An Empirical Analysis of the Happiness Paradox - Your friends are happier than you are.

Neha Mundada
mundada@usc.edu
University of Southern California

Vinit Parakh
vparakh@usc.edu
University of Southern California

Abstract

The friendship paradox states that "your friends have more friends than you, on average." This paradoxical effect can be a result of the topology of network, how they are connected together. The numbers of friends people have are distributed in a way that follows a power law rather than an ordinary linear relationship. So most people have a few friends while a small number of people have lots of friends. Its this second small group that causes the paradox. It would be interesting to know if the paradox hold for human personality traits like happiness, sadness, income as well. This is not so clear because happiness and income are not directly represented in the topology of a friendship network. This paper is a study of the so called happiness paradox "Your friends are happier than you are". On social networking platforms like Facebook, Twitter, Instagram, etc there is this natural intuition that our friends are happier than we are. Is this really true, is the question to be asked. Using a sample dataset from twitter we confirm that the happiness paradox holds for more than 70% of Twitter users. A possible explanation for this can be, active users are more interested in consuming content, on average, tending to add more friends in the network. Considering the fact that people post more happy content than sad, such active users gets exposed to more happy content than sad getting the feel of their friends being happier than them.

Introduction

Notifications on social media of career success from an acquaintance who is in the same field as you, but younger definitely makes you sad if you are struggling. Similarly a relaxing picture on a nice beach of a classmate when you are dying to finish your submission makes you feel that he or she is luckier and happier than you are. This is a natural feeling that most of us experience in our day to day lives. The feeling that your friends are happier than you are based on social media updates. This is what is known as the happiness paradox. This paper particularly tries to empirically analyse if the paradox holds true on twitter data. Even though it is a debatable fact whether everything that is posted on social media denotes the true state of the user in real life or not, this paper tries to check this empirically. Psychologically people treat social media as red carpet and post content when they are happy than they are sad. Thus, active users always get a feeling that their friends are happier than they are by just seeing their happy posts or pictures. As per one of the Stanford researcher Jordan this arises as we wildly overestimate of other peoples happiness based on the images and accomplishments we see on their Social Media. With the constant comparisons, users tend to see themselves as the losers, as compared to their friends and become sad. As this is one of the increasing problems of social media we feel a need to confirm this phenomena. In present work we consider Twitter data to check if the generalised paradox holds or not. Twitter is a bidirectional network, where each user have friends and followers. In our analysis we consider only friends of a user, as what friends posts impact a user more than what follower posts. This is because, while scrolling through the news feed users only see posts of their friends and get affected with what they post. Its rare for a user to go the followers page and check what they have posted.
In this paper, we explain our approach as how to measure the happiness for each user and compare it with its friends, to check if the paradox holds.

Data Collection

We collected Twitter data which contains 81.9 million tweets. For all these tweets we also gathered the Twitter social network which included links between users. Our data has approximately 34,52,009 unique users. Since we need the user tweets for performing the sentiment analysis, we only consider users which have at least 10 tweets and have at least 5 friends. This graph is then used to test the happiness paradox on Twitter. We choose twitter as a platform for performing the hypothesis because of the following 2 reasons:

- The Twitter API is easily available to use and hence makes it the first choice to perform research.
- Also with the limit on the tweet text (140 characters) we feel that user express their emotions concisely and hence it serves as a good platform to do sentiment analysis.

Approach

We have come up with following approach after trying a couple of other approaches for sentiment analysis.

1. **Approach 1** One of approaches included directly classifying the tweet as happy, neutral or sad. However after going through the tools online we did not find good results and so decided to go ahead with the alternative, more sensible and feasible approach.
2. **Approach 2** The second approach that we followed has a psychology basis to it. We found out tools like Alchemy API and Stanford NLP which instead of classifying a tweet as happy, neutral or sad would classify it as positive, negative or neutral. Psychological evidence say that there is a positive correlation between positive person and happy person. We used this as ground truth for proving our hypothesis and thereby consider that the person who tweets more positive as a happy person.

The overall work flow of our system can be explained as follows:

Step 1 We picked random users as our seed users. We picked 20% of the total users randomly using mysql random selection.Twitter being a bidirectional social network each user has followers and friends. The followers are people who follow the user and friends are the people in the network who the user follows. In our case we have considered only the friends of the seed users and not their followers. This makes intuitive sense since a user only sees notifications of his/her friends on its news feed. It is very rare that a user would go to the followers page and look at what they have posted. So we can conclude that the users sentiment are majorily affected by what it sees on his/her newsfeed.
Step 2 We then monitor all the tweets of the seed users and its friends and thereby calculate the sentiment for every tweet and classify them as happy, neutral and sad.
Step 3 Once we have sentiments calculated for all the tweets, we move ahead to calculate what we call as the happiness coefficient for each user (seed user and its friends). This is a value associated between 0 to 1, with every user of the system that quantifies how happy the user is.
Step 4 Aggregate the results of the happiness coefficient to see if the paradox Your friends are happier than you are holds on the twitter social network. However to get a real sense of the network, we only consider the seed user while aggregating the values, since we have all their friends and we do not guarantee the same for the non-seed users.

Architecture

1. **MySQL Database** We used mysql as a persistent storage. This consists of tables like user, usergraph, tweets, retweets which were later use in the evaluation process.
2. **Happiness Coefficient Engine** This is the crux of the project which would calculate the happiness coefficient for the user. This engine performs 3 tasks which are stated as follows:

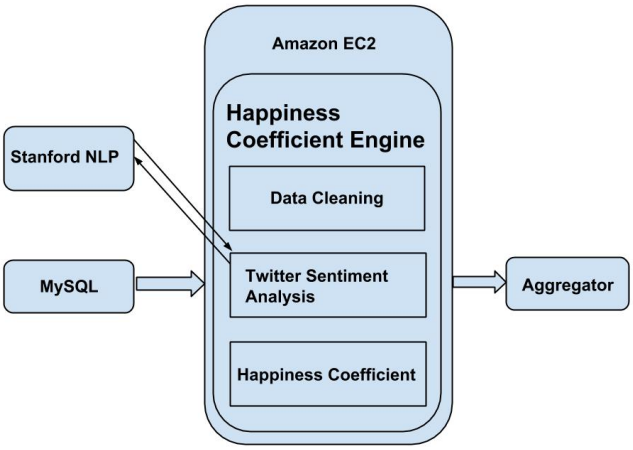

Figure 1: System Architecture

(a) **Data Cleaning** The process of data cleaning is an important part of the system architecture. Before passing the data to the sentiment analysis tool it is better to clean data and remove all the noise. We performed the following steps to make sure we removed the maximum amount of noise from the tweet in order to get accurate results.

i. **Emoticons Replace** Emoticons succinctly expresses the sentiment of the tweets and so we incorporated this it into our system. The sentiment analysis tool that we used (Stanford NLP), did not take into consideration the emoticons while determining the tweet sentiment. So we have adopted a simple but intelligent method to replace the emoticons so that they are considered when sentiments are calculated. We replaced the emoticons with the emotion that is expresses. We scraped a list of emoticons and the emotions is expresses from wikipedia. The list contains a list of all the emoticons along with their keyword mapping which we replaced in the tweet text by using regular expressions. This helped in increasing the accuracy by 5% which we think is significant. The following table shows an example of how we replaced emoticons.

:) :-) :D :-]	Smiley or happy face
:(:-(:'(:-c	Frown, Sad

ii. **Slang Replace** Slang are words and phrases that are regarded as very informal. Since these are not english dictionary words they add noise to the tweet text thereby decreasing the accuracy. Similar to emoticons we replace slang words using regular expressions. We found a list of frequently used slang words on twitter from syhex which served our purpose really well.

iii. **Stopwords Removal** The stop words like 'and',

lol	laughing out loud
bff	best friends forever

'when', 'which', etc were removed from the tweet text as a part of the data cleaning process by adding an annotator to the stanford nlp tool that given a tweet removes the stop words before calculating the sentiment of the tweet.

iv. **Punctuation Removal** Punctuations were also removed except for '!' because we found that the tool we are using for sentiment analysis gives weightage while calculating the sentiments to exclamation mark.

v. **Hashtag Removal** The hashtag symbol was removed from the tweet text to reduce the noise.

(b) **Sentiment Calculation** After cleaning the data the happiness engine then calculates the sentiment of every tweet using the Stanford NLP tool. The accuracy of the hypothesis entirely depends on the tool we use. Hence choosing the right tool was one of the most important step in the architecture workflow. We used stanford nlp because it is an easy to use, open source sentiment analysis tool. It gave us an accuracy of about 80% when we tested it manually by tagging about 500 tweets. The tool runs locally i.e. does not require network bandwidth and has no limits as compared to Alchemy API which gives an accuracy of about 86% but is not open source, has limits on the number of rest calls and is slow due to the network latency. The advantage of using Stanford NLP is that it is already trained on a dataset which fits our purpose really well, there by reducing the development time.

(c) **Calculate Happiness Coefficient** The happiness coefficient for all the user in the system is calculated.

3. **Aggregator** The aggregator finds the mean or median for all the users and then eventually determine if the paradox holds true.

4. **Amazon EC2 Instance** Due to the huge volume of data (tweets) we decided to use an amazon ec2 instance for the happiness coefficient engine. It took us about 3 days for the task to run which would calculate the sentiments for the tweets.

Evaluation

The happiness coefficient for a user U_i is calculated as follows

$$U_i = \frac{NumberofHappytweetsofUser_i}{TotalNumberoftweetsofUser_i}$$

The mean of the of the happiness coefficient of friends of i^{th} user is given by

$$F_i^{(mean)} = \frac{1}{M}\sum_{j=1}^{M} U_j$$

$$F_i^{median} = \forall_j Median(U_j)$$

j are all the friends of $User_i$
F_i is the mean/median of the happiness coefficient of friends of $User_i$

The accuracy(Happiness Paradox) that we got for our system is as follows:

Happiness Paradox (Mean)	73%
Happiness Paradox (Median)	79%

The accuracy(Sadness Paradox) that we got for our system is as follows:

Sadness Paradox (Mean)	23%
Sadness Paradox (Median)	20%

The number of tweets with emoticons is just 7% of the total tweets.

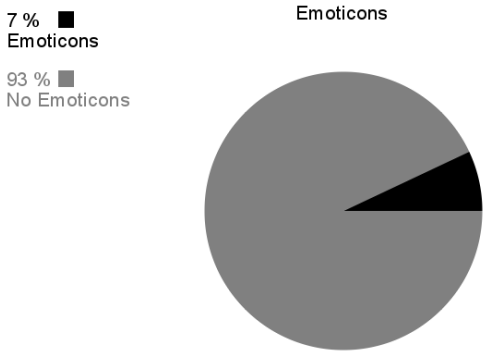

Figure 2: Emoticons Pie-Chart

The number of tweets with slang words is just 2% of the total tweets.

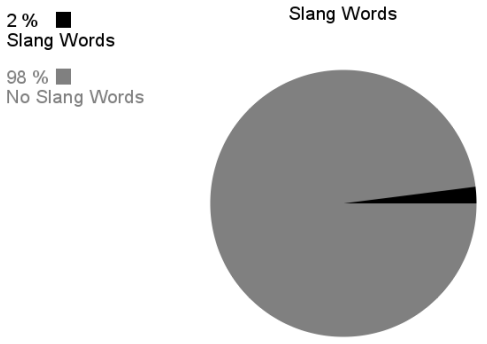

Figure 3: Slang Word Pie-Chart

The tweets are equally distributed along all the 3 sentiments.

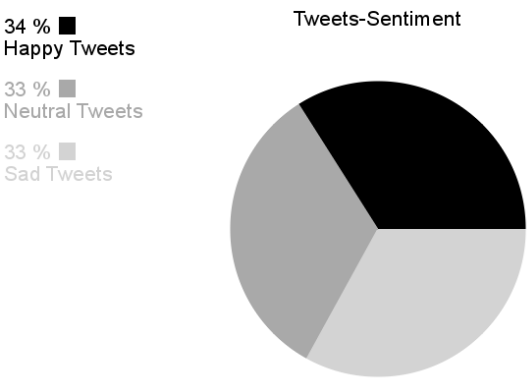

Figure 4: Tweet Sentiment

The evaluation involved considering the psychological point of view, which says that their is a positive corelation between positivity and happiness. The Stanford NLP tool that we are using gives us back values from a range of (very positive, positive, neutral, negative, very negative). We mapped each of these values to happy or sad depending on the table.

very positive, positive	happy
neutral	neutral
very negative, negative	sad

RelatedWork

There has been work going on, on parallel topic like the friendship paradox and other sentiment analysis on twitter, but we feel this is the first of its kind experiment to test this on twitter social network empirically. Researchers from Finland are trying to prove a more generalised version of the friendship paradox, but they are trying to do this mathematically using network properties. They are trying to corelate the network properties with the human attributes like happy, income, wealth. Our approach is to actually measure the values, so as to see if the paradox holds in real on the twitter social network. We believe that it is hard to find a corelation between human trends and network properties. This is because human trends vary from person to person and depends on a lot of external factors which cannot be visible in a network. This is study is a way of getting closer and see if our intution is true. Something that we go through daily.

Conclusion and Future Work

1. **Conclusion** The present work has demonstrated that the happiness paradox holds for approximately 73% of the active Twitter users when considering the mean and 79% of the active users when considering the median. Similarly we have also calculated the sadness paradox "your friends are more sad than you are" and found out that 23% users by mean and 20% users by median follow that the so called 'sadness' paradox. After comparing the 2 measures we feel that mean is biased towards the power law assumption in a social media network, while median gives us the true picture of the scenario. If you are an active Twitter user feeling unhappy because your friends seem to be doing better than you are, remember that almost everybody else on the network is in a similar position.
2. **Future Scope** We plan to consider tweets other than english, as of now we are ignoring them. Also to get better sentiment analysis we would like to consider more parameters like the tweet-time and url content. The tweet time can be used in a way considering the fact that people generally tend to put more positive content during the day than during the evening. We can model this to get better accuracy. Also as of now we have removed all the urls. In future we plan to consider the text from the url and use that to get the sentiment of the tweet. This would improve the accuracy of the system. We can also consider the communication between the friends while considering their sentiments.

References

[1] Nathan Oken Hodas, Farshad Kooti, and Kristina Lerman. Friendship paradox redux: Your friends are more interesting than you.

[2] Sonja Lyubomirsky and Laura King. 2005a). The benefits of frequent positive affect: Psychological Bulletin, 2005.

[3] Duyu Tang, Bing Qin, Ting Liu, and Qiuhui Shi. Emotion analysis platform on chinese microblog.

[4] Young-Ho Eom - University of Toulouse, France and Hang-Hyun Jo - Aalto University, Finland. Generalized friendship paradox in complex networks.