# Venmo: What happens in Vegas, stays in Venmo

Venmo is a peer-to-peer (P2P) mobile payment app owned by PayPal. The Venmo app allows its users to exchange money with just a click of a button. In the fourth quarter of 2019, Venmo's net payment volume amounted to 29 billion U.S. dollars, representing a 56 percent year-on-year growth, and its user-base had more than 40 million active accounts[1]. What made Venmo so popular in the US is its social flavor; users are required to accompany their transactions with a message describing what the transaction was about. This social twist has allowed Venmo to transform financial transactions into sharing experiences.

In this assignment, you are given a sample of Venmo's dataset, and you are called to answer the following questions. All of your code should be done in Spark. For the social network questions, you are allowed to use the package networkx from Python or write your own UDFs in PySpark. For your plots, feel free to use Python or R.

**Disclaimer: Do not distribute this homework and/or the Venmo dataset to anyone outside this class.**

## Text Analytics

Venmo data offers a unique opportunity to see what people are spending their money on. However, the most challenging part in figuring out such a thing is to have the right "lexicon" (most of them are proprietary and for a reason, as you will hopefully realise). Venmo messages can be 1) emoji, 2) text and 3) a combination of emoji and text.

For emoji, I am attaching here an emoji classification dictionary that includes all emoji and their categories (I scraped that through emojipedia.com). This should make your life very easy when classifying messages with emoji.

Text, however, is harder to classify. I have created a word classification dictionary, which you can find here: https://docs.google.com/spreadsheets/d/110FGFJpel_5tpa-UUVXviZ7IMv5akPxOiXUSscdccYQ/edit?usp=sharing. In this dictionary, words are divided into 9 topics: 1) People (usually, words that indicate emotions towards others), 2) Food, 3) Event, 4) Activity, 5) Travel (note that Travel is a subset of Activity, but in my opinion required to be a different topic), 6) Transportation, 7) Utility, 8) Cash and 9) Illegal/Sarcasm (as you will notice, there are many bad words in there, and it's hard to distinguish if they are used as sarcasm or if they are used for illegal activities. *My German colleague was banned from Venmo for one of them, but that's another story...*)

[1] https://www.statista.com/statistics/763617/venmo-total-payment-volume/

**Q0 [5 pts]**: Your first task is to open your Venmo app, find 10 words that are not already in the dictionary and add them to it. Make sure you don't add to the dictionary a duplicate word by hitting Control+F before adding your word.

**Q1 [2 pts]**: Use the text dictionary and the emoji dictionary to classify Venmo's transactions in your sample dataset.

**Q2 [3 pts]**: What is the percent of emoji only transactions? Which are the top 5 most popular emoji? Which are the top three most popular emoji categories?

Q1 and Q2 allow you to get an aggregate view of Venmo's users transaction profile. Let's now create individual spending behavior profiles.

**Q3 [2 pts]**: For each user, create a variable to indicate their spending behavior profile. For example, if a user has made 10 transactions, where 5 of them are food and the other 5 are activity, then the user's spending profile will be 50% food and 50% activity.

**Q4 [3 pts]**: In the previous question, you got a static spending profile. However, life and social networks are evolving over time. Therefore, let's explore how a user's spending profile is evolving over her lifetime in Venmo. First of all, you need to analyze a user's transactions in monthly intervals, starting from 0 (indicating **their first transaction only**) up to 12.

For example, assume a user's first transaction was a pizza emoji. Then, her user profile at 0 would be 100% food. Now, by the end of her first month in Venmo, she has transacted 4 times, 2 of them are food and 2 are activity related. Her speding profile in 1 month is 50% food and 50% activity. Following this logic, you need to create a user's profile up to 12 months (**Hint**: You can use window functions to do this).

If you do this right, you will create a dynamic spending profile for each user. However, this is meaningless to plot. Let's plot instead the spending profile of the average user. To do this, for each time point, you need to compute the average and standard deviation of each spending category across all users. Therefore, in your y-axis, you will have time in months (from 0 up to 12). In your x-axis, for each time point, plot the average for each category surrounded by its confidence interval (+-2 * standard deviation). **What do you observe? Does the spending profile of the average customer stabilize after some point in time?**

# Social Network Analytics

Let's now look at a user's social network.

**Q5 [5 pts]**: Write a script to find a user's friends and friends of friends (**Friend definition:** A user's friend is someone who has transacted with the user**,** either sending money to the user or receiving money from the user). **Describe your algorithm and calculate its computational complexity. Can you do it better?**

**Q6 [10 pts]**: Now, that you have the list of each user's friends and friends of friends, you are in position to calculate many social network variables. Use the dynamic analysis from before, and calculate the following social network metrics across a user's lifetime in Venmo (from 0 up to 12 months).

i) Number of friends and number of friends of friends [very easy, 2pts].

ii) Clustering coefficient of a user's network [easy, 3 pts]. (**Hint**: the easiest way to calculate this is to program it yourselves. Alternatively, you can use "networkx" python package. The latter approach will slow down your script significantly).

iii) Calculate the page rank of each user (hard, 5 pts). (**Hint**: First of all, you need to use GraphX to do this. Moreover, notice that page rank is a **global** social network metric. If you go ahead and calculate the page rank for each user at each of her lifetime points, you will soon realize it will be a dead end. **Can you think of a smart way to do this?**)

# Predictive Analytics with MLlib

If you have survived this assignment so far, well done! Now, let's put all this work into a problem that every company that deals with customers wishes to solve. One of the biggest questions in Customer Relationship Management (CRM) is whether you can predict in advance how many times your customers will transact. In this subsection of the homework, we will investigate how the different set of metrics that you have created above (text and social) can help us predict the total number of transactions a user will have by the end of their first year in Venmo.

**Q7 [1 pt]**: First, create your dependent variable **Y**, i.e. the total number of transactions at lifetime point 12. In other words, for every user, you need to count how many transactions s/he had committed during her/his twelve months in Venmo.

**Q8 [2 pts]**: Create the recency and frequency variables. In CRM, this predictive framework is known as **RFM**. Here, you don't have monetary amounts, so we will focus on just **RF**. Recency refers to the last time a user was active, and frequency is how often a user uses Venmo in a month. You need to compute these metrics across a user's lifetime in Venmo (from 0 up to 12).

For example, if a user has used Venmo twice during her first month in Venmo with the second time being on day x, then her recency in month 1 is "30-x" and her frequency is 2/30.

**Q9 [2 pts]**: For each user's lifetime point, regress recency and frequency on **Y. Plot the MSE for each lifetime point.** In other words, your x-axis will be lifetime in months (0-12), and your y-axis will be the MSE. (**Hint**: Don't forget to split your data into train and test sets).

**Q10 [5 pts]**: For each user's lifetime point, regress recency, frequency **AND her spending behavior profile** on **Y. Plot the MSE for each lifetime point like above. Did you get any improvement?**

**Q10 [5 pts]**: For each user's lifetime point, regress her social network metrics on **Y. Plot the MSE for each lifetime point like above. What do you observe? How do social network metrics compare with the RF framework? What are the most informative predictors?**

**Q11 [5 pts]**: For each user's lifetime point, regress her social network metrics **and the spending behavior of her social network** on **Y. Plot the MSE for each lifetime point like above. Does the spending behavior of her social network add any predictive benefit compared to Q10?**

# Bonus question: Matching [10pts]

If you have survived this far, you can take a deep breath and think if you want to stop or continue. If you decide to continue, let's revise question 10. In that question, you found that some social network metrics are strong predictors. For example, you might have found that users who joined communities (friends of friends) with high clustering coefficient tend to turn out high volume users for Venmo. However, this is just correlational evidence. We don't know if a user joined this highly connected community, and s/he became good because her/his community was very active or because the user was intrinsically good and ignited a lot of activity for her/his community. **Can you try to make these findings causal by using matching?**

For example, you can create functionally equivalent group of users by matching their communities on their social network characteristics and their transactional activity. More specific, assume you have users A and B. They both join (**at time =0**) communities with similar network characteristics. However, the transactional activity of user's A community is low, while the transactional activity of user's B community is high. Now, you are in position to figure out how user A and B affect the transactional activity of their community.

Similarly, you can match communities on their transactional activity levels. Assume users' A and B communities have similar transactional activity levels, but differ in their network characteristics. Then, you can figure out the effect of the community's social structure on the transactional activity of users A and B.

**Try to program your matching algorithm in Spark, and re-run your regressions. Summarize your findings.** (**Hint**: Note that your matching will take place at t=0, but your Y variable will still be at t=12.)