Wharton 2018 GroupMe

MTKG776: Applied Probability Models in Marketing

2017-02-23

Contents

Word Count: 2,419

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 Background

2.1 Objective

Our objective with this Wharton 2018 GroupMe analysis is to answer the following questions:

1. Is the rate of posting, being mentioned, and liking posts different?

2. Is there variation in the concentration of posting, being mentioned, and liking posts?
3. Do females and males interact differently with the platform through the lens of posting, being mentioned, and liking posts?

2.2 GroupMe Platform

You have just decided to get your MBA 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 created in 2010 and 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 screenshot of the Wharton 2018 GroupMe that shows the following three primary actions:

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

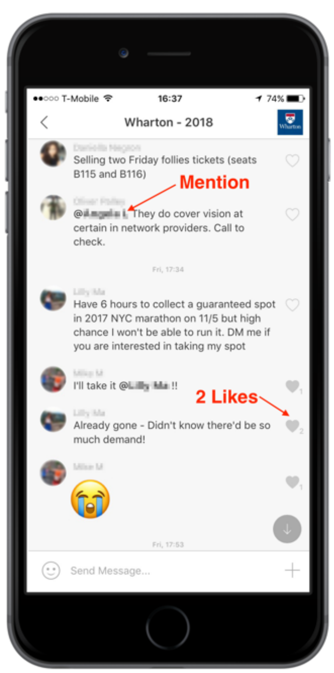


Figure 1: Screenshot showing posts, mentions, and likes

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:

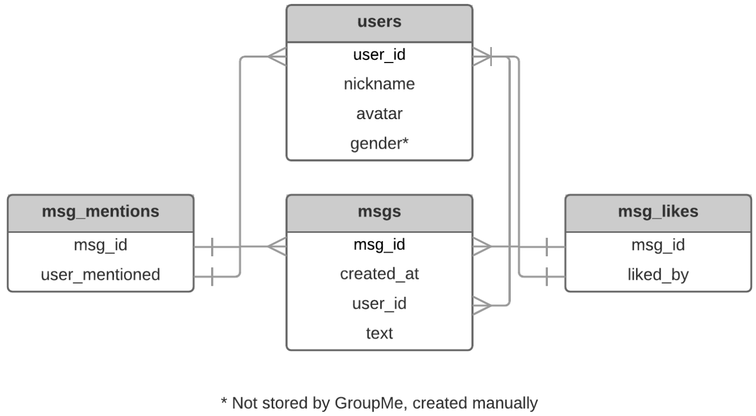
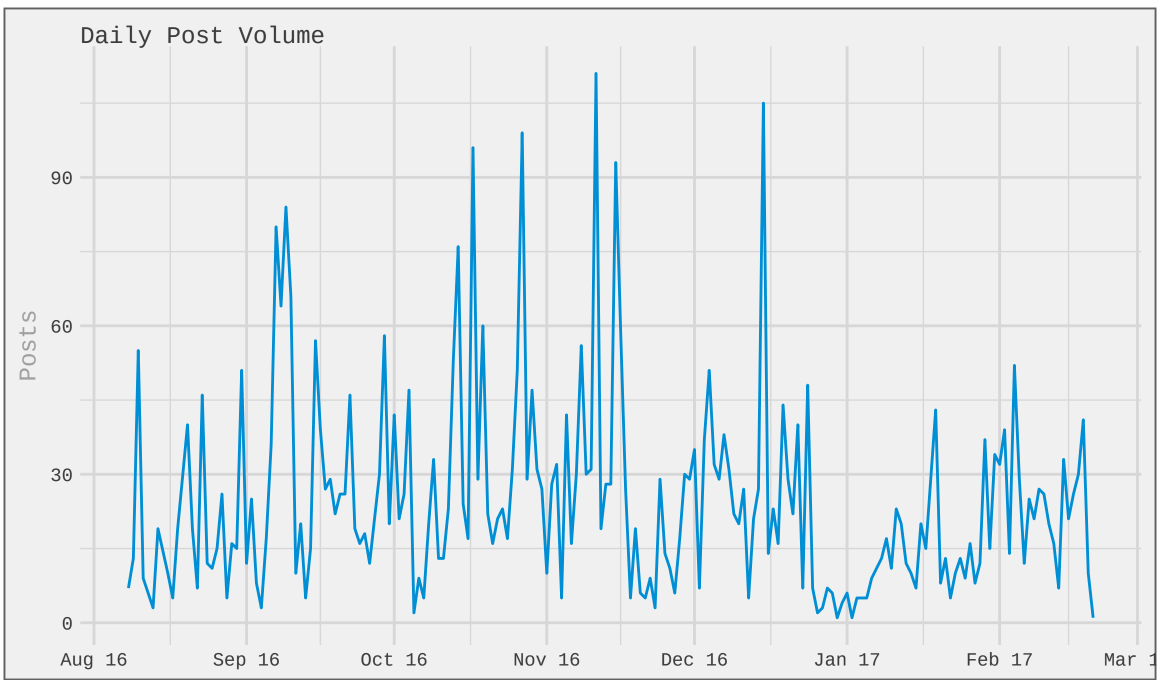


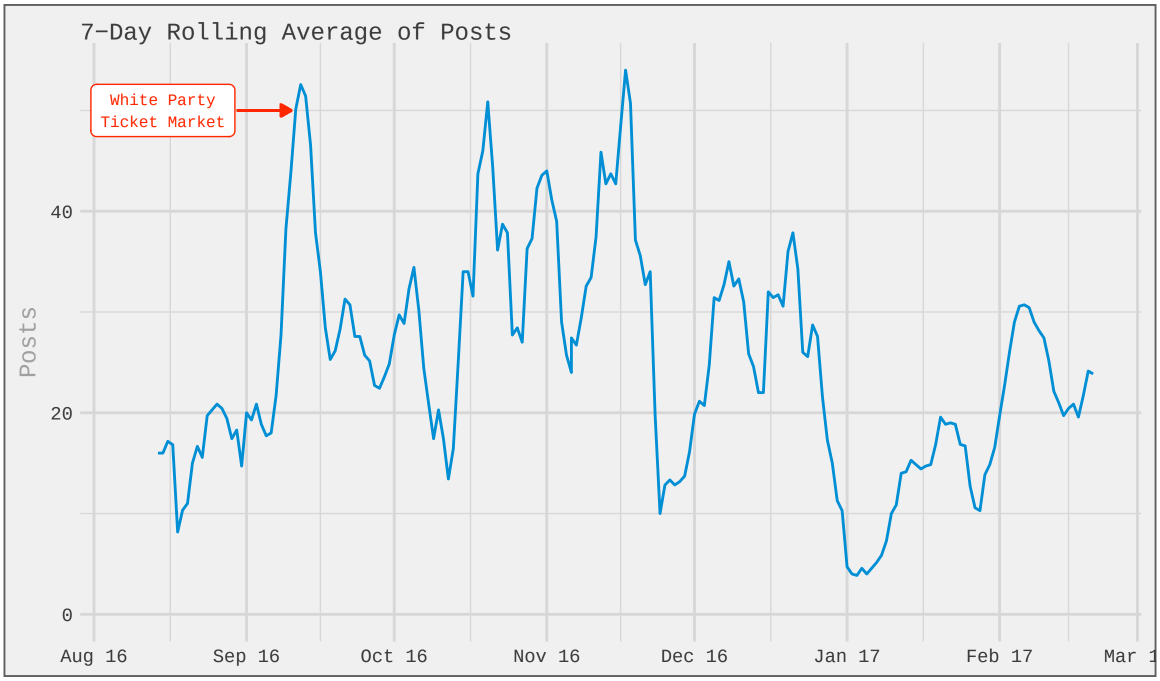
Figure 2: GroupMe data organization

2.3 Wharton 2018 GroupMe

There are **812** users in the Wharton 2018 GroupMe, covering the approximately 850 members of the Wharton 2018 MBA class. Though the group was created in January 2016, we trimmed the dataset to start on 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 Limitations section. The last post in our dataset is **2017-02-20 00:48:37**, thus covering **196** days or **28** weeks. There have been **4,855** posts by **565** 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.



2.4 Count Datasets

2.4.1 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

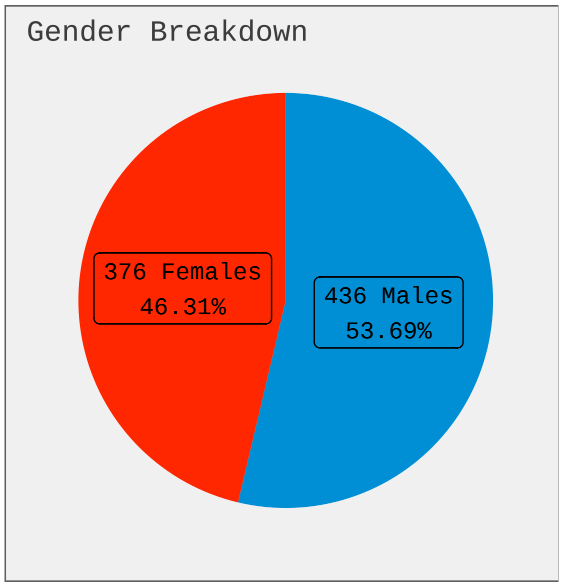
Event	Individual-level Story	Source of Heterogeneity
Posts	Users in the Wharton 2018 GroupMe can post as many times as they would like - there is no upper bound. Thus we can think of each user as having a **post rate**, λ, in the observed time window.	Users interact with GroupMe differently. Some post 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 might expect to observe differences 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.

2.4.2 Gender

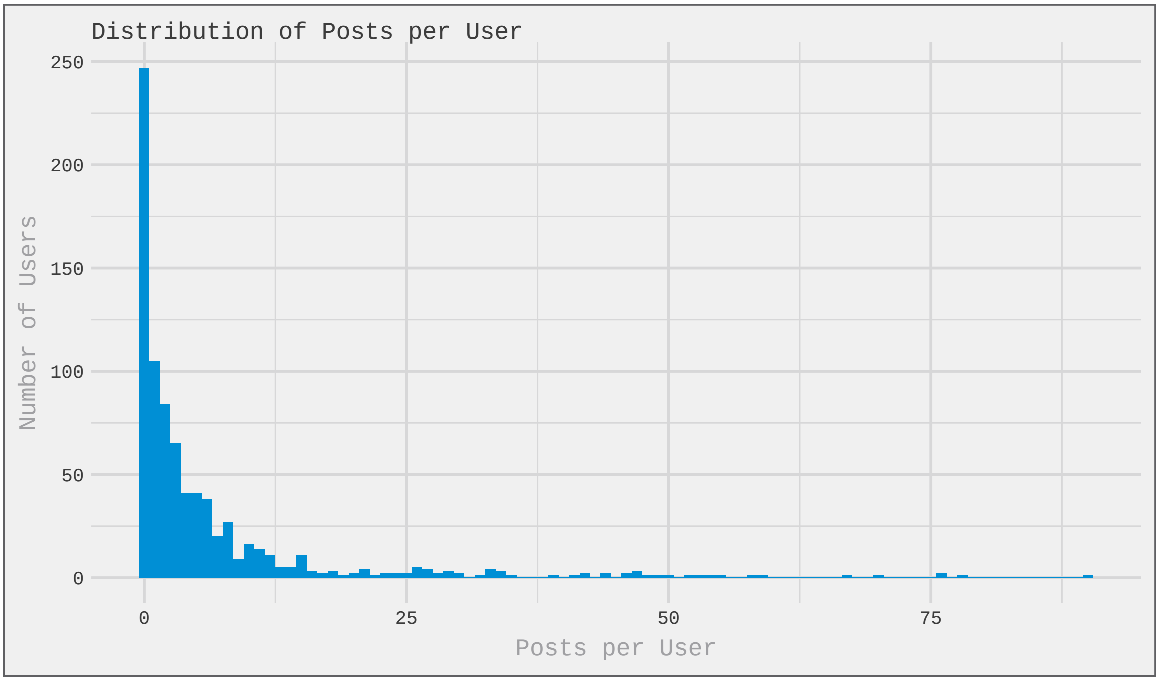
In addition to three behaviors that are the primary interest of this analysis, we included an attribute of the user: gender. We will use this to identify if there are differences in posting, being mentioned, or liking between male and female Wharton students.



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 **5.98** posts (sd = **11.17**).



The data is of the form:

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

posts	users
0	247
1	105
2	84
3	65
4	41
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

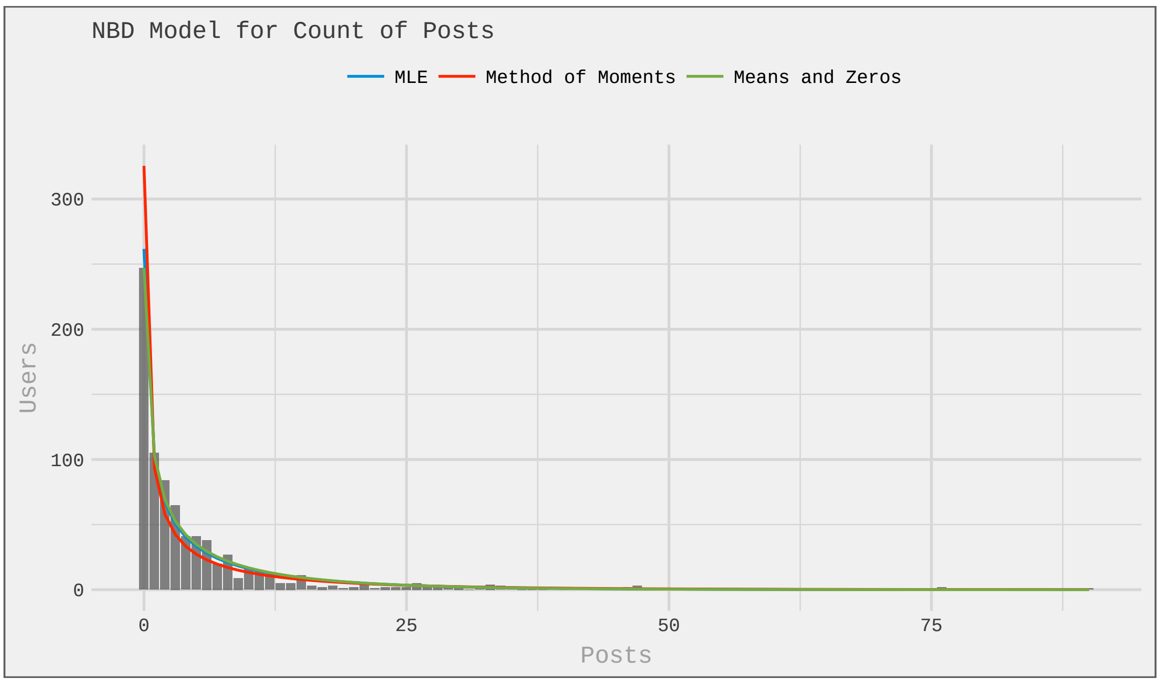
model	r	alpha	pi
MLE	0.4136	0.0692	
MLE (Zero-Inflated)	0.4136	0.0692	0
Method of Moments	0.3009	0.0503	
Means and Zeros	0.4462	0.0746	

We note the divergence between the method of moments and MLE/means and zeros parameter estimates. The large standard deviation, **11.17**, 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 post 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	247	262	325	247
1	105	101	93	103
2	84	67	58	69
3	65	50	42	52
4	41	40	33	42
5	41	33	27	35



In order 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 25+ bucket so that 84.6% expected counts are greater than 5. We calculate the χ^2 test statistic and p-value for each parameter 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	51.9	0.000322
Method of Moments	94.8	0.000000
Means and Zeros	50.1	0.000569

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 (sd = **3.63**).

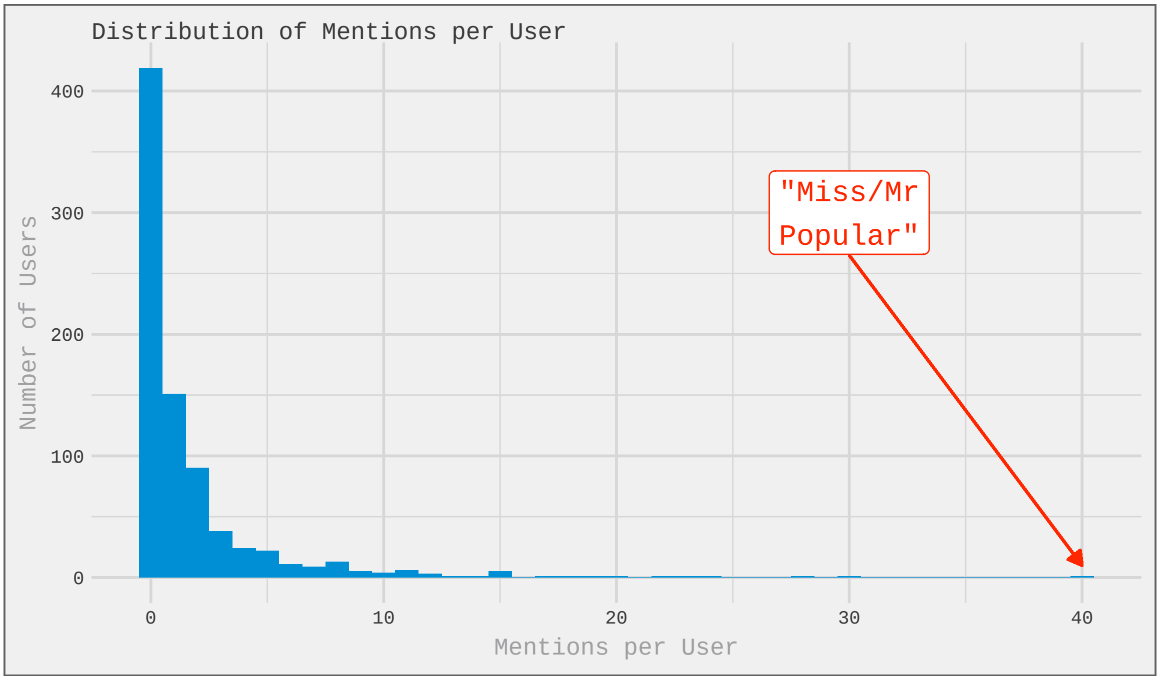


Table 6: Number of users for count of mentions in period (bottom 10)

mentions	users
0	419
1	151
2	90
3	38
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.

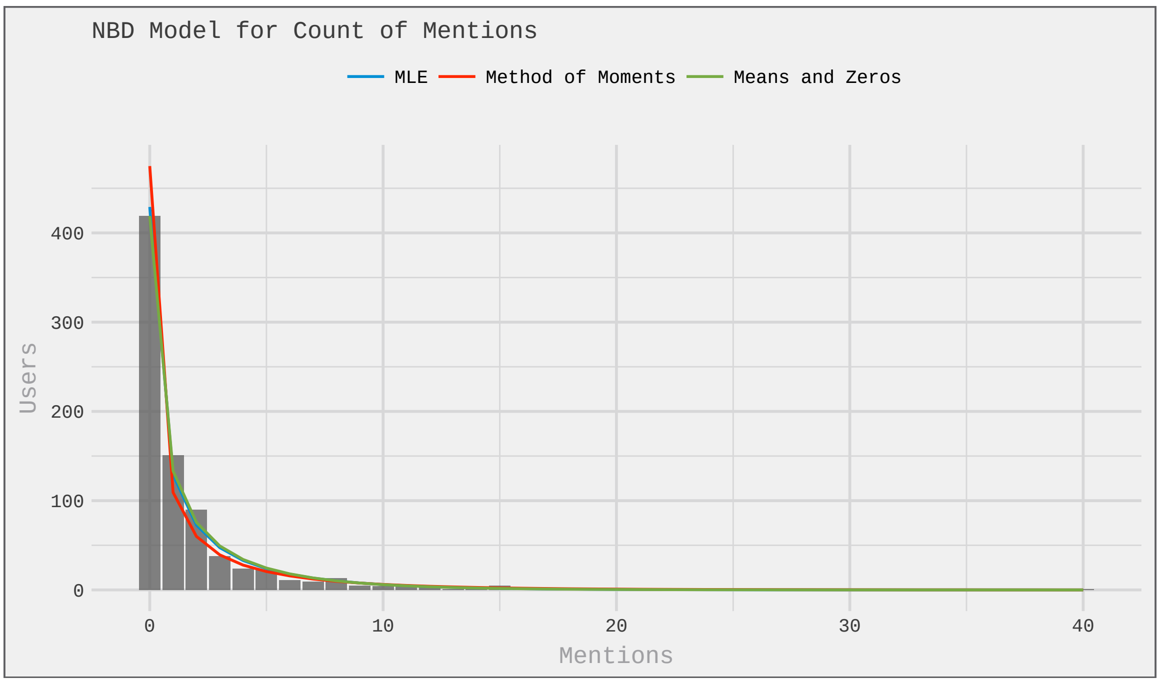
Table 7: NBD parameters estimates for different methods

model	r	alpha	pi
MLE	0.3629	0.2084	
MLE (Zero-Inflated)	0.3629	0.2084	0
Method of Moments	0.2650	0.1522	
Means and Zeros	0.3892	0.2235	

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

mentions	Actual	MLE	Method of Moments	Means and Zeros
0	419	429	475	419
1	151	129	109	133
2	90	73	60	76
3	38	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 so that 100% expected counts are greater than 5. We 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 do not have strong evidence that the data came from the NBD model, ignoring the method of moments as a poor fit.

Table 9: Goodness of fit test for mentions received

model	chisq	p.value
MLE	18.32	0.01061
Method of Moments	43.33	0.00000
Means and Zeros	17.14	0.01654

3.3 Likes

We rinse and repeat, following 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 **46.65** likes (sd = **78.21**).

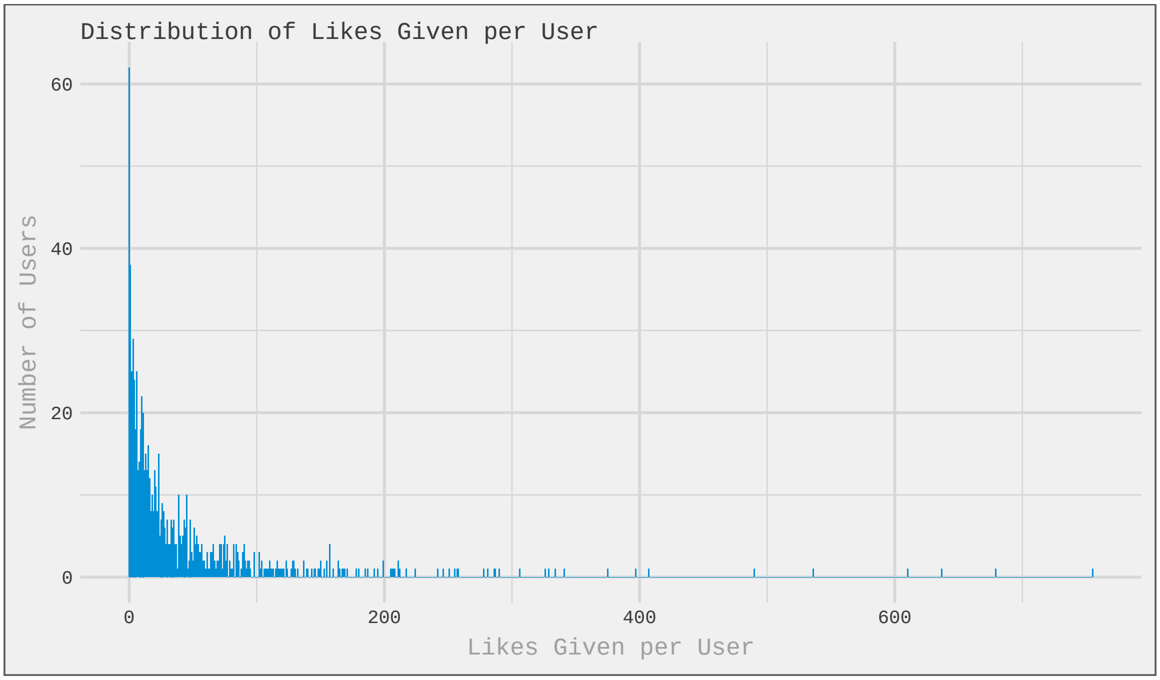


Table 10: Number of users for count of likes given in period (bottom 10)

likes	users
0	62
1	38
2	25
3	29
4	24
5	18
6	25
7	13
8	14
9	18

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, $\Gamma(100) = 9.3e^{155}$. Now imagine $\Gamma(600)$. 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

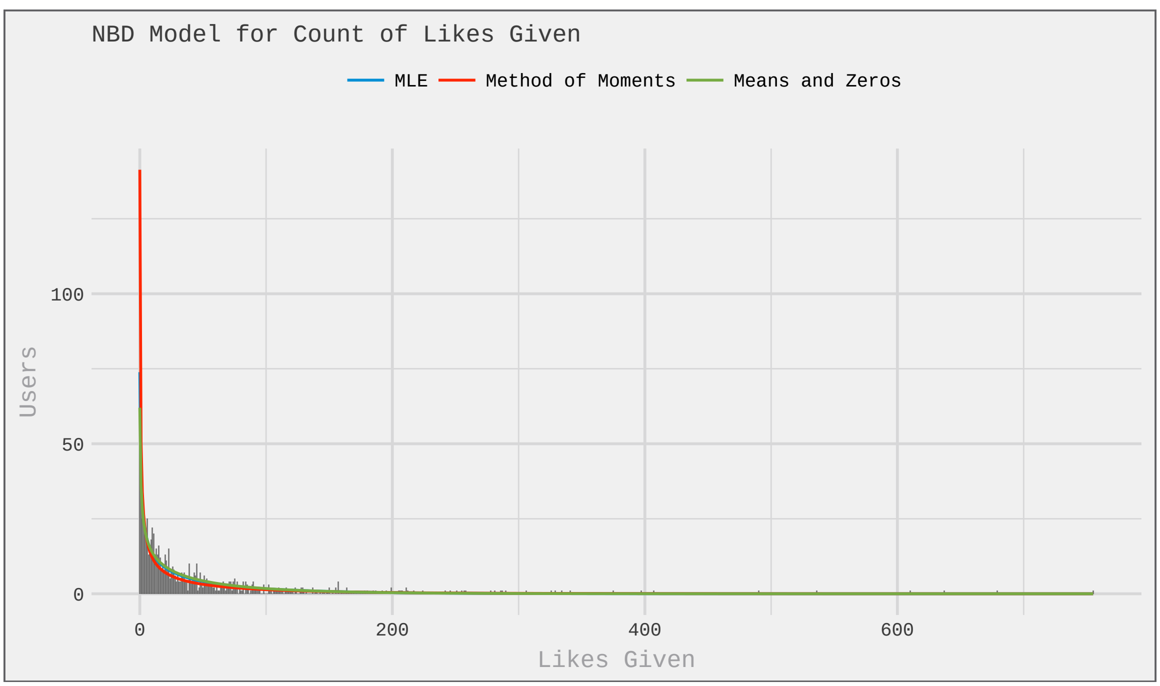
model	r	alpha	pi
MLE	0.5352	0.0115	
MLE (Zero-Inflated)	0.5352	0.0115	0
Method of Moments	0.3586	0.0077	
Means and Zeros	0.5860	0.0126	

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	141	62
1	38	39	50	36
2	25	30	34	28
3	29	25	26	24
4	24	22	22	21
5	18	19	19	19
6	25	18	17	18
7	13	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 so that 88.9% expected counts are greater than 5. We 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 evidence that the data came from the NBD model for the MLE and means and zeros estimation methods. The model created by the method of moments fits poorly.

Table 13: Goodness of fit test for likes given

model	chisq	p.value
MLE	39.20	0.1781
Method of Moments	111.47	0.0000
Means and Zeros	36.83	0.2552

4 Results

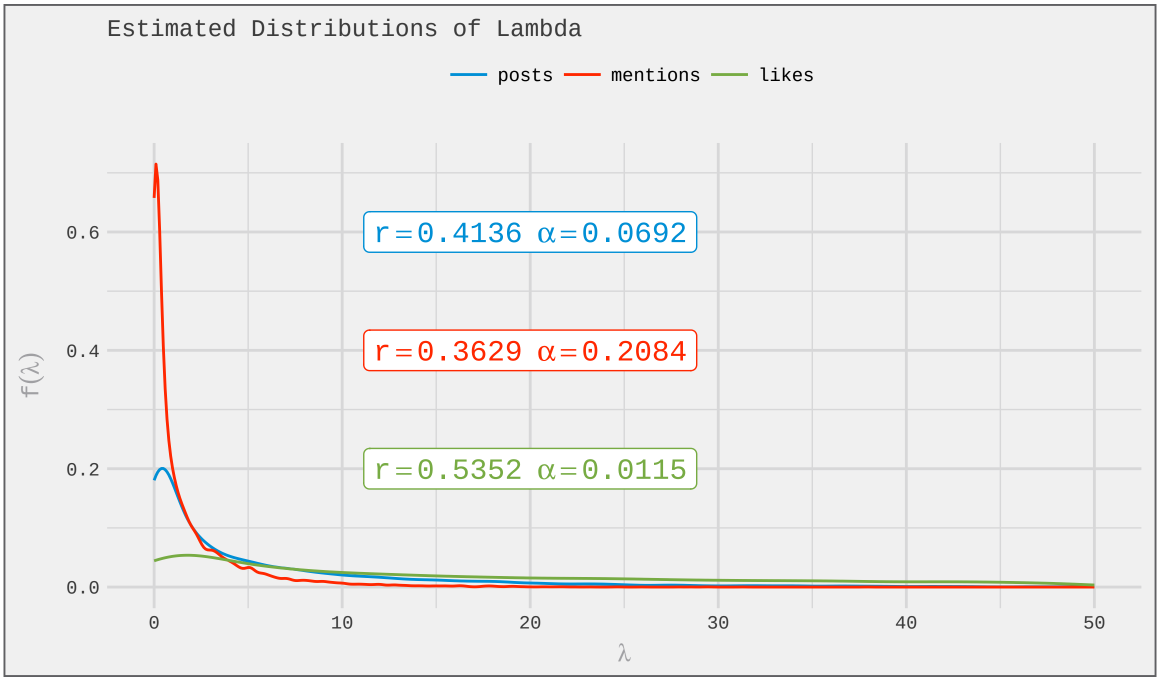
4.1 Activity Rates

Let's get to the answers to our questions. Below is a summary of the parameter estimates (using MLE) for the three behaviors in question. We see that there is in fact a different in rates of each activity. On average, a user posts 5 times more than they get mentioned. Users also like posts 5 times more than they post. So, for the **196** days thus far, you have likely liked 45 posts, posted 5 times, and been mentioned once. The magnitude of the variance (and standard deviance shown above), follow this hierarchy.

Table 14: Summary of model parameters, mean, and standard deviation

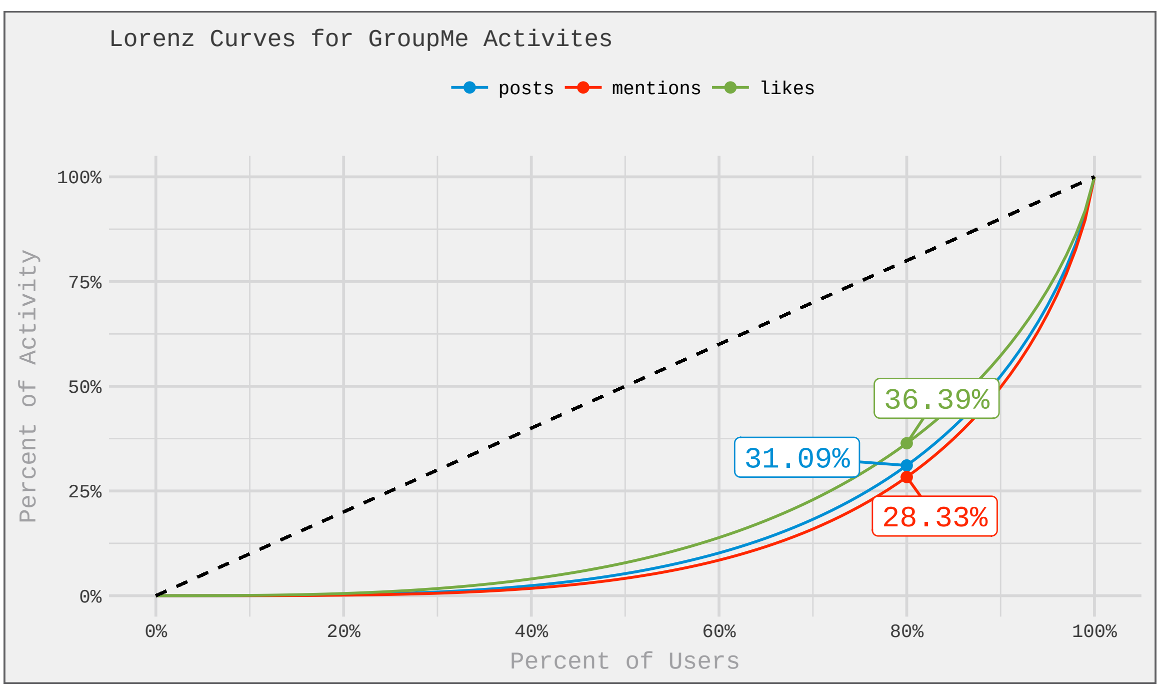
variable	r	alpha	E[X]	sd[X]
posts	0.4136	0.0692	5.979	9.614
mentions	0.3629	0.2084	1.741	3.178
likes	0.5352	0.0115	46.651	64.132

We can also look at the distributions of the three rates, identified as λ in our NBD model. In the plot below we see that there is the most heterogeneity in like rate, the least heterogeneity in mention rate, and the post rate is in the middle. As $r < 1$ for all distributions, each have an interior mean (do not go to ∞ near zero). At this point we have answered question 1: there are differences in post, mention, and like rates.



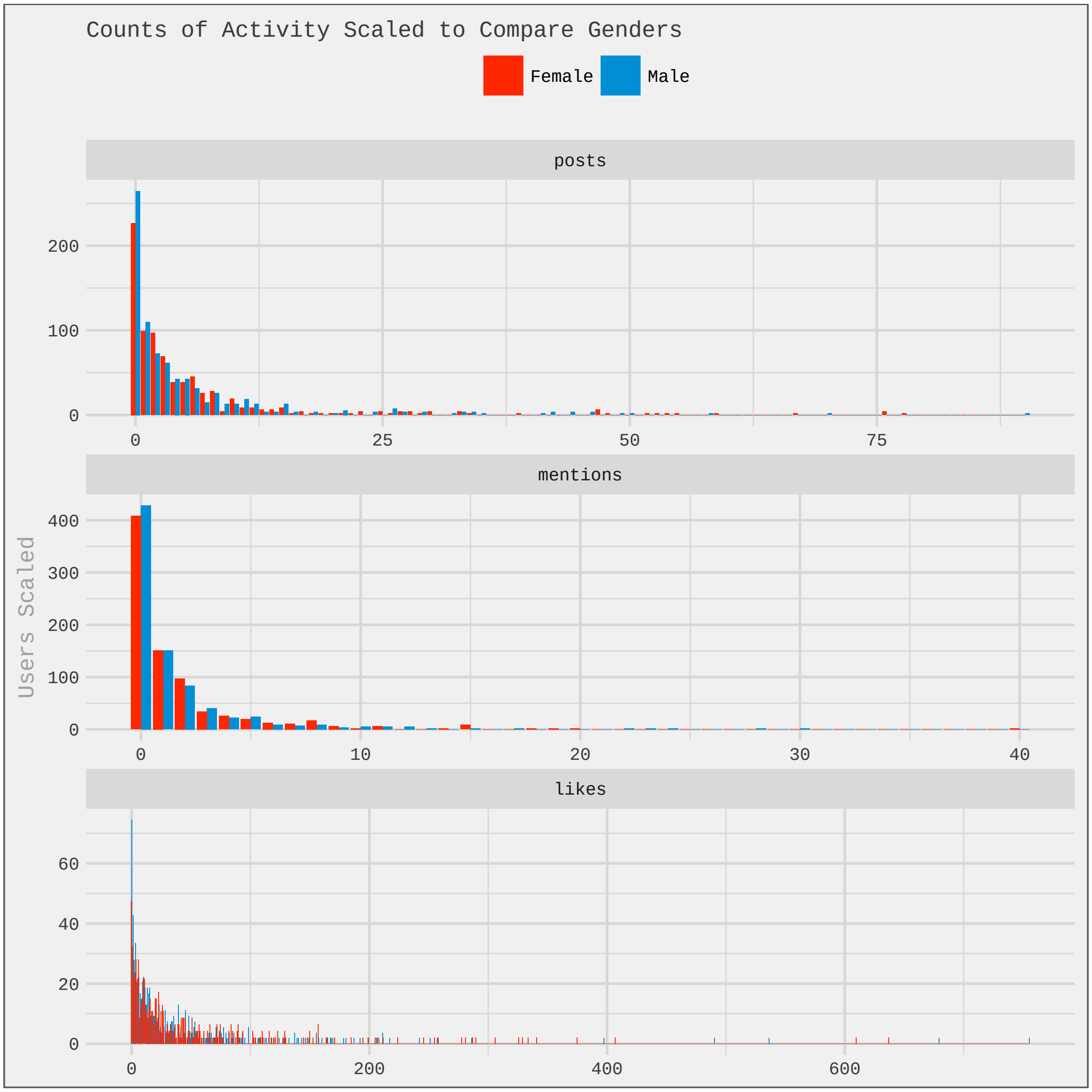
4.2 Concentration

Now we can answer question 2: is there variation in the concentration of each of the GroupMe activities. Using the Lorenze curve and the 80/20 rule highlighted below we see there are some difference, but the differences are not stark. We see that being mentioned is concentrated in the fewest number of users (20% of users account for 1 - 28.3% = 71.7% of the mentions). This follows intuitively from the histogram in the NBD model section. In contrast, likes are the least concentrated (20% of users account for 1 - 36.4% = 63.6% of the likes). So, we found that mentions are more concentrated than likes, with post in between. However, the differences are not tremendous.



4.3 Gender Differences

We move into treacherous waters: asking if there are differences between the genders. To answer question 3 we start with side-by-side histograms of each of the three activies, scaled for differences in the number of females (376) and males (436).



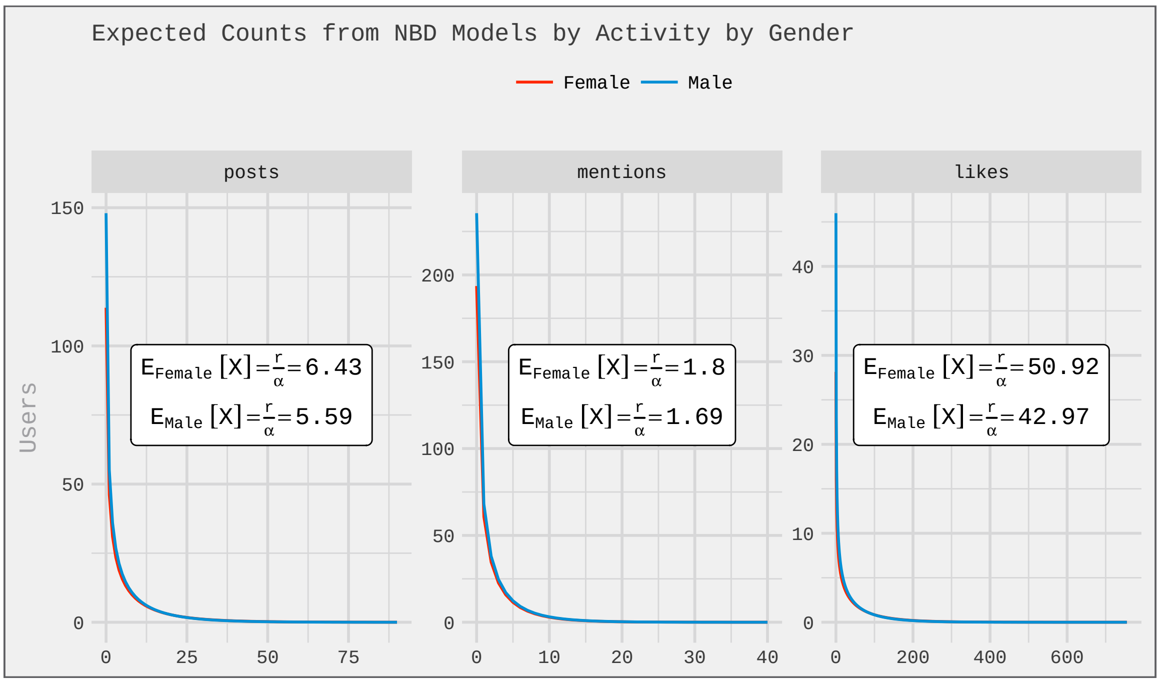
Next, we fit an NBD model for each activity, for each gender and combined, using MLE without and with a spike at zero. Based on the (large) table below we immediately see that none of the zero-inflated models are appropriate.

Table 15: Estimated model parameters and log-likelihood for each model by gender

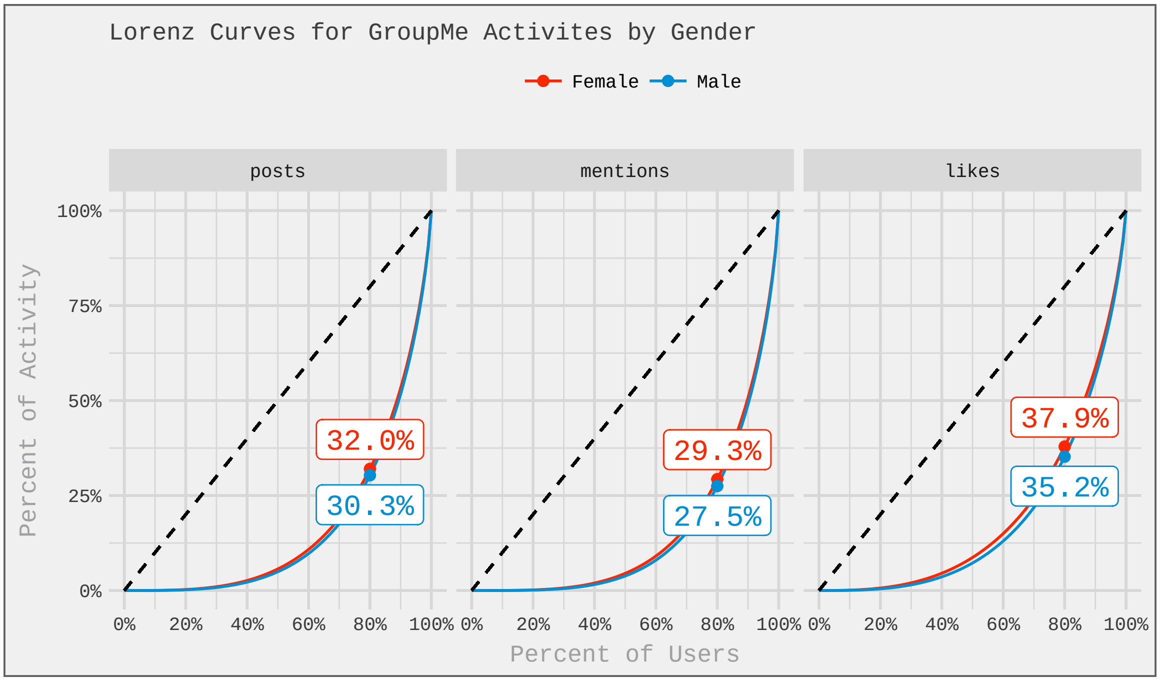
activity	gender	model	r	alpha	pi	ll
posts	Female	MLE	0.4323	0.0672		-1045.3
posts	Male	MLE	0.3988	0.0714		-1148.2
posts	Combined	MLE	0.4136	0.0692		-2194.6
mentions	Female	MLE	0.3800	0.2111		-649.4

activity	gender	model	r	alpha	pi	ll
mentions	Male	MLE	0.3488	0.2064		-724.7
mentions	Combined	MLE	0.3629	0.2084		-1374.4
likes	Female	MLE	0.5772	0.0113		-1818.1
likes	Male	MLE	0.5048	0.0117		-2010.3
likes	Combined	MLE	0.5352	0.0115		-3831.1
posts	Female	MLE (Zero-Inflated)	0.4323	0.0672	0	-1045.3
posts	Male	MLE (Zero-Inflated)	0.3988	0.0714	0	-1148.2
posts	Combined	MLE (Zero-Inflated)	0.4136	0.0692	0	-2194.6
mentions	Female	MLE (Zero-Inflated)	0.3800	0.2111	0	-649.4
mentions	Male	MLE (Zero-Inflated)	0.3488	0.2064	0	-724.7
mentions	Combined	MLE (Zero-Inflated)	0.3629	0.2084	0	-1374.4
likes	Female	MLE (Zero-Inflated)	0.5772	0.0113	0	-1818.1
likes	Male	MLE (Zero-Inflated)	0.5048	0.0117	0	-2010.3
likes	Combined	MLE (Zero-Inflated)	0.5352	0.0115	0	-3831.1

So, we move to plotting the expected counts of each activity for females and males based on the NBD model. We see that the expected counts are remarkabely similar for each activity, though the r and α parameters are slightly different.



Using the Lorenz curves below we see that the activites are a bit more concentrated for males than for females. We can see this in the histogram at the beginning of this section. For each activity, there are more males than females that are not involved. However, this difference is minimal.



Lastly, we use the likelihood ratio test (with degress of freedom $4 - 2 = 2$) to identify if the individual model are better at explaining the behavior than a combined movel. We see from the p-values below, two separate model are not different from the combined model. So, we have answered question 3: females and males uses the platform in a similar fashion.

Table 16: Likelihood ratio test to determine in separate female and male models are appropriate

activity	Female	Male	Combined	chisq	p.value
posts	-1045.3	-1148.2	-2195	2.0475	0.3593
mentions	-649.4	-724.7	-1374	0.5583	0.7564
likes	-1818.1	-2010.3	-3831	5.2614	0.0720

5 Limitations

1. *Users that joined after August 8, 2016 or left and rejoined.* While we removed users that left during the observation period, this analysis assumes that all **812** users were in the Wharton 2018 GroupMe for the entire duration of the observation period. While we know a few cases where this is not true, this occurance is minimal. Unfortunately, GroupMe does not have a join data in the API for users to a group. There are system created posts when users join or when existing users add new users. Unfornately, the way that GroupMe has stored this data has changed over time. The true source of truth if GroupMe generated human-readable text which would need to be extensively parsed and was not done so here.
2. *Interdepedence of activites.* Not discussed in this paper is interdependence of posts, mentions, and likes. An extension of this analysis would explore the interaction between the three.