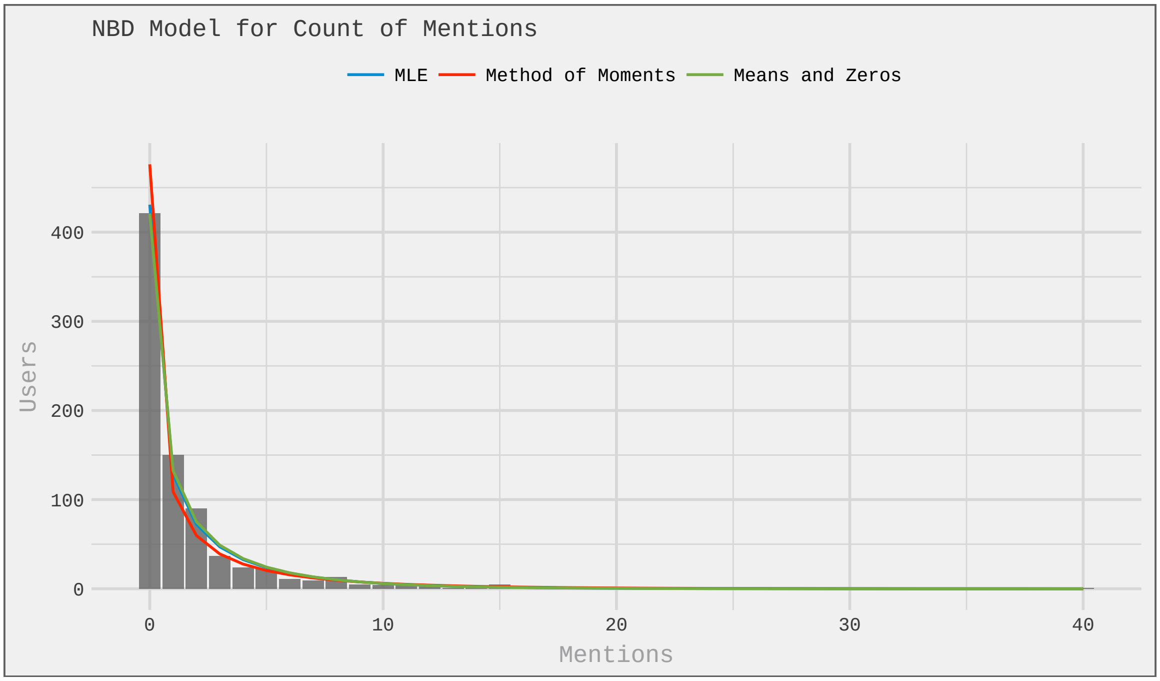
Table 7: NBD parameters estimates for different methods

model	r	alpha	pi
MLE	0.3593	0.2069	
MLE (Zero-Inflated)	0.3593	0.2069	0
Method of Moments	0.2632	0.1516	
Means and Zeros	0.3848	0.2216	

Table 8: Estimated number of users for mentions (<= 5) by different estimation methods

mentions	Actual	MLE	Method of Moments	Means and Zeros
0	421	431	476	421
1	150	128	109	133
2	90	72	60	75
3	37	47	39	49
4	24	33	28	34
5	22	24	21	24
6	11	18	16	18
7	9	13	12	13
8	13	10	10	10

The plot below shows the parameter estimates by MLE and means and zeros fit quite well.



Like before, to perform the χ^2 goodness of fit test for the NBD model, we need rollup the tail so that 80% of the expected counts have more than 5 counts. We create a 10+ bucket and calculate the χ^2 test statistic and p-value for each paremeter estimation method using 10 - 2 - 1 = 7 degrees of freedom. Though the plot above looked quite good, based on the p-values shown below, we have no evidence that the data came from the NBD model.

Table 9: Goodness of fit test for mentions received

model	chisq	p.value
MLE	166.63	0
Method of Moments	91.65	0
Means and Zeros	219.25	0

3.3 Likes

We follow the same process for likes given as we did for posts and mentions. We note that the tail is a bit longer for likes as some users do lot of post-liking. The median number of likes given is **20** likes though the mean is **46.55** likes.

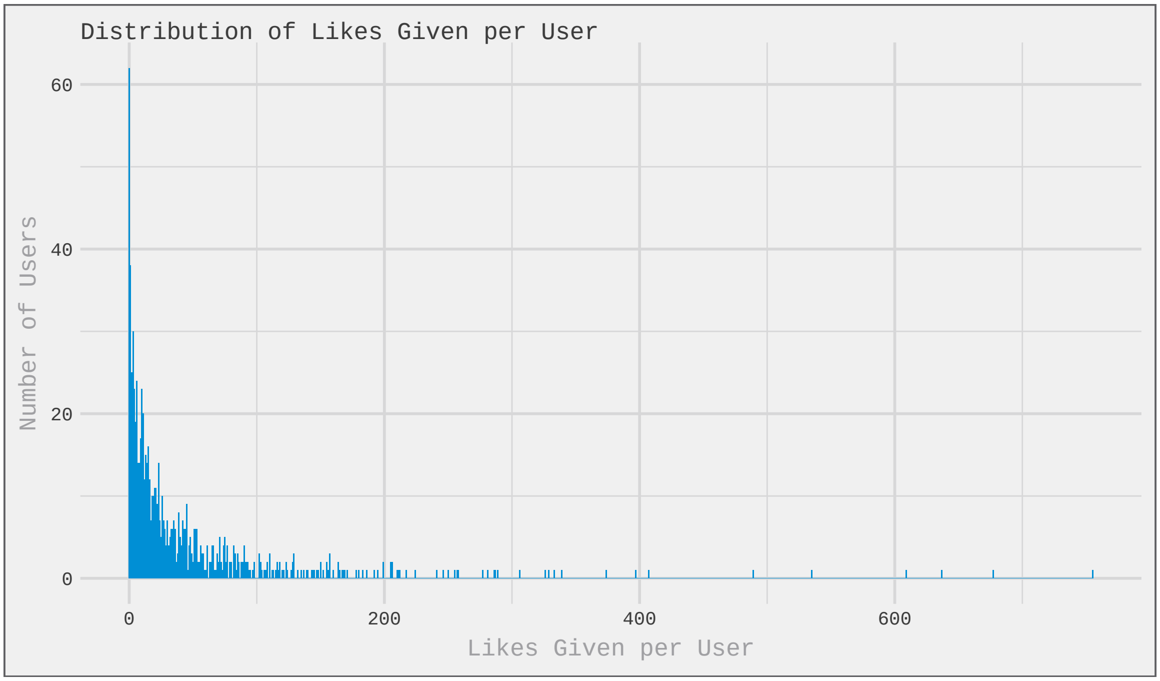


Table 10: Number of users for count of likes given in period (bottom 10)

likes	users
0	62
1	38
2	25
3	30
4	23
5	19
6	24
7	14
8	14
9	17

A careful observer of the plot above may have noted the magnitude of the counts are quite large. This is problematic when calculating gamma functions. For example, $\Gamma(100) = 9.3e^{155}$. To handle this, we used log-gamma and log-factorial functions and restated the first term of the NBD equation as

$$\frac{\Gamma(r+x)}{\Gamma(r)x!} = e^{lgamma(r+x)-(lgamma(r)+lfactorial(x))} \tag{1}$$

We estimate the parameters using each of the three methods as before and again find that the zero-inflated model does not fit the data and that the method of moments estimate is quite different from the MLE and means and zeros estimate.

Table 11: NBD parameters estimates for different methods

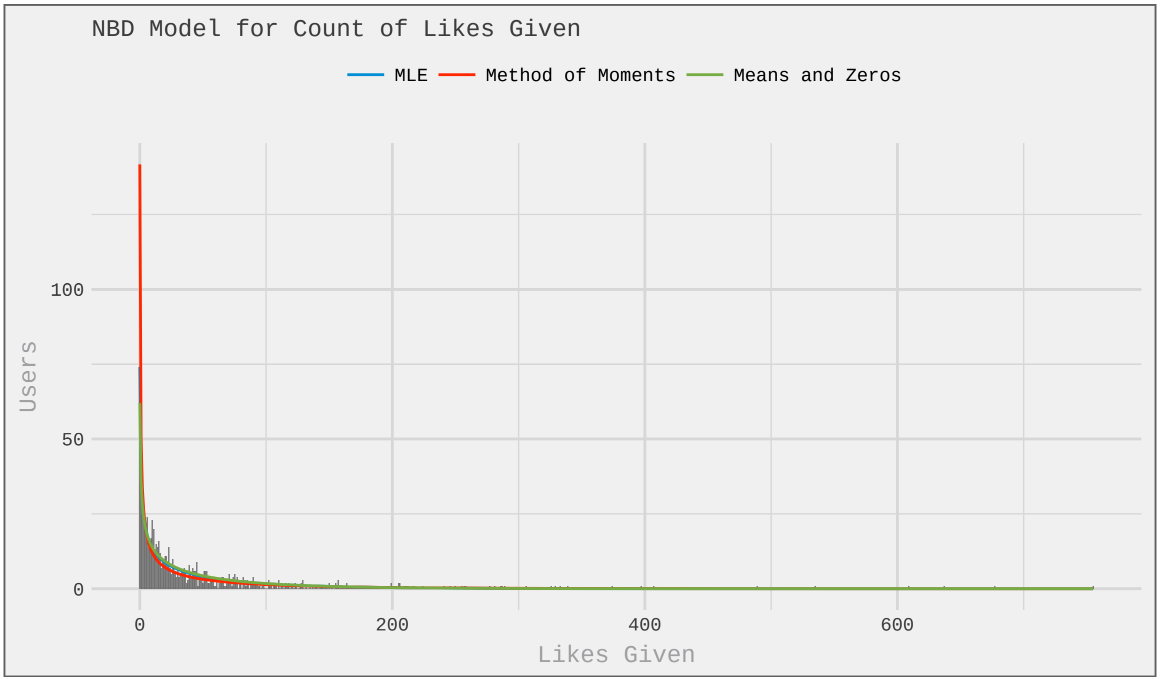
model	r	alpha	pi
MLE	0.5350	0.0115	
MLE (Zero-Inflated)	0.5350	0.0115	0
Method of Moments	0.3580	0.0077	
Means and Zeros	0.5864	0.0126	

Below is a comparison of the expected counts for the left-end of the likes distribution:

Table 12: Estimated number of users for likes given (<= 5) by different estimation methods

likes	Actual	MLE	Method of Moments	Means and Zeros
0	62	74	142	62
1	38	39	50	36
2	25	30	34	28
3	30	25	26	24
4	23	22	22	21
5	19	19	19	19
6	24	18	17	18
7	14	16	15	16
8	14	15	14	15

Aside from the large spike for the method of moments, the MLE and means and zeros model do not look too bad. However, we can see quite a few gray spikes above the blue and green lines in the 10-30 range indicating poor fit there.



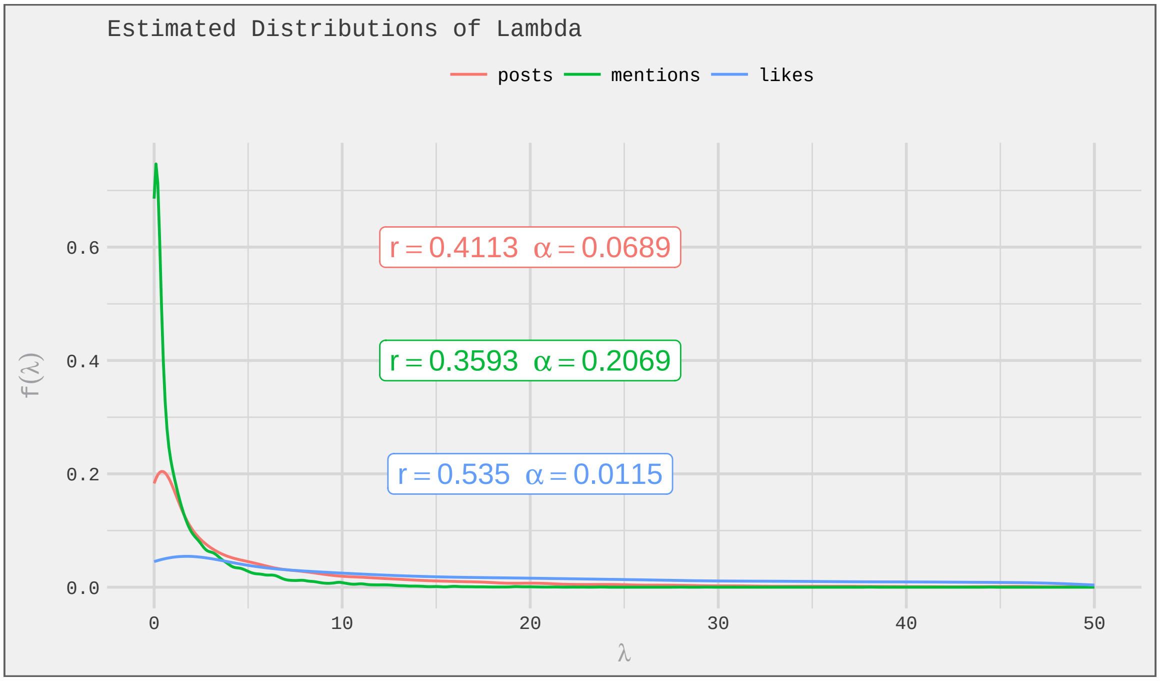
Finally we perform the χ^2 goodness of fit test and first rollup the tail so that 80% of the expected counts have more than 5 counts. We create a 35+ bucket and calculate the χ^2 test statistic and p-value for each paremeter estimation method using 35 - 2 - 1 = 32 degrees of freedom. Based on the p-values shown below, we have no evidence that the data came from the NBD model. Moreover, we note how large the χ^2 test statistics are for the likes given in comparison to posts or mentions, indicating how poor a fit the NBD model is for the likes given data.

Table 13: Goodness of fit test for likes given

model	chisq	p.value
MLE	5614	0
Method of Moments	1387	0
Means and Zeros	9180	0

3.4 Compare Parameter Estimates

variable	r	alpha
posts	0.4113	0.0689
mentions	0.3593	0.2069
likes	0.5350	0.0115



3.5 Lorenz Curve

3.6 Gender

4 Limitations