The GroupMe Economy: A Snapshot

Messages, Likes, Mentions, & Memes

by Mike Lee, Rosalie Reuss, Michael Wedd, & Ben Weinberg

Disclaimer:

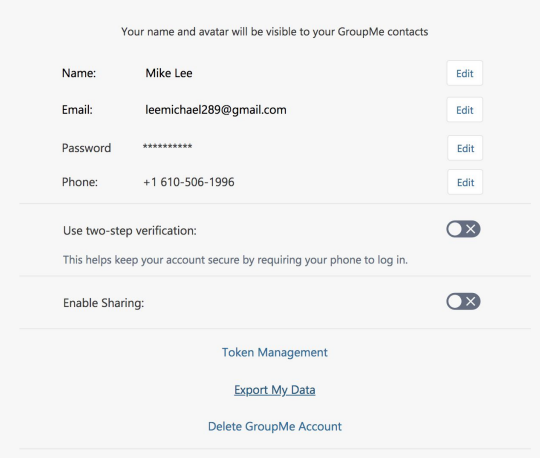
"Take this with a grain of salt. We, the listed group members, recognize the hard limitations of statistical analysis in the absence of large numbers of true randomly sampled observations, methodologies which control for multiple factors, and analysis of the variance in our observations. We have done our best to simply visualize, not infer, trends in our social network, and make no attempt to produce any scientific conclusions from this analysis. Whenever doing this type of work, it is extremely important to be responsible in the methodologies we use and the way in which we present our findings. That being said, please enjoy."

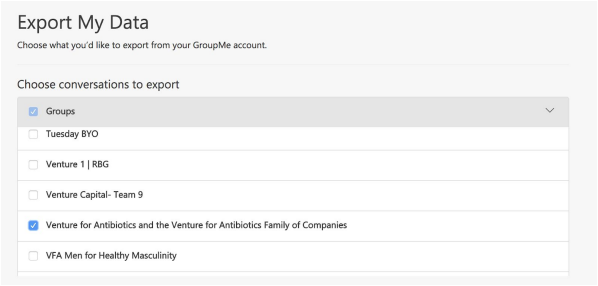
Introduction:

Prior to the age of social media, social interaction, popularity, and influence were considered "soft" and qualitative things which existed solely in the minds of our peers. For better or for worse, these are now some of the more quantifiable aspects of our existence. Every action we perform on social media platforms is tracked and used to target advertisements, train machine learning models, influence elections, profile criminals, and generally map human behavior at scale. The purpose of this project is to provide some entertaining visualizations and demonstrate how easy it is for individuals, corporations, and governments to scrape data from our accounts.

How I got this data:

I was going through GroupMe to manage something on my profile and stumbled upon an interesting link.

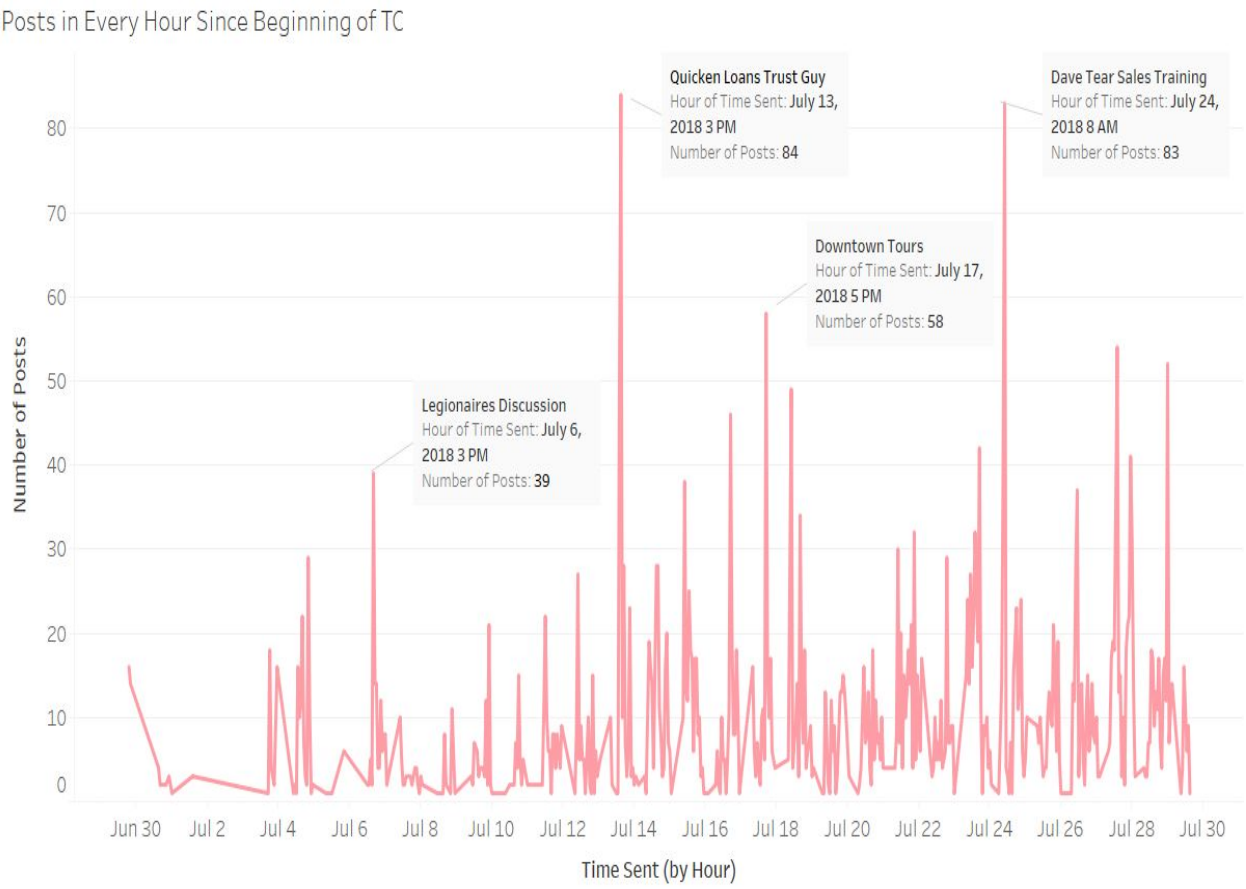




Apparently, GroupMe has an "export data" feature that anyone, as a part of any group, can use to immediately download entire message histories. This includes names, text data, mentions, number of likes, who likes who, and every meme or other questionable thing you've ever posted. Yikes.

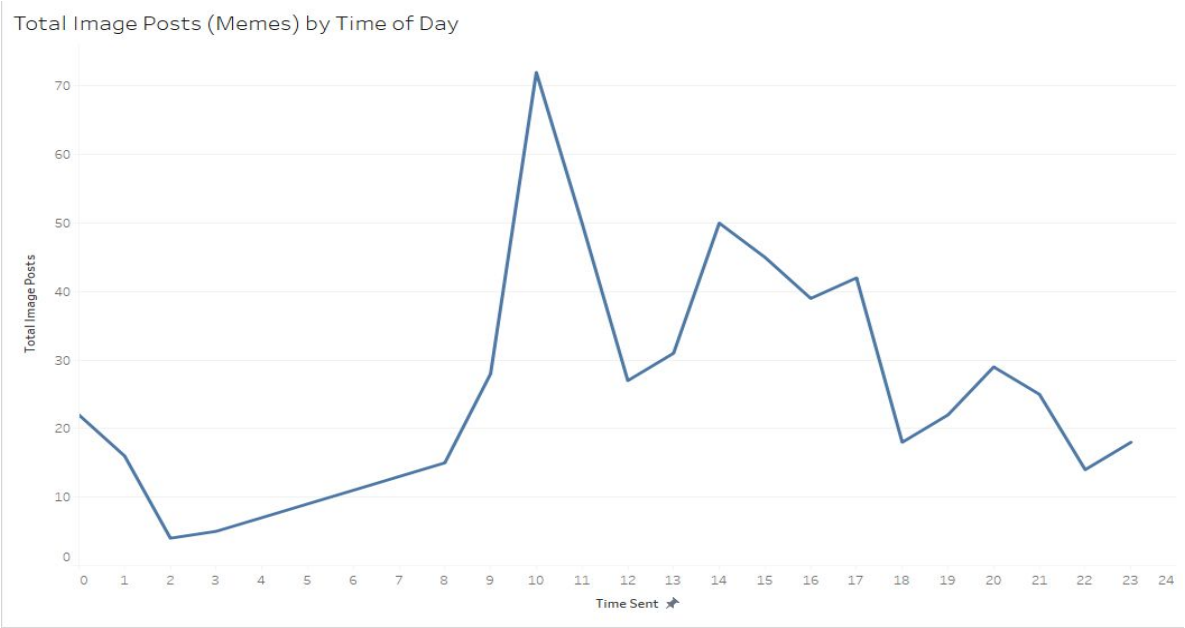
The data was downloaded as a .json file, which was then opened, formatted, and cleaned into a .csv with Python and Microsoft Excel. From there, we transformed, analyzed and visualized our data with R and Tableau. The code to do this has been open sourced on GitHub here: https://github.com/MikeHLee/GroupMeAnalysis , and it can be run with JuPyter Notebooks and R Studio. For privacy reasons, the dataset itself is not hosted on GitHub, but you can get it through GroupMe using the methods shown above.

Total Post Activity Over Time:



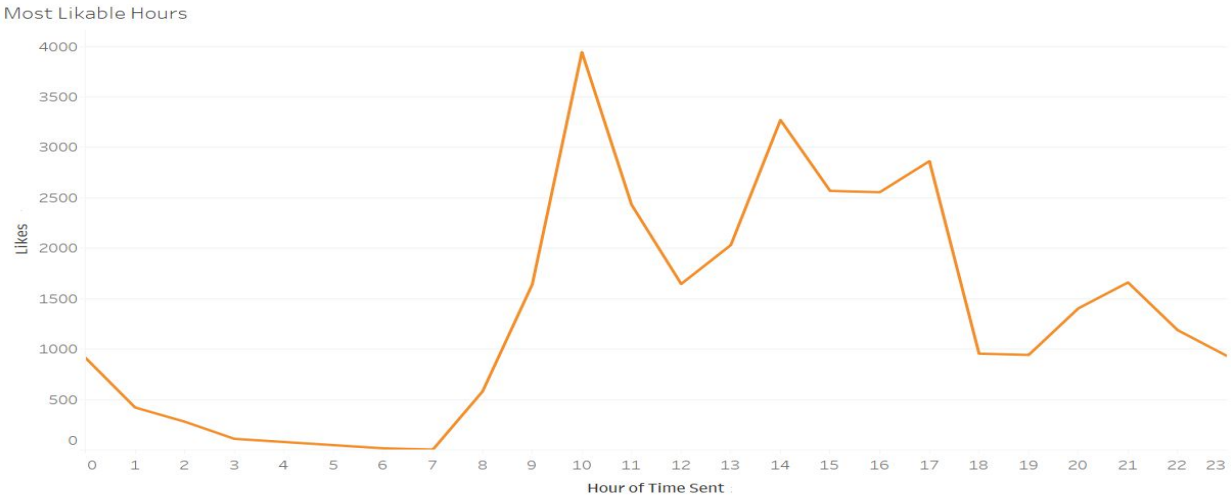
Let's start by looking at our total post volume in every hour since the beginning of training camp. We can see certain periods where activity peaks and lulls. Post frequency is typically at its highest when there is an unpopular speaker or programming event. VFA 2018 **loves** to roast people onstage, but we tend to do it quietly.

Meme Traffic



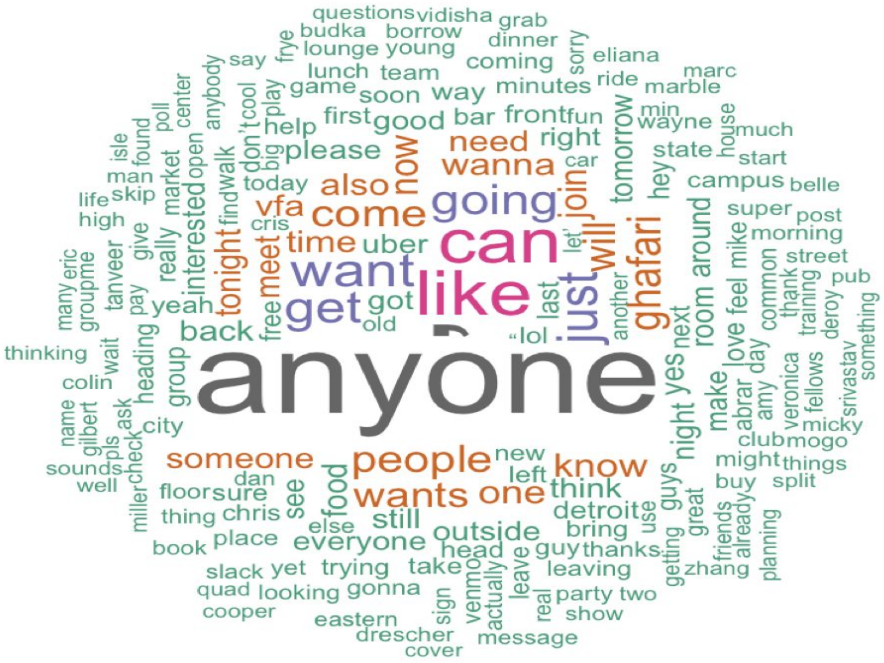
Let's now observe the total number of image posts for each hour of the day over time. Taking image posts as a proxy for memes sent, the most activity tends to occur around 10:00-11:00 am and 2:00-3:00 pm. These tend to be our longest speaker and lecture periods. "Peak Meme" tends to occur at our most bored and restless moments. Possible feedback for Cris Landa with regards to programming.

As might be expected, the most meme-able hours are also when the most likes are awarded.



Is Anyone There?

Most Common Words From Venture for Antibiotics and The Venture for Antibiotics Family of Companies



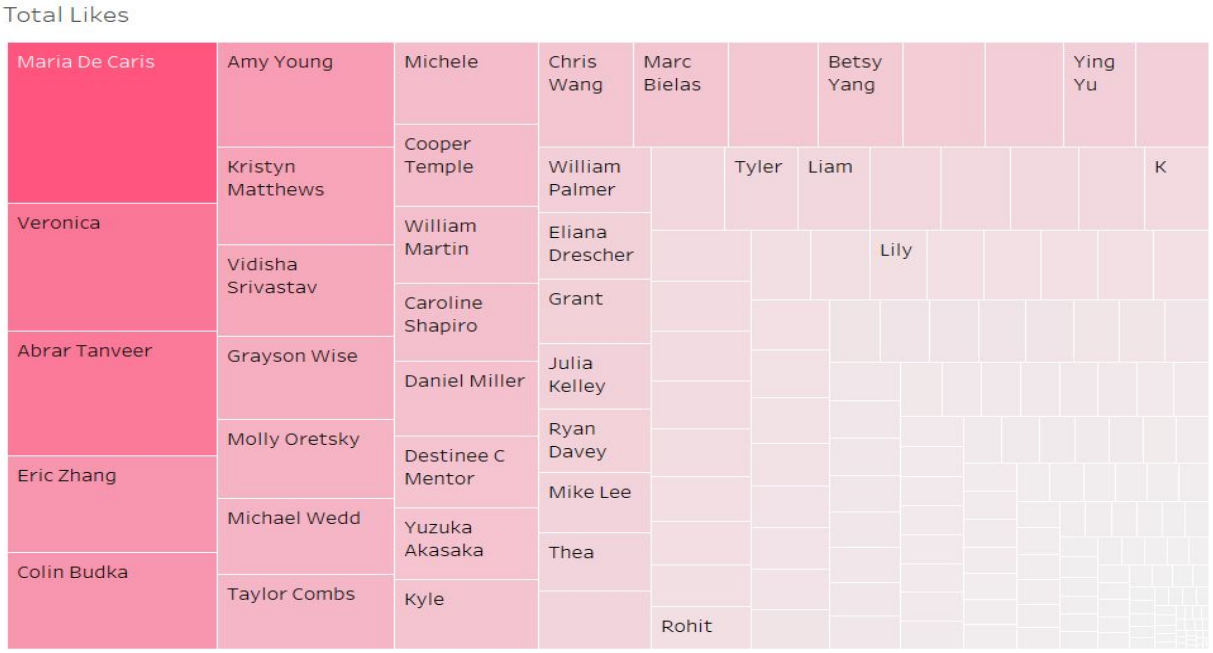
We're a pretty needy group. This is a word cloud of the most commonly mentioned words in "Venture for Antibiotics and The Venture for Antibiotics Family of Companies." Notable chart-toppers are "anyone, someone, people, food, want, come, tonight, group, join, need, wanna, free, meet, and **Cris**."

Also interesting: Most Common Words Among the 200 Most Liked Comments (top 5%)

Post Activity by Individual:

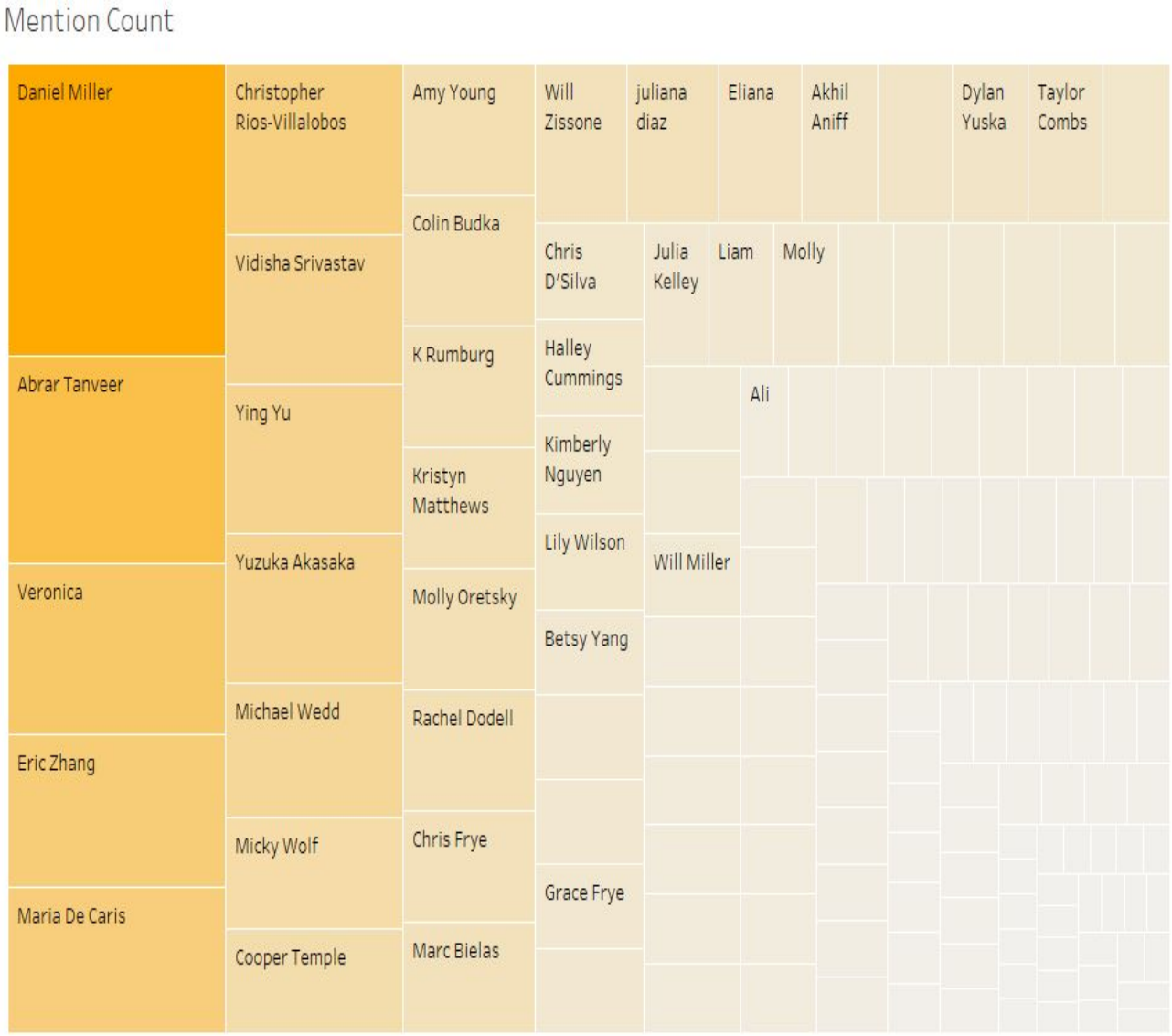
Counting the total number of posts by named individual, Abrar leads the pack followed by Eric, followed by Veronica, followed by Colin, followed by Liam, followed by Maria. In this plot, a larger & darker space corresponds to a higher total post count.

Total Likes and Average Likes/Post by Individual



Self explanatory. Worth noting that the distribution for total likes by person is heavily right-skewed. That is, the vast majority of likes have been accumulated by a very small percentage of the class. Likes/post is slightly more even, however the overall shape and trend remain the same. However, different people appear at the top of the charts for both of these metrics. We wanted to include the bar graph, but the tails were so long that they would not fit on a single page.

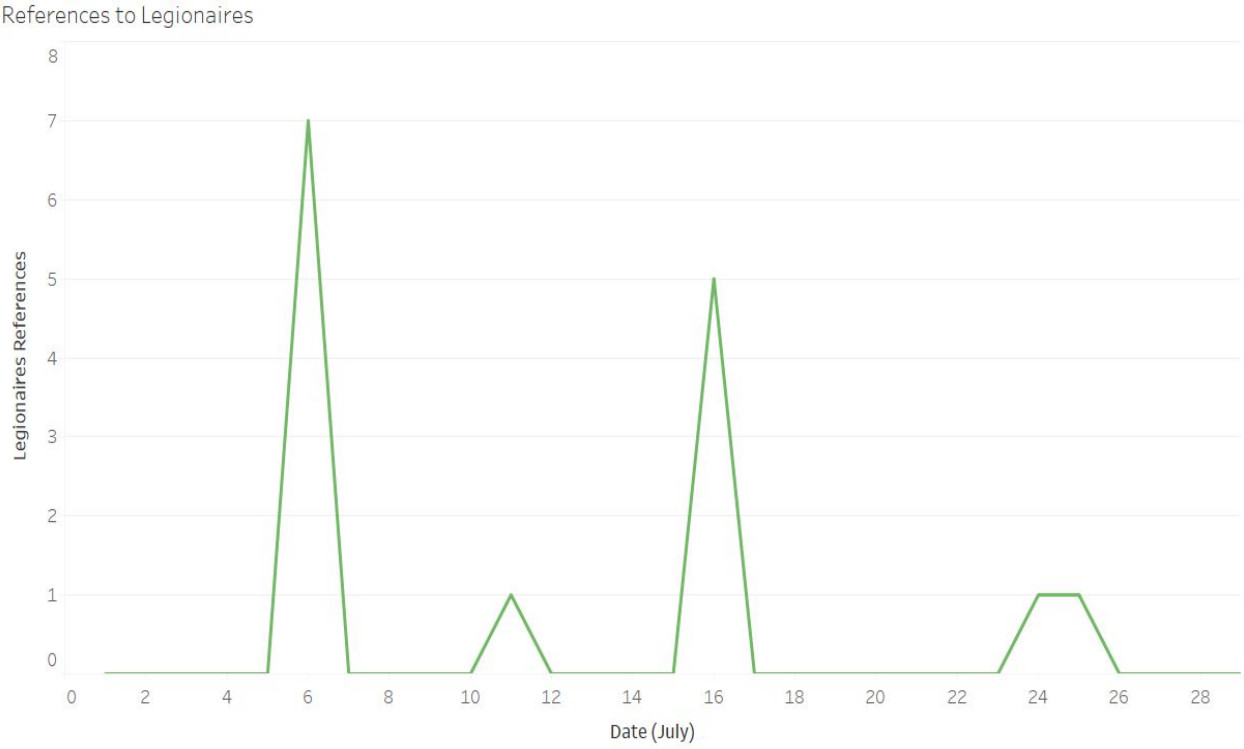
Who is Mentioned the Most?



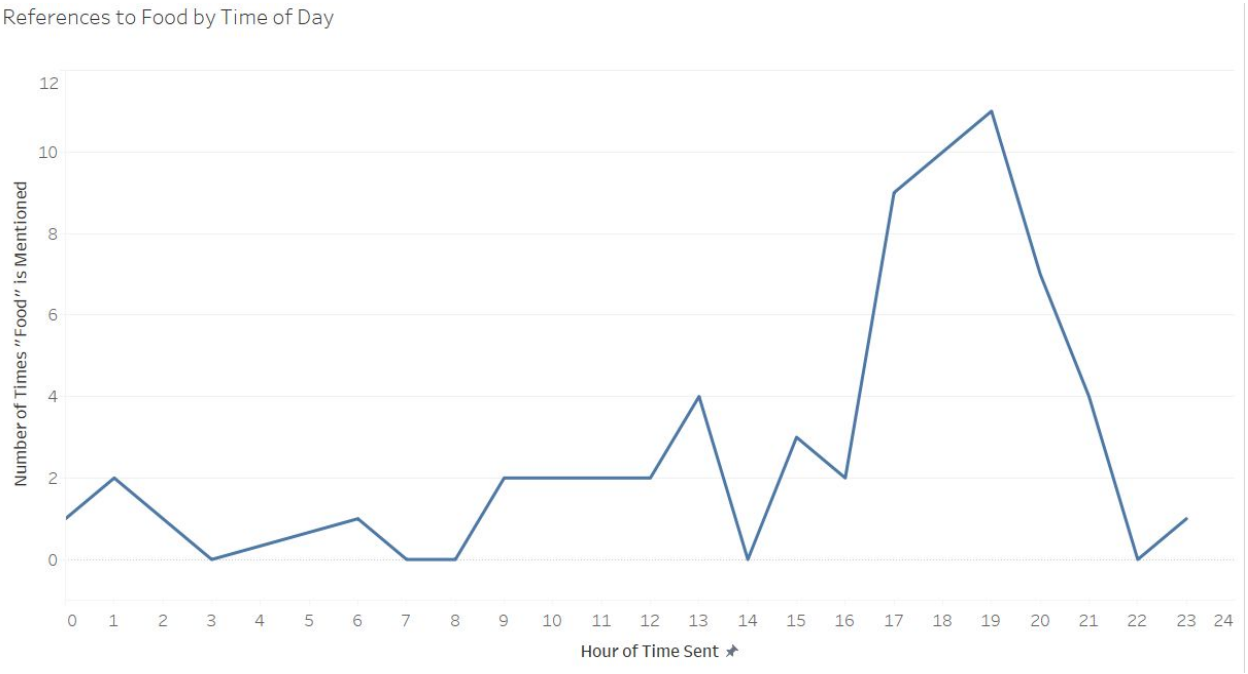
This graph shows how many times an individual is called out in chat with the "@" feature. The most mentioned individuals include Dan Miller, Abrar Tanveer, Veronica, Eric, Maria, Chris, and Vidisha. This distribution is much more even than both total like/person and likes/post/person.

Topics of Interest:

Here, we parsed the data to count the number of times a certain word, word base, or collection of words appeared in message text.



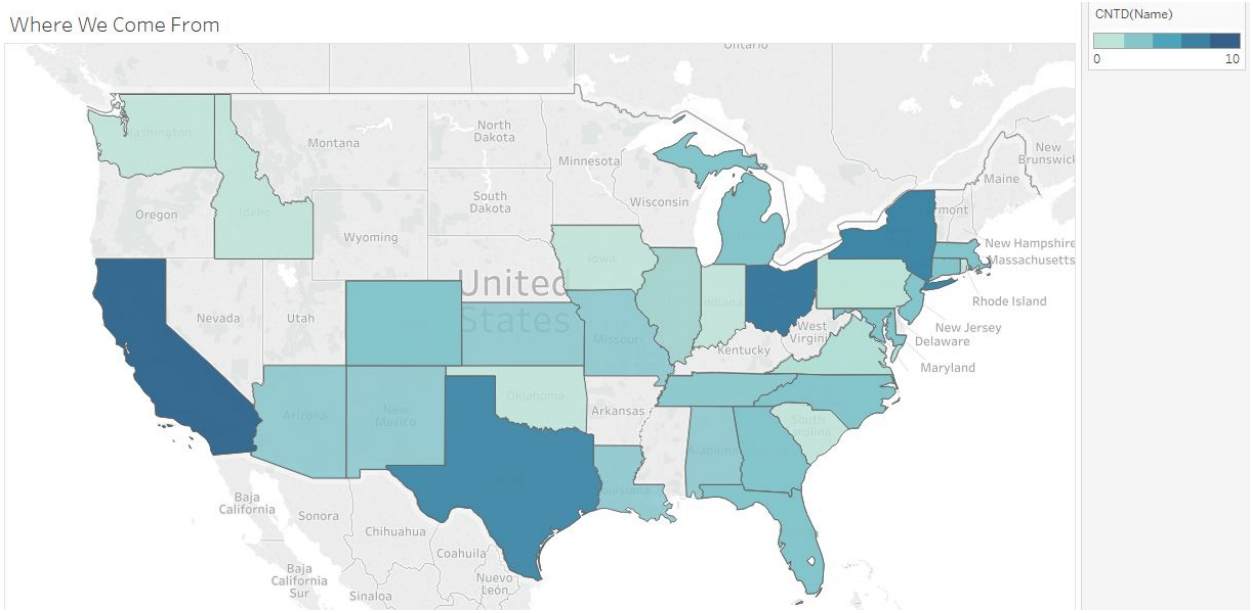
References to Legionaires peaked on July 6th and have declined since.



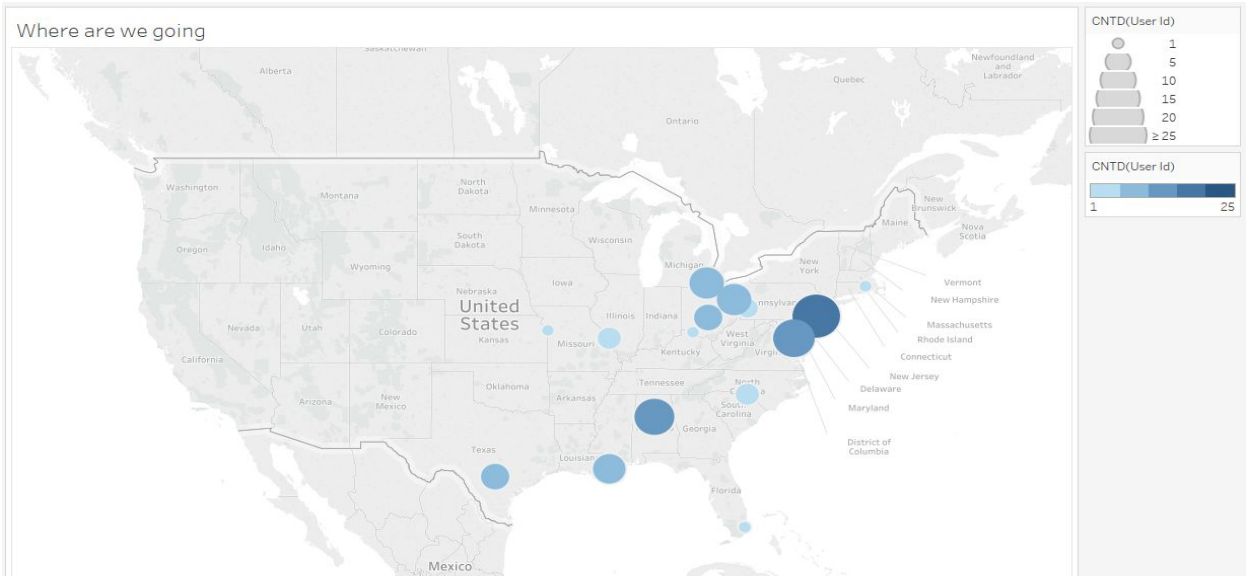
We tend to be at our hungriest around 7:00 pm. Valuable info for local restaurant owners.

Geographic Data:

To enrich our analysis we acquired a dataset of fellow home cities, schools, fellowship locations, etc. and merged it with our groupme tables by name in Tableau. Let's take a look.



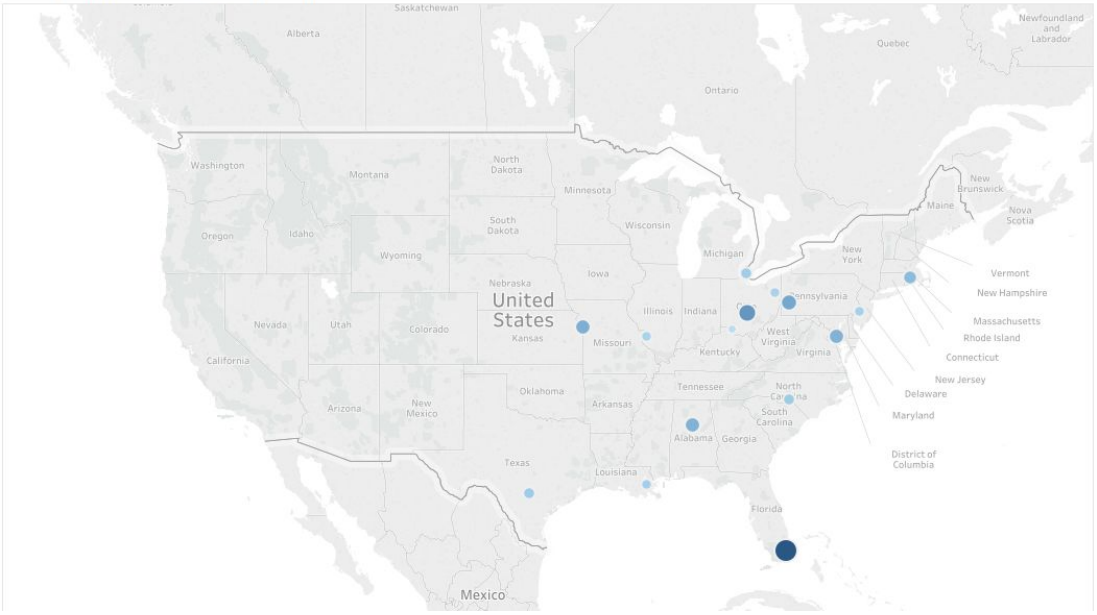
Here's where we are from in the US. Darker states have higher representation among the 2018 class.



Here's where matched 2018 fellows are heading. Good luck to those who are still on the hunt.

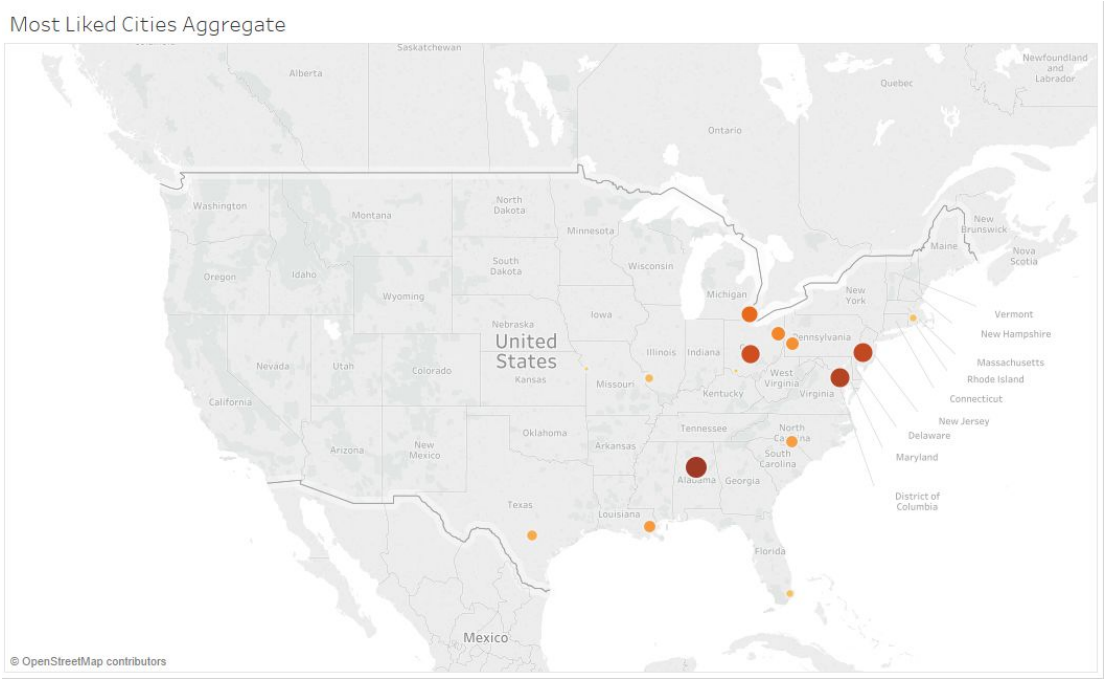
Now let's look at the maps in conjunction with GroupMe data. Here is the aggregate level of participation (total post count) for each VFA city destination. Larger and darker circles indicate that all of the fellows who are heading to that location contribute the most posts.

Most Participatory cities per capita

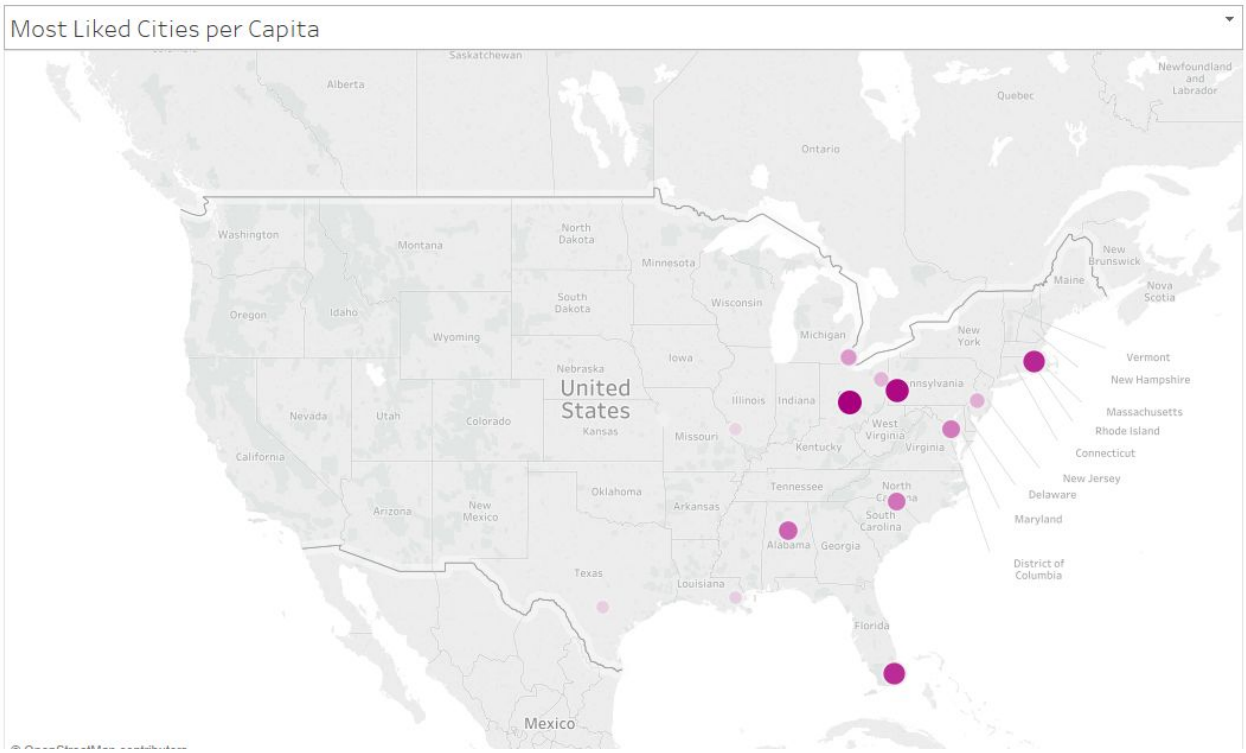


Here are the most popular cities by total likes. Birmingham comes in 1st as the most likeable city.

Most Liked Cities Aggregate



Also interesting: total likes/person/city



Conclusion:

There is a lot more that could be done with this. Features of the data we didn't touch include "who likes each comment," "who likes who's comments," "apple or android device user, " "fellow undergrad schools" and various system messages. Correlations, regressions, and statistical learning were also not applied, as these methods can be highly misleading if done improperly. To explore further, please request from Mike Lee, Rosalie Reuss, or Ben Weinberg. In all honesty though, you probably don't want to.

Take issue with this? Think it's creepy? Email support@groupme.com