Experts v.s The Crowd: Examining News Prediction and User Behavior on Twitter

Chu-Cheng Hsieh
University of California, Los Angeles
4732 Bolter Hall
Los Angeles, CA 90095, USA
chucheng@ucla.edu

Christopher Moghbel
University of California, Los Angeles
4732 Bolter Hall
Los Angeles, CA 90095, USA
cmoghbel@ucla.edu

Jianhong Fang
University of California, Los Angeles
4732 Bolter Hall
Los Angeles, CA 90095, USA
sjtufjh@ucla.edu

Junghoo Cho
University of California, Los Angeles
4732 Bolter Hall
Los Angeles, CA 90095, USA
cho@cs.ucla.edu

ABSTRACT

{Change Log:
version 3.0 – clean up based on 2.6.1 rv1 until the end of section 3
version 3.1 – until the end of section 4 }

Things want Dr. Cho to help:
- Comments on 1 4 (P.7)
(not proof reading yet): discussion(sec. 5) & conclusion(sec. 7)
- Do you think that I may miss anything in the discussion or conclusion?
- We are running out of 10 pages, do you think we should add the missing 4.5 (hybrid) to the paper

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

Efficient market, Twitter, Social networks

1. INTRODUCTION

Life in human society relies on collective decisions – from a juries' verdict to lawmaking. Often people believe two (or more) heads are better than one when making a decision that demands deliberation because independent judgments may suffer from bias or limited information. For example, doctors [13] discovered that a group of inexperienced physicians may make better decisions than one experienced physician because they can take advantage of their complementary abilities and avoid extreme estimates. Political scientists [15] believe that democratic forms of governance benefit from the wisdom of the populace. What is more, computer scientist [8, 14] show that the crowd wisdom can be used to design recommender systems, namely, folksonomies.

In this study, we further investigate the problem of crowd wisdom – we attempt to discover whether it is possible to pick a group of experts who consistently make better decision than the crowd. If such an argument holds, one may take advantage of this fact, as polling a small group of experts can be accomplished much easier than polling the population. For example, passing legislation through a parliamentary system is often faster and easier than having the populace vote directly [3]. However, whether experts exist itself is a long debatable question: "Efficient Market Hypothesis(EMH)" [5], a famous hypothesis proposed in the 1960s in the finance domain, draw a debated argument that no expert can consistently outperform the market with regards to making stock investments in an informationally efficient market. Our work provides a direction for harnessing social wisdom, and helps researchers understand user behaviors as well as their power on social sites.

Forming large groups to conduct experiments has been expensive or infeasible. Now with the help of social websites, we are able to compare experts' wisdom against the crowd's in a domain where the qualities of decisions can be verified. In particular, we employ Twitter as a polling source, form expert groups based on varying aspects, and evaluate their performance in predicting popular news withing a given time constraint. Twitter is the largest social, micro-blogging website, allowing users to publish text-based messages consisting of at most 140 characters, called tweets. According to on-

SAC'12 March 25-29, 2012, Riva del Garda, Italy.

line traffic report[1], 27.5 million (unique) users visited the Twitter website in May 2011. We may take advantage of Twitter to conduct polling because among these visits, news reading/sharing is one of the most important activities. Previous research by Java et al. [9] shows that, except chatting, sharing URLs and spreading news are two main purposes of using Twitter.

We conducted experiments on 2.83 million tweets collected from September 1st to December 31st in 2011. All of them contains a link pointing to any page belonging to *The New York Times* website[2]. We developed expert selection strategies based on various aspects, and then compared wisdom of these expert groups against wisdom of the crowd. We summarize some of the main findings as follows:

1. News on Twitter:
   - Most news pages attract only short attention from the public. Our result shows that the public loses interest in 85% news stories no matter whether the stories are popular ones or not.
   - 42% of twitter users read tweets on a web browser, but mobile devices are getting popular. Moreover, about 29% of users publish a tweet via auto-posting service.
   - Users with strong influence, namely lots of followers, bias the numbers of tweets about news stories these users tweeted.
2. Polling expert groups:
   - We propose three methods for selecting experts, and out of the three, the confidence-interval strategy performs the best (Section 4.4).
   - Polling from experts does not outperform polling from the public, but expert wisdom can be used to improve the results from the public (Section 4.5).

The remainder of this paper is organized as follows. Section 2 explains how we define news tweets and why we see tweets as votes. Section 3 shows our experiment design and how we collect data, and Section 4 discusses how experts are selected. Section 5 studies the characteristics and interesting findings from analyzing news data, followed by related research and our conclusion.

2. POLLING ON TWITTER

To keep our discussion concrete, we formulate our problem in the context of Twitter, though extending polling notion on other social-network sites is straightforward. In this section, after presenting several definitions, we discuss user recommendation activities we observed on Twitter, and then introduce the targeted problem.

2.1 Definition Of Terms

Tweets are messages published on Twitter. When a tweet (denoted as x_i) is created, it often embeds some URL linking to a web page for disclosing more details about the message. Embedding an URL is a very popular practice because Twitter has a length constraint on Tweets. After an tweet x_i is published, one may now get notified of the tweet x_i if he/she follows the creator of the tweet x_i. In other words, the creator broadcast a message to his/her followers.

If a follower reads the tweet and decide to promote it, the user may broadcast the tweet to his/her followers by retweeting it. With regards to the retweeting activity, conceptually one may consider another tweet x_j now being created, but the content is identical to the tweet x_i.

Definition 1. **Origin tweet**: We name a tweet **an origin tweet** if it is not a retweet. An origin tweet serves to bring outside (original) content into Twitter, while a retweet serves to spread a message within Twitter.

Definition 2. **News tweet**: If a tweet contains a link that points to a news website, for instance, *http://www.nytime.com*, we say the tweet is a news tweet. Here we have a broad definition of news tweets, including tweets pointing to reader letters, opinions, and other news-related articles.

Definition 3. **News thread**: Many different URLs may lead to the same article. Thus, we define a news thread as the collection of all tweets with links pointing to the same content. Furthermore, if two or more pages in a website share one `<TITLE>` meta tag in `<HEAD>` section, we view them as identical ones (from the perspective of content) in that the difference may come from embedding some advertising snippet(s) or from some minor modifications.

Definition 4. **Expert group**: Experts are a group of users that are selected based on our proposed strategies. Users are sorted by some ranking algorithm and top G% of users are then selected to from an expert group.

Definition 5. **Market**: Following a similar notion mentioned in efficient market hypothesis [5], here we define the market decision as the decision made by the all users within Twitter. {Chris believe the crowd is a better word.}

Definition 6. **Popular News Prediction**: Given a time constraint $t_\triangle$, the task is to find top V% popular news articles n_i within the time constraint, where the popularity of a news article is determined by the total number of tweets associated with the news thread of the article with no time constraint.

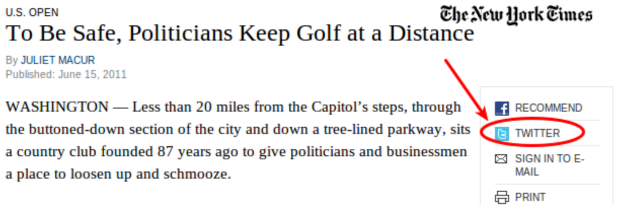
2.2 (Re)Tweeting is a Recommendation

We believe that both tweeting activities and retweeting activities can be viewed as making recommendations. Here we explain our observations on user behaviors on Twitter, and therefore just why drawing this analogy is reasonable.

2.2.1 Tweeting

Tweeting a web page means to create a tweet for broadcasting the page to the user's followers on Twitter. Supposedly,

[1] http://siteanalytics.compete.com/twitter.com/
[2] www.nytimes.com



U.S. OPEN

The New York Times

To Be Safe, Politicians Keep Golf at a Distance

By JULIET MACUR
Published: June 15, 2011

WASHINGTON — Less than 20 miles from the Capitol's steps, through the buttoned-down section of the city and down a tree-lined parkway, sits a country club founded 87 years ago to give politicians and businessmen a place to loosen up and schmooze.

Figure 1: A twitter button on a news page of *The New York Times* website

the creator intends to broadcast the page to all his/her followers on Twitter. Previous study {cite here} shows that people tend to tweet what they care, we can make a conjecture that one's tweets reflect his/her recommendations.

Recently, many websites embed a `TWITTER` button along with every page so that a user's clicking on the button, as shown in Fig 1, the site automatically creates and publishes the (news) tweet though an Twitter API(Application Programming Interface) for its reader. For instance, the TWITTER button offer by *The New York Times* create a tweet composed of the title of the news page and a (short) URL pointing to it. Given the fact that the TWITTER button is arranged next to "RECOMMEND on Facebook" and "email to friends", it is rational to assume that the creation of an origin tweet is a recommendation.

2.2.2 *Retweeting*

By default, a user on Twitter views tweets posted by persons he follows. When the user see an interesting tweet posted by someone else, retweeting forwards a tweet to the user's followers. Since tweets created by retweeting intends for creating nothing new, but only broadcasts one tweet to different audience, retweeting can be reasonably viewed as a recommendation if we embrace the notion that one broadcasts something if he/she recommends it.

In the past, to retweet another tweet someone must firstly copy-paste the contents of the original tweet, add a prefix "RT" (standing for retweeting) followed by `@<username>` (acting as a citation). For example, seeing a tweet starting with `RT: @john ...` says that this tweet is actually a retweet, and it originates from reading (and then forwarding) tweets from the user `john`.

Nowadays, twitter provides a **Retweet** button next to every tweet, and it is now an official mechanism provided by twitter. However, when a user's clicking on the button, Twitter simply broadcasts and add a retweet decoration on top left corner but not adding prefix (RT) anymore.

2.3 Wisdom comparison

In the essence of the efficient market hypothesis , expert cannot outperform the market with regards to making prediction of the future. The hypothesis says that "*one cannot consistently achieve returns in excess of average market returns on a risk-adjusted basis, given the information available at the time the investment is made*" – from Wikipedia.

To reason by analogy, our work examine whether or not a group of experts can consistently achieve better performance in terms of predicting top quality news in the future. We set a time constraint and collection recommendations from different group of experts. We aim to see whether polling experts a better idea than polling everyone (market).

Though the problem is different, the spirit is similar. If experts make better predictions, we should find these experts and monitor them to save the cost of polling the entire Twitter. Later in our experiments, we will show that very likely a similar hypothesis exists on Twitter, because the market wisdom is superior than expert wisdoms forming from many aspects.

In our experiments, we say one wisdom is better than the other if the wisdom could more accurately predict popular news pages in the future. Note that the popularity of a news pages is measuring by the number of tweets with a link pointing to the page. We will discuss how we define popular news in Section 3.2.

3. EXPERIMENT SETUP

We now describe our experiment setup. Starting with data downloading and cleaning, followed by how we identify popular news, and about how we evaluate the qualify of wisdom.

3.1 Downloading Data

We limit our experiments to news data on Twitter because downloading all tweets is beyond the capacity of our system and is also prohibited by Twitter. As a result, we set up a keyword filter, extracting tweets containing (all) the following keywords: "**http nyti ms**". Note that by default, all URLs pointing to *The New York Times* website are converted to short URLs prefixing with `http://ntyi.ms/`.

We collected news tweets for six months, starting from August 1st, 2011 to January 31th, 2012. We collected 4,234,899 raw tweets in total. We use the streaming API provided by Twitter together with the filtering keywords to download data. The API allows us to get near-realtime access to public and protected Twitter data.

3.1.1 *Data Cleaning*

Step 1: *URL prefix check*
The simple keyword filter brings some noise data; namely, tweets with content happening to satisfy the keyword filter. Thus, we purify data by applying a prefix pattern matching process that removes any tweet not containing a URL with the prefix.

Step 2: *Fixing Incomplete Data*
We decided to further purify our data so that we keep only tweets of news we can study for their entire tweeting activities. Therefore, we assign 2011 August as the warming-up period. Namely, if news pages are discussed in the warming-up period, we exclude all tweets about these pages so that we will not miss the early stage tweeting activities of any news thread.

In a similar fashion, we assign January 2012 as a cooling-down period. In this period we exclude tweets pointing to untracked news page, namely pages we had not seen before December 31st, 2011. Therefore, for every news page we were tracking, we ensure that we keep at least 30 days activity after the first mention of a page.

3.1.2 Training Set & Testing Set

Our data contains tweets about news pages where the first tweet of each news thread is are reported between September 1st and December 31st. Our study shows that after data cleaning, from September 1st to December 31st in 2011 we have 2,837,026 tweets originated from 18838 distinct news threads published on *The New York Times* website.

To avoid over-fitting, when evaluating the performance of polling experts, we divide the dataset into a training set and a testing set. If the first tweet of a news thread published between September 1st and October 30th, tweets about the thread is assigned to the training set. On the contrary, if the first tweet of a news thread published between November 1st and December 31st, tweets about the thread are assigned to the testing set.

3.2 Popular News

After sorting news threads based on numbers of tweet in each thread, the top $T\%$ are defined as popular news. Later in our experiments, we try setting $T\%$ to 2%, 5%, and 10%. Table 1 shows the top five news pages in our testing set. Though we only show the top five, it is worth mentioning that even in the top 100 news, very few stories appeared as headline news on *The New York Times* website. Some of them even did not appear in *The New York Times* portal page.

3.3 Examining the Source of Tweets

Our dataset contains 2,837,026 tweets, and among these tweets, 76% are origin tweets. Table 2 shows the break down of the source devices used to publish tweets.

About 30% of origin tweets are created from `twitterfeed.com`, a service that automatically publishes a tweet on Twitter when a blogger publishes an article in his/her blog. The second position is “Tweet Button” as we explained in Section 2.2.1. It is worth mentioning that a great portion, 13.42%, of users still favor creating tweets on Twitter through a browser.

{I rewrite this paragraph to make my point more clear:} We see a difference of device usage between origin tweets and retweets in Table 2. A possible speculation is that blogger service are popular and the writer by nature tends to post tweet on Twitter to engage with more audience. In such a case, when a user creates an origin tweet, it is possible that not only he/she believe something is valuable to his/her followers, but also that he/she would like to promote the article he/she wrote. On the other hand, people tend to retweet if the news are valuable to their followers.

As shown in Figure 2, if we sort by number of origin tweets, `tweeterfeed` would rank in the top spot. But most tweets from `tweeterfeed` are not related to popular news. On the contrary, when considering origin tweets, the source “Tweet Button” becomes very insightful. We observed that not all sources are equally important in terms of identifying popular tweets (or experts).

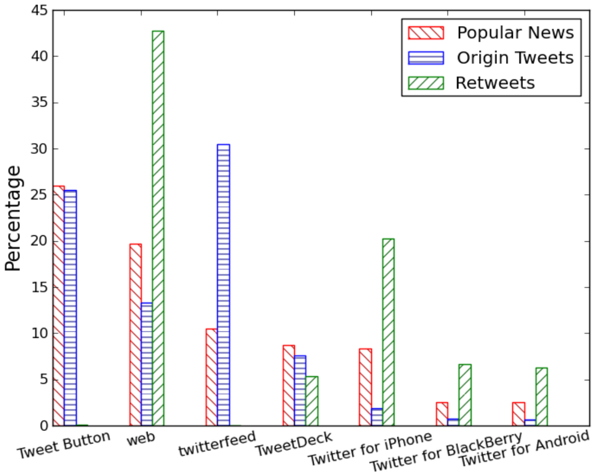


Figure 2: Breaking Down Source for All Tweets

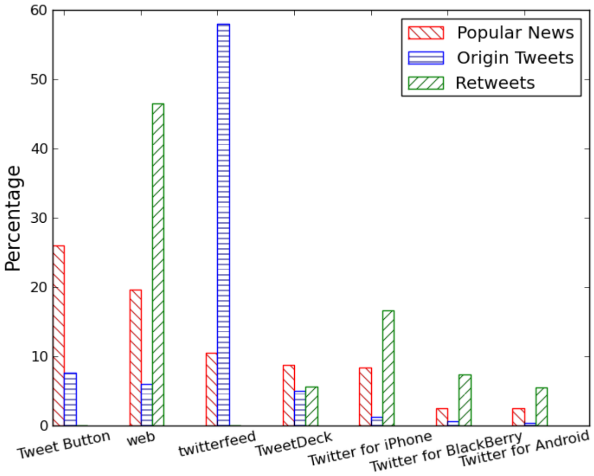


Figure 3: Breaking Down Source for First Hour Tweet

Retweets tend to come from different devices than origin tweets. Firstly, from Table 2 we may conclude that 42% of users read Twitter in front of computers. But in Figure 2, we see mobile devices are getting more and more popular; in particular, mobile users are very active retweeters. Intuitively, we may see these tweets come from real readers, when compared to `tweeterfeed`.

{Figure 3 shows the same graph with the first hour data.}

{future work: use only retweet to pick experts and show the graph}

3.4 Evaluation – Precision Recall Curve

Following [6, 7], we evaluate news recommendation results based on precision and recall. Precision refers to the fraction of recommended news pages existing in the popular news collection, and recall refers to the fraction of popular

Table 1: Top five news pages in the testing set

Title	Category	Short URL	# of tweets
Penn State Said to Be Planning Paterno Exit Amid Scandal	Sports	http://nyti.ms/vC9Ccg	6535
The Joy of Quiet	Opinions	http://nyti.ms/tZr7p2	4933
Google's Lab of Wildest Dreams	Technology	http://nyti.ms/uj0cpu	3286
Touch Of Evil	Gallery	http://nyti.ms/udI0wJ	3218
A Dispute Over Who Owns a Twitter Account Goes to Court	Technology	http://nyti.ms/uEWPRD	2989

Table 2: Sources of Tweets

Source	Origin(%)	Source	Retweet(%)
twitterfeed	28.96	web	**42.38** ⇑
Tweet Button	24.55	Twitter for iPhone	**20.25** ⇑
web	13.42	Twitter for BlackBerry	**6.31** ⇑
TweetDeck	7.40	Twitter for Android	**6.29** ⇑
Twitter for iPhone	3.33	TweetDeck	5.65
The New York Times	2.05	NYTimes for iPad	**3.14** ⇑
HootSuite	1.82	Mobile Web	**2.88** ⇑
NYTimes on iOS	1.68	Echofon	**2.32** ⇑
ifttt	1.38	Twitter for Mac	**1.19** ⇑
NYTimes for iPad	1.34	TweetCaster for Android	**1.04** ⇑

news collection that are retrieved, as shown in the following equations. Note that $|News\ recommended|$ is the only parameters that grows gradually to draw the curves, shown in Section 5.3.

$$precision = \frac{|News\ recommended\ are\ popular\ news|}{|News\ recommended|} \quad (1)$$

$$recall = \frac{|News\ recommended\ are\ popular\ news|}{|Popular\ news|} \quad (2)$$

We evaluate news recommendation made by different groups of users. In the next section we will explain how these groups are formed. A special group, named "market group", refers to conduct polling on the entire population. Apart from the market group, all groups are formed by the same number of users to guarantee fairness.

4. EXPERT SELECTION

In an attempt to capture the notion of an expert from different perspective, in this section we elucidate our three proposed strategies and one hybrid approach for selecting experts. We start with introducing characteristics of experts, following by the proposed strategies.

4.1 Characteristics of Experts

{to chris: I mostly rewrite the entire section 4.1. BTW, do you like this title?}

4.1.1 Accuracy

We would like to identify those who accurately tweet popular news at all times. We suggest this characteristic, accuracy, to be represented by the precision of tweeting popular news bases on someone's tweeting history. That is to say, we want an expert to be a user who keeps showing what he/she likes is what majority users on Twitter love.

4.1.2 Activeness

We would like to identify those who actively participate in news sharing activities. Therefore we measure someone's activeness based on how frequent he/she posts a news tweet on average. Accounts like *nytimes, breakingnews, latimes, ...* are representative examples because those accounts share one main purpose: publishing news tweets. Without considering activeness, we may end up with picking someone with a perfect accuracy but his/her voice rarely contributes to the prediction when conduct polling.

We notice a long-tail graph on user activeness. Figure 4 shows that only 3% of users in our dataset tweet more than once a week, even though the top 1% of users are tweet at least once a day. Most users are rarely interested in tweeting news because their accounts are not dedicated to sharing news. Conceptually, setting up a high bar of activeness favors picking experts from those who dedicate to news sharing, while a low bar welcome voices from folks.

4.1.3 Promptness

We would like to identify those who promptly report breaking popular news articles, but not those who report popular news in the past. Otherwise, a spammer could easily impersonate an expert by tweeting well-known popular news. As shown in Figure 5, we may categorize a spammer into accurate users but they should not be experts. In addition, news value decreases over time in general. Our experiment results (Section 5.1) shows that most news page are time sensitive because people quickly lose their interests in a news story once it has been widely disseminated.

In practice, when picking experts by reviewing their activities in the past, we impose a promptness threshold. We consider only tweets posted within the threshold are impactful, for example, within four hours after seeing the first tweet of a news thread.

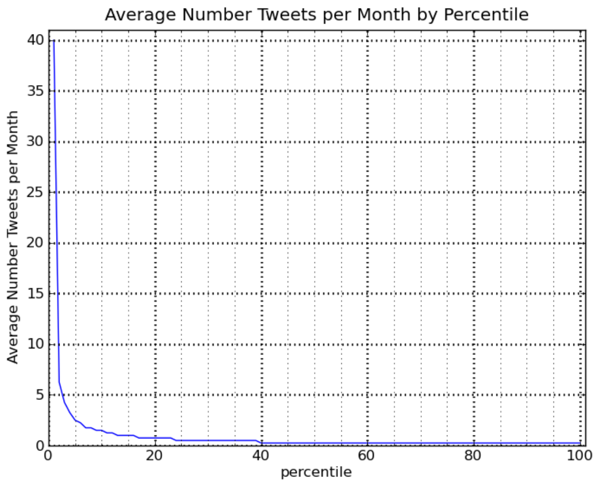


Figure 4: Average Numbers of Tweets per Month

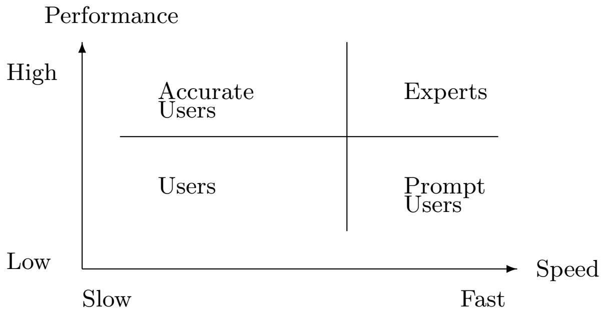


Figure 5: User Types

4.2 Precision-based Model

Precision is a quantitative way for describing accuracy. It is a popular measure for evaluating performance based on correctness. Basically, it computes the fraction of outcomes for which the desired result is returned. We believe high precision of tweeting popular news pages by a user suggest good potential for being an expert.

Equation 3 shows the formal definition of precision, a metric based on hit numbers and miss numbers, where we view a hit as recommending a story in the popular news; otherwise, a miss.

$$precision = \frac{\#\ of\ hit}{\#\ of\ recommendations\ (hit + miss)} \tag{3}$$

In this model, a user's performance in recommending good news is mainly dominated by precision. In its spirit, users with 100% precision are the best candidates in this model.

In practice, we have to add two setups to implement the model:

1. So many users are inactive and thus we cannot completely neglect the influence from activeness. We deice to consider a user as an candidate only when the activeness of the user is in the top 25% of users when ranked by activeness. This corresponds roughly to tweeting one per week. As a result, most inactive user are excluded because we believe their contribution to voting are too limited to be useful.

2. Having a tie is possible, because there exists a large amount of users on Twitter, and a user simply gets 100% precision as long as he/she make no miss in the training set. To break a tie, we make the user with more recommendations prevail.

4.3 F-score Model

F-score is another quantitative way for evaluating the potential of an user to be an expert. It does not only consider precision but also recall, a measurement computing the fraction of popular news tweets that are retrieved. Equation 4 shows the formula for computing recall. A high recall means that we miss only a few good news; on the contrary, a high precision means that most news we report are actually good.

$$recall = \frac{\#\ of\ hit}{\#\ of\ popular\ news} \tag{4}$$

F-score put into consideration both precision(correctness) and recall(retrieval). Equation 5 shows the general formula of F-score, where β is a tuning parameter, a parameter of adjusting the weighting of combining recall and precision. For example, $\beta = 0.5$ implies that precision is more importance than recall, and $\beta = 0$ makes F-score ranking be the same as the precision-based ranking.

Table 3: Confidence Interval Chart

Hit	Total	Low	High	Middle	Rank
1	1	0.224	1.000	0.612	4th
9	10	0.574	1.000	0.787	3rd
90	100	0.824	0.947	0.886	2nd
900	1000	0.880	0.917	0.899	1st

$$F_\beta = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (5)$$

In essence, here we want to consider user's possible contributions in the future. For example, it is arguable to say a user of 99% precision from 100 news tweets is inferior to another user of 100% precision from 10 news tweet, despite the former is with higher precision, because news tweets by the later cover much more news pages than by the former. This model suggests that the user in the later case may contribute more in terms of news recommendation.

4.4 Confidence-interval Model

Confidence intervals are used to tell the reliability of an estimate in statistics. From a different angle, picking an expert can be viewed as a process of identifying reliable prophets. Precision alone is not reliable because a precision of 90% may come from 90 out 100 or 9 out of 10, where the former is more reliable.

Our proposed ranking selects experts based on a similar philosophy to the confidence interval concept mentioned above. If we view a short-term user tweeting history (training data) as a sampling from the population (the user's forthcoming tweets in the future), confidence intervals are indicators of how well a user will perform in terms of tweeting good news articles in the future.

Here we use Table 3 to explain the notion of confidence intervals. Table 3 shows the low and high boundary of 95% confidence interval. Referring to the second row, if a user tweeted 9 popular news tweets out of 10 tweets in his/her history, the statistic tell us that 95% chance the user's "true" average precision in the future is between 57.5%(a low boundary) and 100%(a high boundary). As a result, in the first row a precision 100% coming from one hit on popular tweet out of one tweet is essentially a worse idea than picking a precision 90% coming from 9 hits out of ten.

Intuitively speaking, here we care about not only the precision we observed in a period, but also the reliability of the precision, determined by the number of tweets in a time window. In essence, the approach downgrades the score of non-active users, because we should not trust their precisions, which might be caused by, from the perspective of statistics, sampling errors.

In practice, we implement our model with the following settings:

1. We suggest to compare two confidence intervals through computing the middle points of the intervals, namely, the average of a low boundary and a high boundary. We say a user with a confidence interval 0.8 ∼ 0.9 (average 0.85) is better than another user with the confidence interval 0.6 ∼ 1.0 (average 0.8). If there is a tie, we use the low boundaries to break it.

2. We use the adjusted Wald method [1], proposed by Agresti and Coull, to compute the confidence interval. Their work shows that the adjusted Wald method provides better coverage of 95% confidence interval when sampling size is smaller than about 150. We present the confidence interval formula in Equation 6.

$$\begin{aligned} p' &= \frac{\#\ of\ hit + 2}{\#\ of\ trials(= hit + miss) + 4} \\ low &= max(0,\ p' - 1.96\sqrt{\frac{p'(1-p')}{\#\ of\ trials + 4}}) \\ high &= min(1.0,\ p' + 1.96\sqrt{\frac{p'(1-p')}{\#\ of\ trials + 4}}) \end{aligned} \quad (6)$$

4.5 The Hybrid Approach

{Two type: super experts and expert+market, add later}

5. DISCUSSIONS

{give an introductory graph here}

5.1 Lifespan of News Threads

We start our discussion by studying the "longevity" of a news threads. The notion of longevity aims to define the period that people are still interested in seeing a news page. Knowing longevities help us to not only know the fading speed of news value, but also help us to make a decision how fast we should recommend a news to public, namely, how long we should accept votes from polling groups.

Ideally we may set the start (of a news thread) the time the page being created and the end the time none being interested in knowing the page because, for example, they have read the news or they just do not care of knowing news in the past. In Table 1, the first news page "Penn State Said to Be Planning Paterno Exit Amid Scandal" may have a shorter longevity than the second news page "The Joy of Quiet" because the later seems to be an opinion or an essay while the former seems to be a report of a news event. {draw a graph to compare these two?}

In practice, we cannot monitor all pages created in the site so that we define the start of a news thread as the time we see the first tweet linking to the page. Also, we set the end as the time we see 90% of tweets are reported. We assign a cutting point at 90% (but not the very last tweet) because the last few tweets can be very irrational. A user may retweet some tweet happens in few months ago while most tweets gather up in first few hours.

We illustrate our findings in Figure 6. The dash line (top popular news) refers to top 2% of news threads based on their popularity (Section 3.2). And the solid line refers to all tweets.

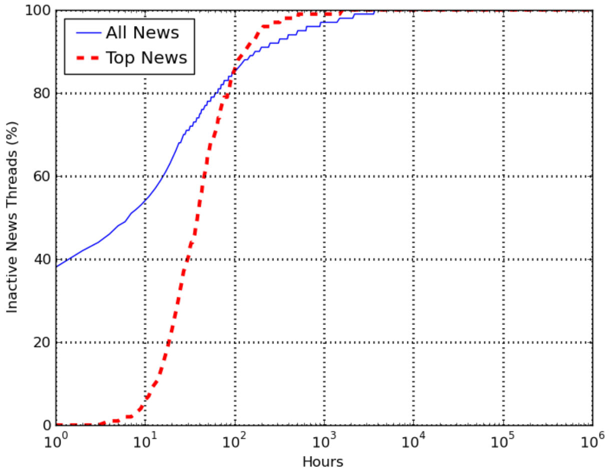


Figure 6: Lifespan of News Tweets

As seen in the figure, a high percentage of news threads, about 38%, are "dead" right after their birth, meaning that none tweet or retweet those news threads except the first (origin) tweet. This does not mean that lots of users have bizarre tastes than others. As we study the source of publishing tweets in Section 3.3, we learned that many news blogger set up automatically publishing services and it turns out that all blogs they write by default publish to Twitter. It becomes a self-promoted behavior then, and possibly these articles themselves are not interesting enough to being spread out.

We also observed that about 85% of tweets loss public interests in four day, no matter whether they are popular news tweets or not. And ten hours or less seems to be a good choice to conduct polling because most of top news are still active, meaning that only good taste users may consistently report popular news.

Indeed that some articles, for instance, *Is Sugar Toxic?* [3] by Gary Taubes, was tweeted or retweeted for months. An popular essay style article may live for a long period, but only less than 1% of news article survive more than a month (720 hours).

5.2 User Activeness

{discuss with John: whether we should keep this section}
Figure 8 shows results when we partition users into three groups based on their activeness. The news-addicted group are users ranked in top 2%, the active group are users ranked in top 25% but not in news-addicted group, and the common-user group are the rest of users.

Intuitively one might think that active users are better choices to get popular news, but the experiment shows a different results. Surprisingly that even the news-addicted group has 25% more tweets than the common-user group, the common-user group perform much well than we expected. From the results we observe here, we have a counter-intuitive conclusion: aggregating voice of active users is not a good idea when doing popular news recommendation , even though some active users are with good tastes.

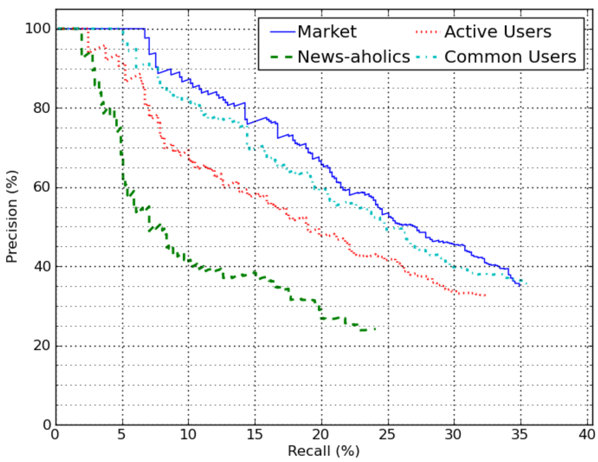


Figure 7: Precision-Recall (News: 2%, $T_\triangle$: 1 hour)

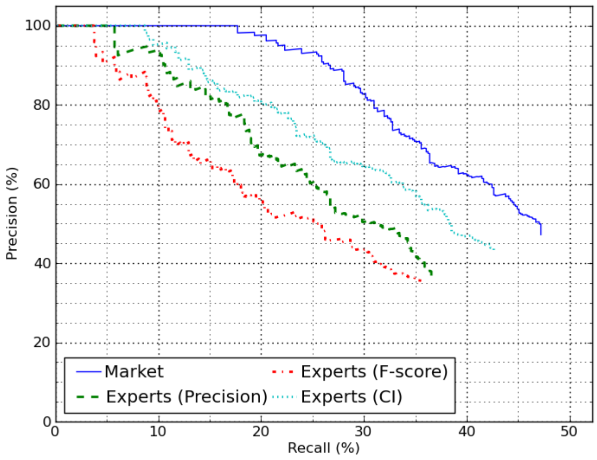


Figure 8: Expert Groups (News: 2%, $T_\triangle$: 4 hours, Expert: 2%)

5.3 Expert Wisdom vs Market Wisdom

In Section 4, we introduced different ways of organizing expert groups. We illustrate the performance of each groups in Figure 8. Note that we set the response criteria to four hours, select 2% users out of the population to form an expert group, and consider a recommendation of top 2% popular news as a hit while evaluating the performance.

We draw two observations here. Firstly, it is almost impossible for any of our expert groups to beat the market. We also increase the size of group, for instance, to 10%, overall the performance is getting better, but still the market performs the best. Secondly, out of all expert selection strate-

[3] http://nyti.ms/pXgxcV

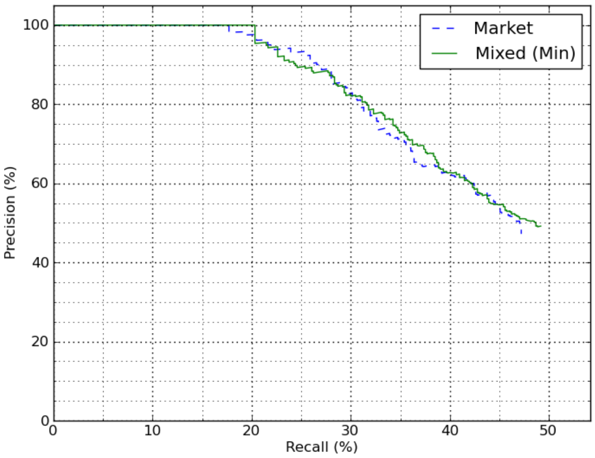


Figure 9: Mixed (News: 2%, $T_{\triangle}$: 4 hours, Expert: 2%)

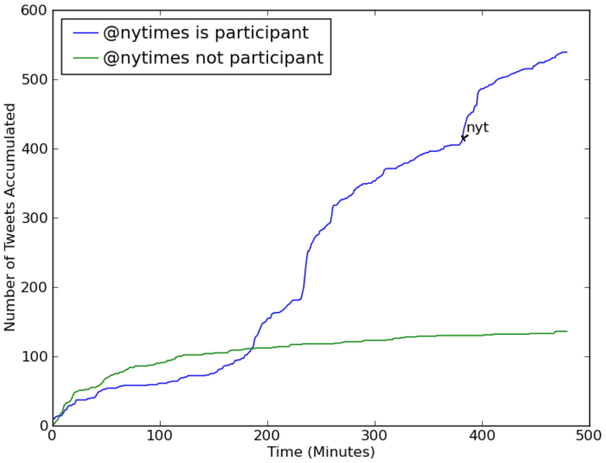


Figure 10: Tweets Accumulation Over Time

gies, the confidence-interval approach is the best. Although we illustrate our observation with Figure 8, we try different combination of parameters and we observe similar outcome on all of them.

Figure 9 shows the precision recall group when we mix up the decisions made by all expert groups and the market. To mix up decision, we use the intellect learned in the training set. For example, suppose that the market decision ranks a news page in 140th position and in the training set we know the precision of the market is 84% at 140th position. If any expert group pick the news page, say, at 10th position and we know the expert group make no mistake in top 20. We boost the news page to 100%. All news page are sorted by highest precision they can get in all models, and top 2% are then returned as the mixed result.

This results shows the expert wisdom help the decision made by the market although the increase is not significant. Moreover, if 2% of users can achieve more than 80% performance of monitoring the entire population, it is resource-wise to adapt the expert selection strategies.

5.4 Social Hub Bias

The spreading of tweets are not only controlled by the quality or interestingness of the news article, but also influenced by who read and (re)tweet the article via Twitter. Therefore, when we view popularity of a news tweet as a quality measurement of the tweet, an obvious objection to this viewpoint is of neglecting the effect from user influence, which has shown to be a critical factor of spreading news[2].

In our experiments, we do see users with many followers, such as the twitter account of *The New York Times*, `nytimes`, bias the number of retweets because the account has over one million followers. In Figure 10, at first tweets are accumulated at an approximated constant speed. {???} minutes after seeing the first tweet, an influential Portuguese users tweet the article and thus introduces the news to another group of audience. With the help of his {???} followers, his role acts as a gateway, forwarding the message and augment the popularity of the news page. Later, the user `nytimes` tweet the article, and right after the tweet `nytimes` boosts the number of tweets associated with the news article again. Intuitively, A large number of audience brings a better chance to diffuse information into social networks, where Bakshy et. al [2] argue that numbers of followers on Twitter are strong indicators.

Therefore, we foresee tweets and retweets are not only affected by the quality of target news page, but also the number of people who may read a tweet, that is, the size of the potential audience. Ideally, to balance user influence, one might propose another reasonable measurement of news quality, for example, the ratio of tweeting after seeing a page – the number of tweets divided by the number of viewers of the page, where the number of viewers compensates the influence some influential users broadcast the page via tweeting. However, to acquire the true traffic requires the cooperation of the domain owner and the consent of every users for tracking their clicking behaviors, which are both not realistic in practice.

6. RELATED RESEARCH

A previous study [16] explains two roles of users on twitter: readers and writers. Abundant writers create abundant data, and users need helps in filtering and discovering interesting contents [4]. A study [10] has shown that if we randomly choose a seed (a user) and a rest (another user), the average path length is 4.12, and the effective diameter [11] (the 90th percentile distance) is 4.8. Comparing to the well known ”six degree of separations” hypothesis [12] in social theory, this shorter effective diameter justify the usage of twitter on propagating time-correlated information, for example, news. Kwak et al. [10] report that 85% of hot topics are actually headline news.

Efficient market hypothesis(EMH) [5] was first introduced to public at the 1969 annual meeting of the American Finance

Association.

Discovery, Filter, Recommendation: [4]

7. CONCLUSION

In this paper, we investigated the problem whether polling from a group of experts is a better idea than from the entire population. We introduced three approaches of picking experts on recommending popular news, and study their performance based on precision-recall graph. The experiment results show that aggregating the voice of a group of experts do not bring better result when making decision of the future in terms on news recommendation. We also show that even though individual expert group might be inferior, they can be used to help the market decision in that experts did see things from different aspect.

We also study the success factors that one should put in mind when developing a popular news recommendation system. The longevity study shows that most news page become inactive after few days. Also, in terms of news tweet spreading on Twitter, we found out that the writer and the reader tweet via different channels. In particular, we show that mobile device channels are getting more popular in reader, and we observe that many writers use auto-posting mechanism.

Although it is controversial and we cannot say that the news predication is equivalent to investment, in this study we do observe a similar conclusion that having experts to make decision does necessarily lead to better results. Though this study, we learned that when trying to tell the future, a group of folks may make better decision that experts who make success decision in the past. Our research also highlight the possibility of building a recommendation system on top of social streaming data.

8. REFERENCES

[1] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, May 1998.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In I. King, W. Nejdl, and H. Li, editors, *WSDM*, pages 65–74. ACM, 2011.

[3] T. S. J. N. Bates. Parliament, Policy and Delegated Power. *Statute Law Review*, 7:114–123, 1986.

[4] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. *Short and tweet.* ACM Press, New York, New York, USA, Apr. 2010.

[5] E. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, May 1970.

[6] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

[7] J. L. Herlocker, J. A. Konstan, L. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[8] C.-C. Hsieh and J. Cho. Finding similar items by leveraging social tag clouds. In *SAC*. ACM, 2012.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. *International Conference on Knowledge Discovery and Data Mining*, 2007.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *International World Wide Web Conference*, pages 591–600, 2010.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, page 177, New York, New York, USA, Aug. 2005. ACM Press.

[12] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[13] R. M. Poses, C. Bekes, R. L. Winkler, W. E. Scott, and F. J. Copare. Are Two (Inexperienced) Heads Better Than One (Experienced) Head? Averaging House Officers' Prognostic Judgments for Critically Ill Patients. *Arch Intern Med*, 150(9):1874–1878, Sept. 1990.

[14] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. X. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Eng. Bull*, 31(2):40–49, 2008.

[15] J. Surowiecki. *The Wisdom of Crowds*, volume 75 of *Anchor Books*. Anchor, 2005.

[16] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, WSDM '11, page 327, New York, New York, USA, 2011. ACM Press.

9. BACKUPS

{chucheng: move unwanted context in previous version here}

9.1 Spreading of Tweets

{this section might be redundant as the context here is very simple} {Note: twitter does not tell where you read the tweet, but only the initial tweet.}

Twitter provides no information about where a user read the information. When a user click the **Retweet** button, Twitter only provides the information about which the corresponding original tweet is.

To study how a tweet is propagated, we create a *tweet-spreading path* graph for every origin tweet and its descendant(s). Figure 11 is an example of the graph, where every node in Figure 11 stands for a tweet, represented by a unique tweet identifier.

When two tweets shares one original tweet, we create an edge if we see (1) a tweet of u_i is after another tweet of u_j and (2) u_i follows u_j. In case two users follow each other, the created time of tweets determine the direction. For instance, $144562298 \longrightarrow 1865567$ means that a user posting the tweet 144562298 follows another user posting the tweet 1865567, and the tweet 1865567 happens before the tweet 144562298. In case that more than one of u_i's friends retweet the origin tweet, we say the earliest tweet prevails, because Twitter actually shows only the earliest one on u_i's portal.

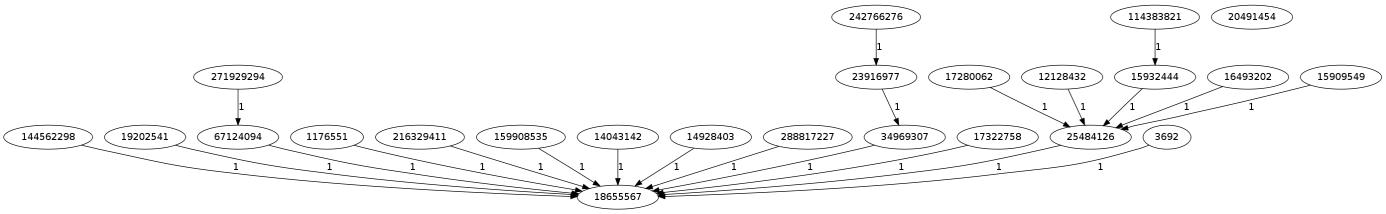


Figure 11: A Tweet-spreading Path Graph

A disconnected node refers to a retweet, where the author of the retweet follows none else in the news spreading path. Disconnected nodes are rare, because most users only read (and thus may retweet) tweets from people they know(follow). However, Twitter actually allows a user to retweet a tweet of a person he/she does not follow.

9.2 Is Polling Experts a Good idea?

The main problem we investigate in this work is to compare different expert selection strategies against the market, and to see whether a group of carefully selected experts behaves well in predicting valuable news.

9.3 Creating News Tweets

From our observation, a user usually create a news tweet (an origin tweet) in the following two major ways: (1) the user could click a **TWITTER** button, as shown in Fig 1, provided by most popular news website for facilitating the sharing of news pages; or (2) the user could write up a tweet, attaching with an additional URL. Since the first approach achieves the same functionality with fewer efforts, it seems to be a preferred way to share news by many users, especially when the operational convenience is a concern, for example, reading news though a smart phone or a tablet. In contrast, the second approach requires a URL shortening service to address the limit of 140 characters per tweet. Moreover, it requires more copy-pastes than the first approach.

{we may shift the following two parts to the experiment section} Tweets generated through the first approach are all news tweets and they follows a convention: each of them consists , in most cases, the title of a news, and a short URL prefix plus a hash value, for instance, *http://ntyi.ms/ABCDEF*. As a result, news tweets can be easily recognized through filtering keywords in the prefix.

(Collecting News Tweets) In this study, we limit our experiments to news tweets originated from the first approach, which is the preferred approach we believe. This limitation is necessary because to extract news tweets generated by the second approach, we have to gather every tweet with a URL, which is beyond the limit of any white-listing policy offered by Twitter.

{move this paragraph to experiments}

9.4 Anything else

Note that when a user creates a tweet, in order to meet the length constraint of a tweet, the original URL to the news page is often encoded(shortened) through some short-url service provider, for instance, `http://bitly.com/`. So, whether a tweet is a news tweet cannot be determined by the contents of the tweet, but requires a name resolution of the short URL.

9.5 Experiments

9.5.1 target news

{move the rest to back up}

We say our targets are those tweets which contains URLs pointing to "popular" news pages (Section 9.6). Note that we discriminate tweets solely based on links they contain in that we assume that a user make his judgment of retweeting based on the content of page but not the blurb in a tweet.

The number of total tweets related to a news page N_i, $R(N_i, t)$, is always unknown because someone can always retweet a old tweet later on. It is possible that some old news pages catch public attention again after a few weeks later, say, because a hearsay becomes the truth. Therefore, to tell the exact $R(N_i, t)$ is impossible. However, with regards to our finding in Section 5.1, we learned that 90% of tweets actually gain their tweeting or retweeting within 4 days from the very first tweet so in our experiments we assume the life of a news page cease if in 4 days we see no tweet that is related to the news pages.

We then sorted news pages based on $R(N_i, t)$ and then select top 2/5/10% as popular news pages. Our goals are (1) to find experts that frequently report these popular news pages, and (2) to compare decisions from experts and folks in terms of finding popular news.

9.5.2 Evaluation – Frobenius Norm

In our study, every recommendation algorithm outputs a list of news article sorted by polling results of a group of users. Given a model x, we denote the ranking results as $\mathbf{R}_x$

9.6 Popular News Pages

News articles are recommended though a voting process after we select experts. Note that we set up a pre-set time constraint. If an expert post a tweet linked to a news page within the constraint, we consider the page of getting one vote. News articles are then ranked by the sum of votes they get.

To compare voice from experts and voice of the market, we rank news articles based on votes from all users within an identical pre-set time constraint, and consider the output as the decision of the market. That is to say, we let everyone vote to mimic the concept of market proposed in the efficient market hypothesis.

Table 4: A list of symbols

Symbol	Definition
u_i	A user
N_i	A news
$R(N_i, t)$	The total number of tweets linked to the news N_i after t hours.
β	A weighting parameter used in F-score formula

To know which news pages are truly popular ones, we let everyone vote without any time constraint. Namely, after a reasonable period of waiting, say 7 days, we counts their numbers of total tweet. And news pages rank in top position, for instance, 10%, are considered as *popular news pages.*

We will continue our discussion of comparing experts and the market in our experiments (Section 3).

9.7 News Residual Value Function

We see the value of a news as the number of recommendations over its life time. For a news page N_i, we define a function $R(N_i, t)$ to represent the total number of tweets containing a URL pointing to the page after t hours of its first-seen on twitter space. That is to say, $R(N_i, 0)$ refers to all recommendations of the $N_i's$ entire lifespan $(t > 0)$. Intuitively , the function $R(N_i, t)$ shows the residual value of news N_i.

In our model, news page are selected by the function $R(N_i, 0)$. If we rank news articles based on their lifespan values. News pages that are ranked in the top are consider as true valuable news articles. In our experiments, a threshold, say 10%, is then selected to make a cut for defining the boundary between a valuable news page set and others. Also, in this paper, *popular news* refers to news pages that ranked in top 10%. These news are considered valuable news and are recommended by many users.

Revise here: Chucheng: I think we can replace this section with lifespan. $R(N_i, 0)$.

9.8 Symbol

We end this section by summarizing symbols for modeling in Table 4.

9.9 News Spreading pattern

A valuable/interesting news article does not correspond to a tree graph with a high depth. The spreading pattern of news tweets are similar to ripples that the spreading pattern looks like a ripple, not a wave.

9.10 Social Bias Backup

{ I'm still thinking what is the best graph to show the social hub bias. Figure 12, Figure 13 are two examples but still thinking of better ways.}

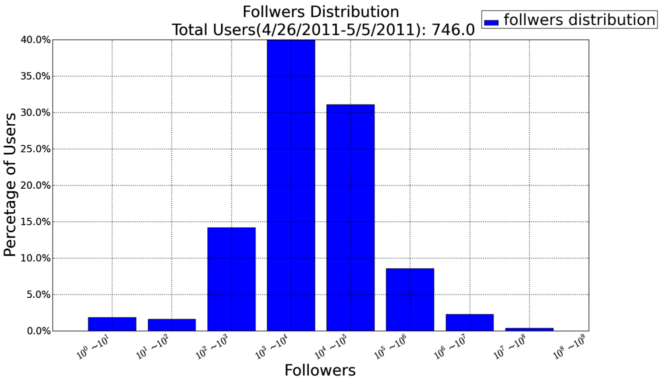


Figure 12: User distribution

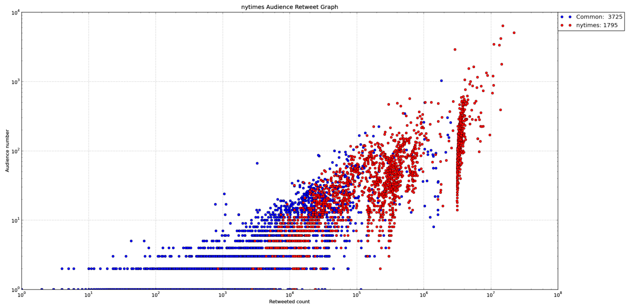


Figure 13: Social Hub Bias