# Milestone1 Report

## — Tweet Lifecycle Analysis

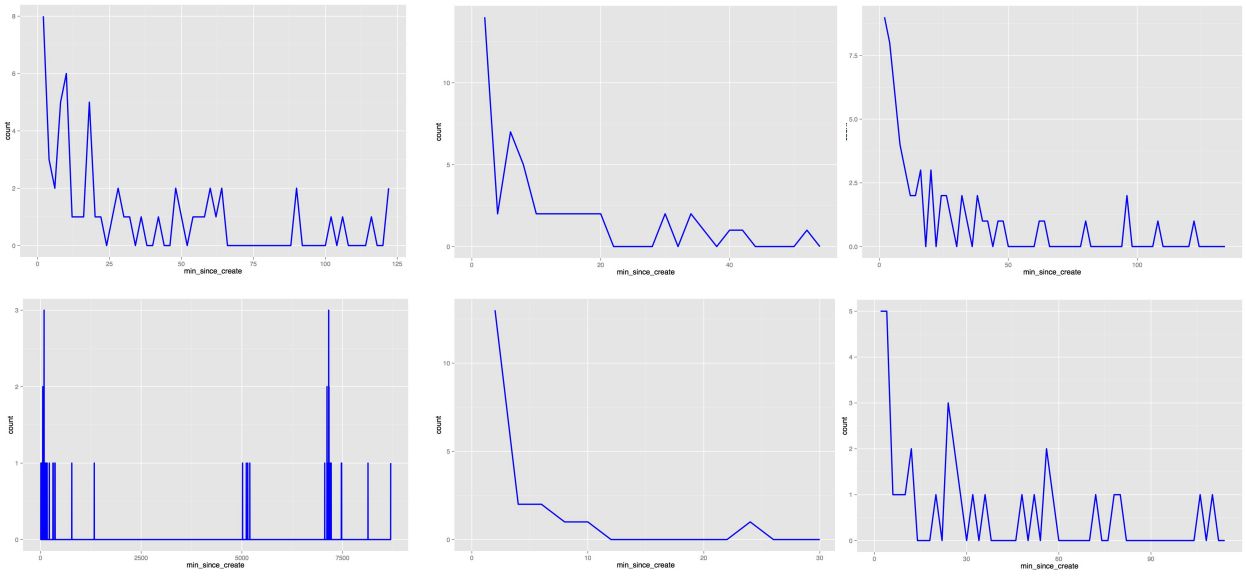
## Research Questions

We want to study the progress of a tweet got retweeted in it's life cycle, especially the timing of its life cycle events. Some example of such events could be tipping points of retweet count and when the tweet got highest retweets in the same time interval.

Knowing the timed events in a tweet's life cycle is interesting. It is also important in understanding how to effectively communicate ideas with audience. We can also find out what features play important roles in a tweet's life cycle.

## Preliminary Analysis

We pulled retweet data for several tweets from Twitter API. Then transformed such data into time intervals and the count of retweets in such intervals.




Within the six tweets we picked, their retweets count charts show a similar pattern, and minor differences. Most tweets reaches their highest retweets count immediately after it was created. Partially approved our hypothesis that a tweet has a life cycle. However, there are different patterns in our data where we will get a clear picture when we got more data.

For details about data collection and cleaning please refer to collection and cleaning section.

Created by Daiwei Lu and Yuan Gao

## Review of the State of the Art

1. *Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks* (by Kristina Lerman and Rumi Ghosh) Retrieved from: http://www.cabdirect.org/abstracts/20133041951.html
   This study focus on how the structure of the social network, e.g. network density, affects dynamics of information spread on social network.
2. *Want to be Retweet? Large Scale Analytics on Factors Impacting Retweet in Twitter Network* (by Bongwon Suh, Lichan Hong, Peter Pirolli and Ed H. Chi) Retrieved from: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5590452&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5590452
   This article try to find out what are the key factors behind the number of retweets. They found that URLs and hashtags have a strong relationship with retweetability.
3. *Predicting Information Spreading in Twitter* (by Tauhid R. Zaman, Ralf Herbrich, Jurgen van Gael and David Stern) Retrieved from: http://incc-tps.googlecode.com/svn/trunk/TPFinal/bibliografia/Zaman%20and%20Herbrich%20(2010).%20Predicting%20Information%20Spreading%20in%20Twitter.pdf
   These article presents a new methodology for predicting the spread of information on Twitter. They found the most important features for prediction is the tweeter and retweeter.
4. *The Pulse of News in Social Media: Forecasting Popularity (by Roja Bandari, Sitaram Asur, and Bernardo A. Huberman)* Retrieved from: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4646%26lt%3B/4963
   This paper studied how to use certain features of an article to predict the popularity of it and evaluate the efficacy of these features to serve as predictors of online popularity.

Those researches focuses on finding out which feature is more effective in the information spreading process, often focuses on the end results. However, information has a life cycle. Its spreading is a dynamic process. Our study focuses on the timeline of such process instead focusing on the end results.

## Dataset Collection

We didn't use any existing dataset. Instead, we directly accessed Twitter's API, and downloaded data from it.

## Dataset Analysis Cleansing

The data we got from Twitter's API is pretty clean, but not in a structure we want. We transformed the data by counting the number of retweets within a series of time intervals.

## Other Contributions

Initially we want to use hashtags to identify popular topics and study their spreading process. But as we experiment with Twitter's API, we found out it only allows us to go back about 7 days

in history, which is way less than a normal hashtag's life cycle. This is not enough for us to extract a complete dataset.

We then turned to single tweet, as it has a much shorter and identifiable life cycle. However, it turns out Twitter's API to retrieve retweets is limited to only 100 most recent retweets. This shouldn't be a problem if start collecting tweets in advance, since Twitter provide a streaming API using which we can track users and aggregate the data along the way. This way we can go beyond 100 limit and increase our dataset size.

Since we didn't plan in advance and didn't collect have dataset with tweets have a complete dataset contains more 100 retweets, we can only hand pick some tweets with less than 100 retweets and using Twitter's API to aggregate the data. We used a Ruby library called "twitter" (https://github.com/sferik/twitter) to interact with Twitter's API. After data collection with used Ruby to transform and summarized our data for preliminary analysis.

## Specify what each member has done for each section

We discussed every step through the process. Each of us is involved in every section.