

DanZarrella's

# The Science of **ReTweets**

2009

# Contents

Why ReTweets Matter 4

## **More Followers = More ReTweets?**

Distribution of ReTweets per Follower 5

RTpF of Suggested Users 6

## **ReTweeting & Links**

Link Occurrence in ReTweets 7

ReTweetability of URL Shorteners 8

## **Linguistics**

Most ReTweetable Words & Phrases 9

Least ReTweetable Words 10

Average Syllables per Word 11

Readability Grade Levels 12

Word Occurrence & Novelty 13

Parts of Speech 14

Punctuation Occurrence 15

Punctuation Types 16

## **Psychology**

RID Content Types 17

RID Attributes 18

LIWC Attributes 19

## **Timing**

Time of Day 20

Day of Week 21

About the Author & Data 22

*“Ideas shape the course  
of history.”*

-John Maynard Keynes

# Why ReTweets Matter

I spend a lot of time working on ReTweets, because I believe them to be one of the most important developments in modern communications, extending far beyond the Twittersphere.

Like “The Matrix” was composed of computer code, the real world is made of infectious information. Your chair, your desk, the computer you’re reading this on, the food you’ll eat today, the money you’ll earn: they all began as ideas jumping from person to person. None of it would exist if the concept wasn’t contagious.

Ideological epidemics have made and lost fortunes, they have saved countless lives and caused horrific wars, they have birthed and destroyed nations. **Clearly the most powerful weapon known to man would be the ability to create powerful mental viruses at will.** The very course of human history would be at your whim.

You don’t spread ideas just because they are “good;” you spread them because of some other trigger or set of triggers has been pulled in your brain. And that trigger fires the biggest gun ever seen.

And yet, a reliable, repeatable method of crafting a contagious idea has not emerged.

The advent of the web changed how memes spread: it made them spread faster, it exposed them to more people, and it removed many of the constraints imposed by the limits of human memory. But there is one change that dwarfs them all: **observability.**

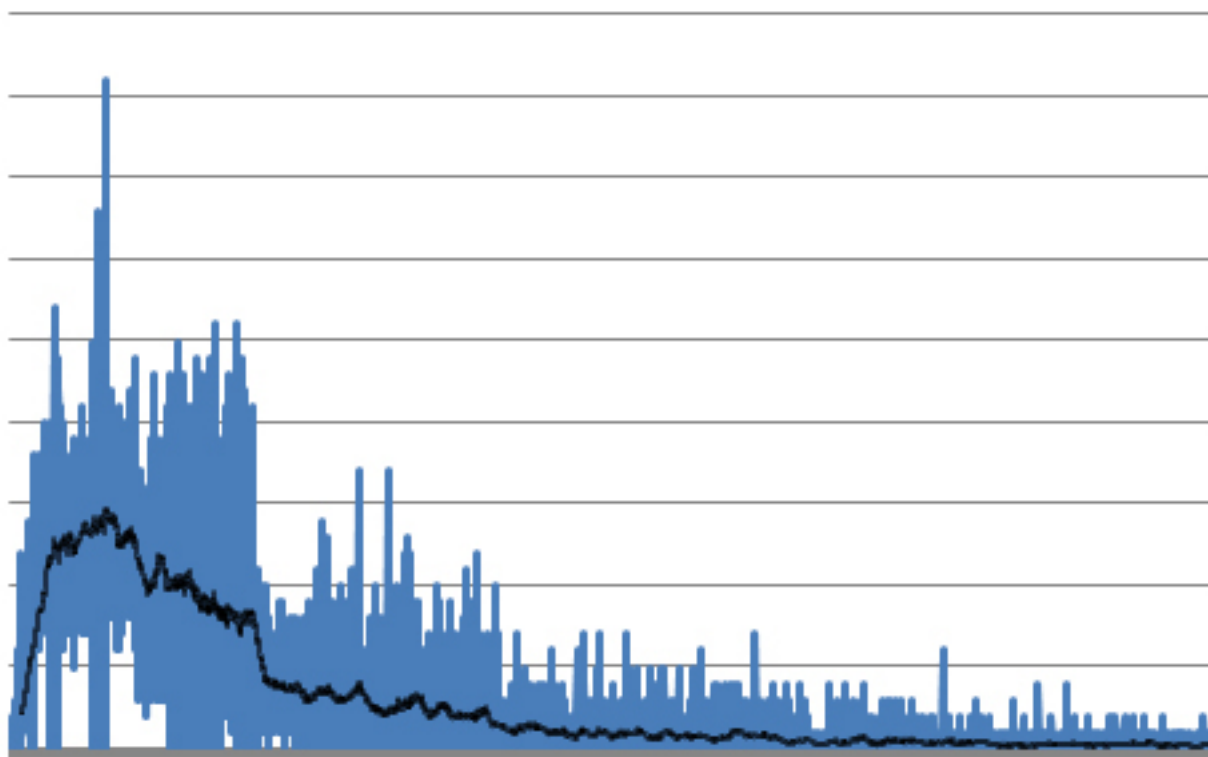
We can now compare millions of viral ideas to uncover the building blocks of contagiousness.

ReTweets may seem like a small idea, and they are in some ways. But that small idea is the first real window into how ideas spread from person to person. We can study the linguistic traits, the topical characteristics, the epidemiological dynamics, and the social network interactions that take place when a person spreads a meme.

Not only can this information help us create more contagious Tweets, but many of the lessons learned through ReTweets will be applicable to viral ideas in other mediums.

**For the first time in human history we can begin to gaze into the inner workings of the contagious idea.** That most powerful force can now be put under our microscope and probed for its secrets.

# Distribution of ReTweets per Follower



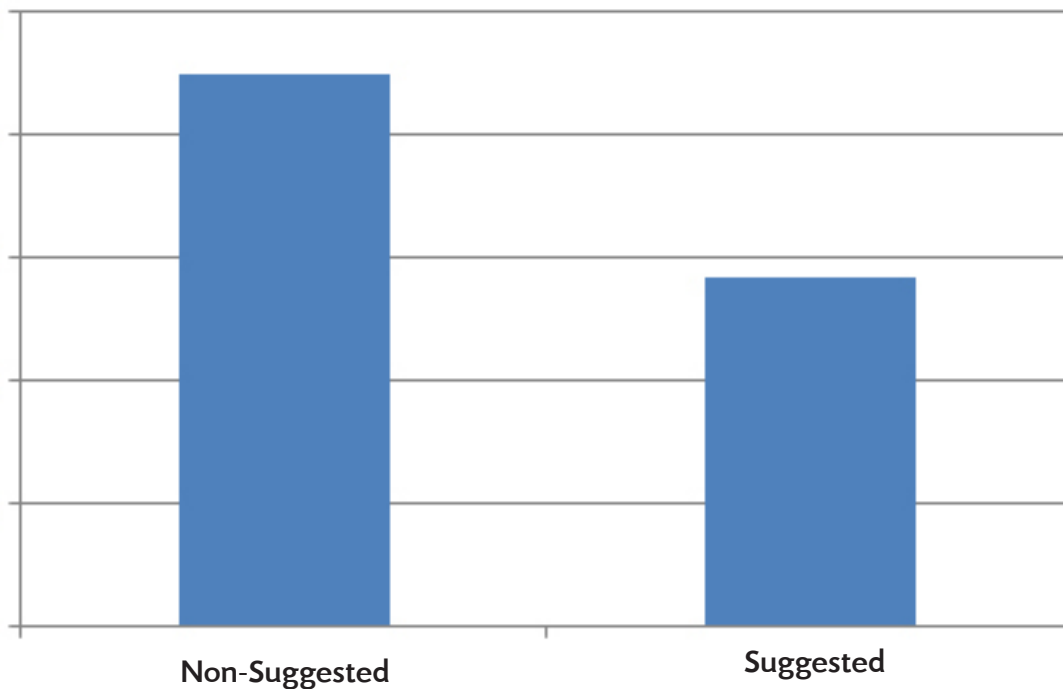
When I started thinking about how to get more ReTweets, my first thought was to get more followers; clearly, more followers would mean more ReTweets. To check this assumption, I looked at ReTweets per Followers (RTpF), the number of ReTweets per day divided by the number of followers.

The graph above shows the distribution of RTpF in the top 9000 most followed users in my database; I've graphed the actual distribution line in blue, with a 30-point moving average over it in black.

Here we see that while most users had an RTpF of under 1% in my dataset, some users showed much larger ratios, possibly indicating that there are a class of users who are more "ReTweetable" than others.

This means that while users with more followers will get more ReTweets, some users are able to get lots of ReTweets without lots of followers; their content must be more contagious.

# RTpF of Suggested Users

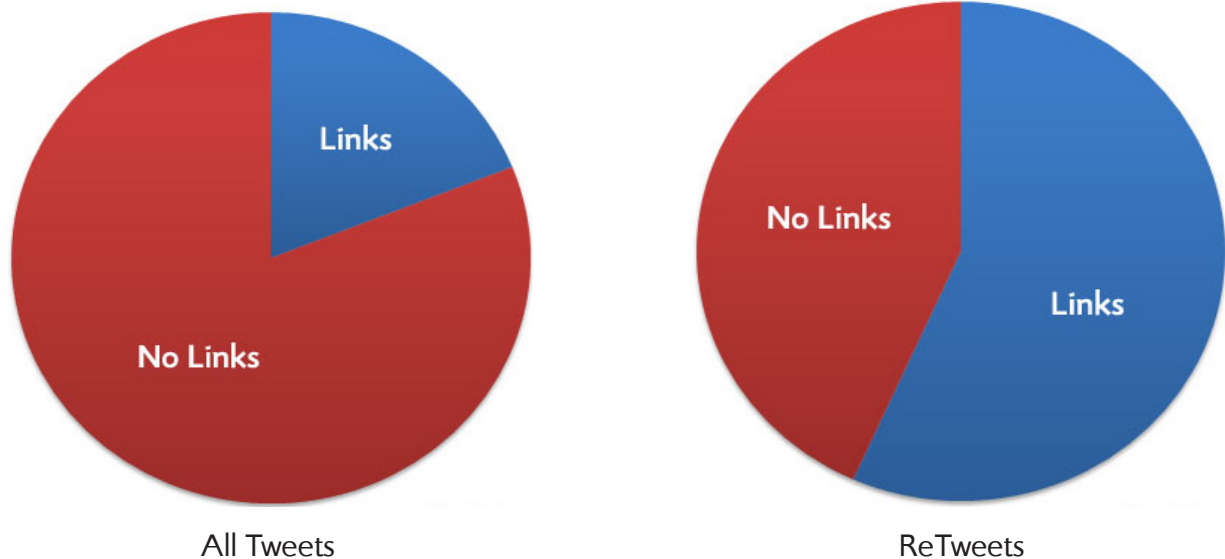


Twitter maintains a list of users as suggested people for new users to follow; the people on this list gain tens of thousands of new followers every day and are among the most followed people on Twitter.

I looked at 200 suggested users and compared them to the 200 most followed users not on the list. Since many of the suggested users are the most followed people on Twitter, they had a much higher average number of followers. So I compared the two groups using the RTpF metric.

The result is clear: suggested users are far less ReTweetable. I think this is likely due to the fact that many of the followers gained by those users on the suggested list are new Twitter users and may be less ReTweet-savvy.

# Link Occurrence in ReTweets

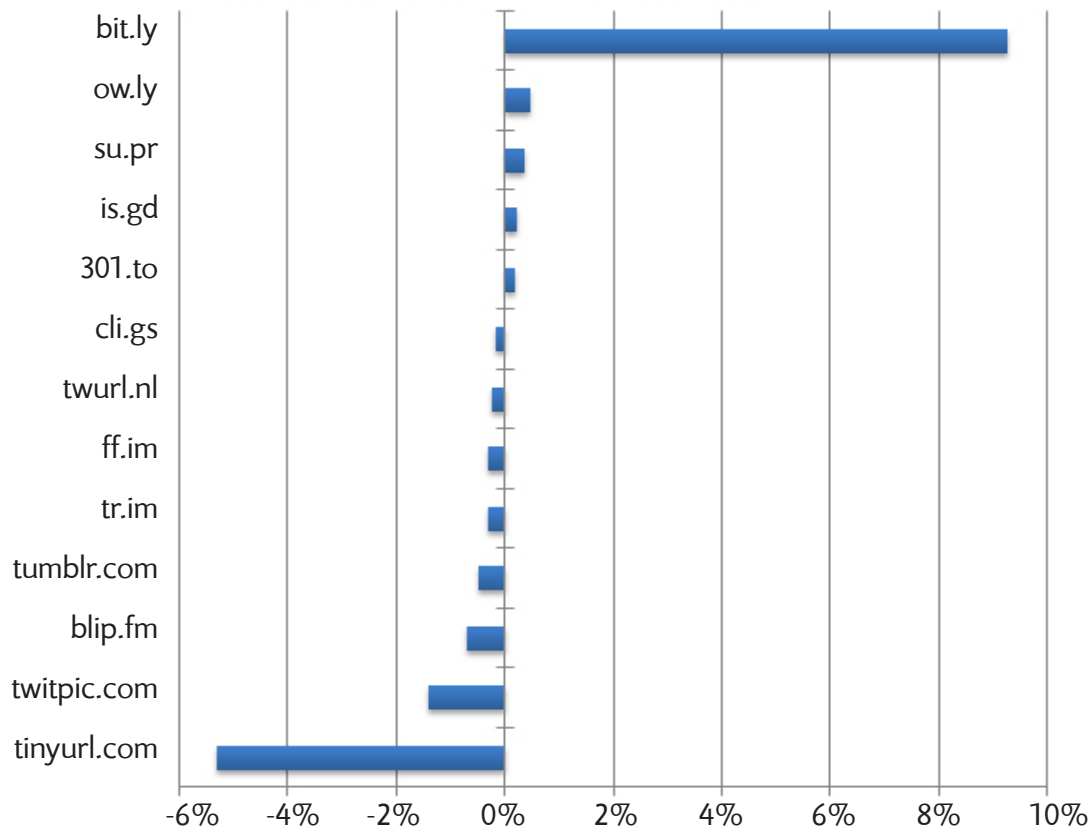


I began to study the content of Tweets to identify traits that are correlated with more contagious, or ReTweetable, content. The first such trait was the presence of a link.

I found that in a random sample of normal (non-ReTweet) Tweets, 18.96% contained a link, whereas 3 times that many ReTweets (56.69%) included a link.

This means that not only are ReTweets an accepted way to spread off-Twitter content, the presence of a link may increase a Tweet's chances of being shared.

# ReTweetability of URL Shorteners



I know that most ReTweets contain a link, but there are hundreds of different URL shortening services available to help you save space with that link. I analyzed my database of over 30 million ReTweets and compared them to over 2 million random Tweets to find which shorteners are the most (and least) ReTweetable.

I calculated how much more or less often each URL shortening service appeared in ReTweets than it did in normal Tweets and presented this value as a percentage. For instance, in my data 9.28% more ReTweets than random Tweets used bit.ly. I took into account the fact that ReTweets tend to contain more links than average Tweets and normalized the occurrence values.

I compared the percentage of occurrence for each shortener in random Tweets to the same shortener in ReTweets, to control for the popularity of services like bit.ly and tinyurl.com.

The short, post-Twitter shorteners, bit.ly, ow.ly, and is.gd were all more ReTweetable than the older, longer, tinyurl.

# Most ReTweetable Words & Phrases

1. you
2. twitter
3. please
4. retweet
5. post
6. blog
7. social
8. free
9. media
10. help
11. please retweet
12. great
13. social media
14. 10
15. follow
16. how to
17. top
18. blog post
19. check out
20. new blog post

I compared common words and phrases in random Tweets and ReTweets to find those words that occur in ReTweets more than they occur in normal Tweets.

The word “you,” while very common, seems to occur especially often in ReTweets, indicating that if you’re talking to “me,” I am more likely to ReTweet it.

Its really not surprising that “Twitter” ranks high, but this is a good reminder that self-reference is always good for buzz in social media.

The words “please” and “please ReTweet” are very ReTweetable (“please rt” also ranked highly). It’s hard to overstate how important it is to ask for the ReTweet when you want it; calls to action work.

“New Blog Post” is the common prefix used when a person Tweets about, well, a new blog post to their site. That this ranks so highly tells us that Tweeting your posts is a very smart thing to do.

# Least ReTweetable Words

1. game
2. going
3. haha
4. lol
5. but
6. watching
7. work
8. home
9. night
10. bed
11. well
12. sleep
13. gonna
14. hey
15. tomorrow
16. tired
17. some
18. back
19. bored
20. listening

What about the words that are least likely to get your ReTweets?

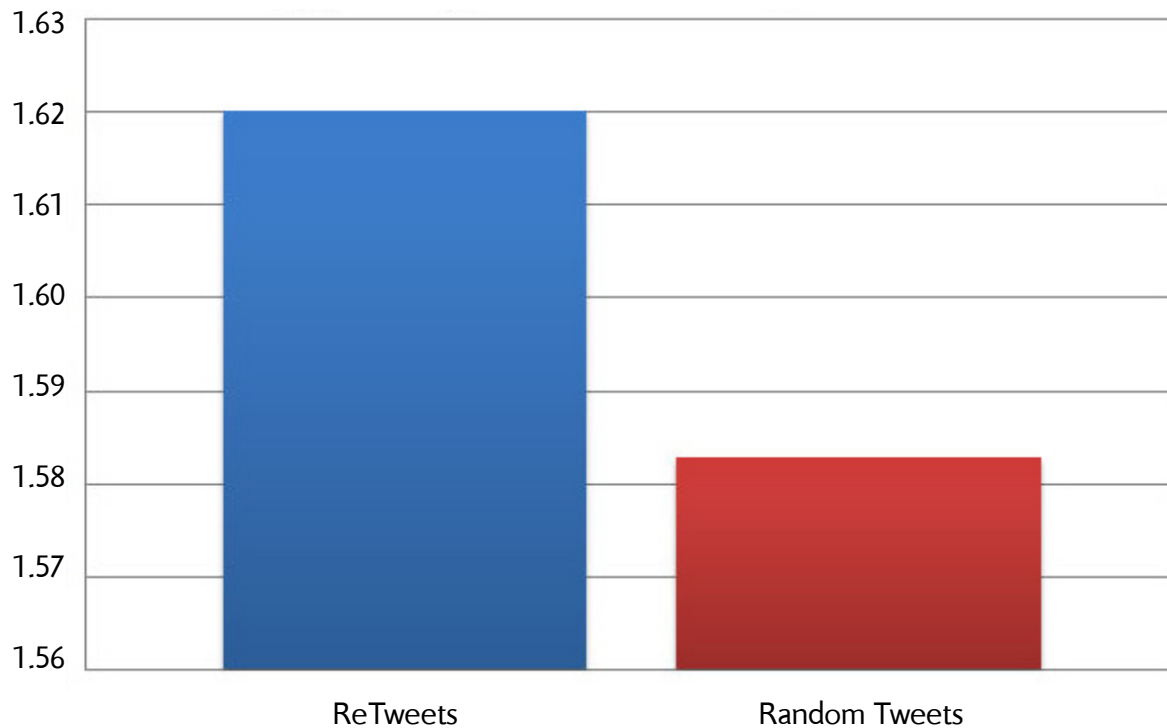
There are a number of “-ing” verbs, including “going,” “watching” and “listening,” which reinforces my understanding that answers to the “What are you doing?” question don’t get very many ReTweets.

The presence of “sleep,” “bed,” “night,” and “tired” indicate that people often Tweet “goodnight” style messages, but generally don’t ReTweet them.

The relatively informal nature of many of the words on the list including “lol,” “gonna,” and “hey,” show that simple or slang conversation is not ReTweetable.

The lesson learned here is that if you’re trying to get more ReTweets, don’t just engage in idle chit-chat or Tweet about mundane activities.

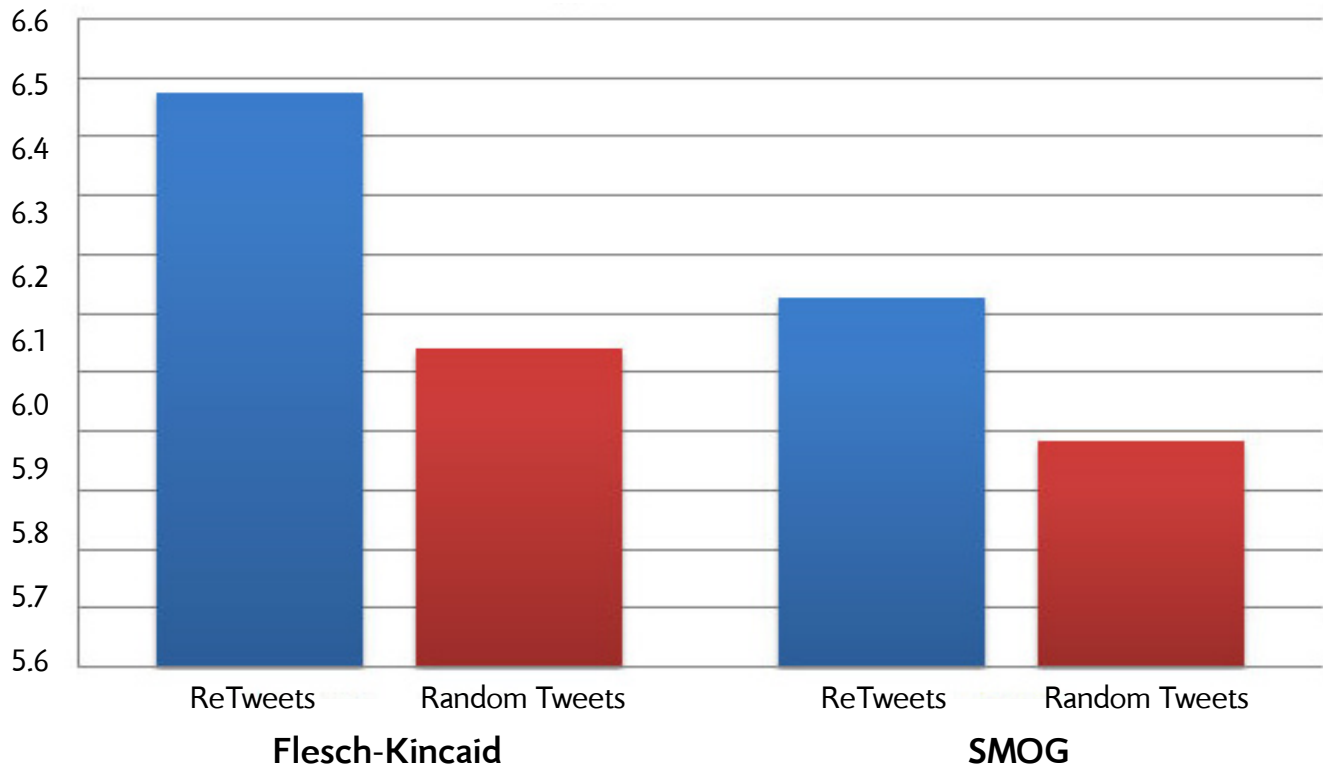
# Average Syllables per Word



I tested the assumption that simplicity is a vital component of ReTweets (as it has been observed in other viral-content types) and I found that random Tweets have 1.58 syllables per word on average, while ReTweets have an average of 1.62 syllables per word. Longer, higher syllable-count words are typically more complex, indicating that ReTweets may be more complex than their less viral counterparts.

Be sure to notice the scale of the graph above, the difference is small, but it certainly challenged my hypothesis that ReTweets are less-complex.

# Readability Grade Levels

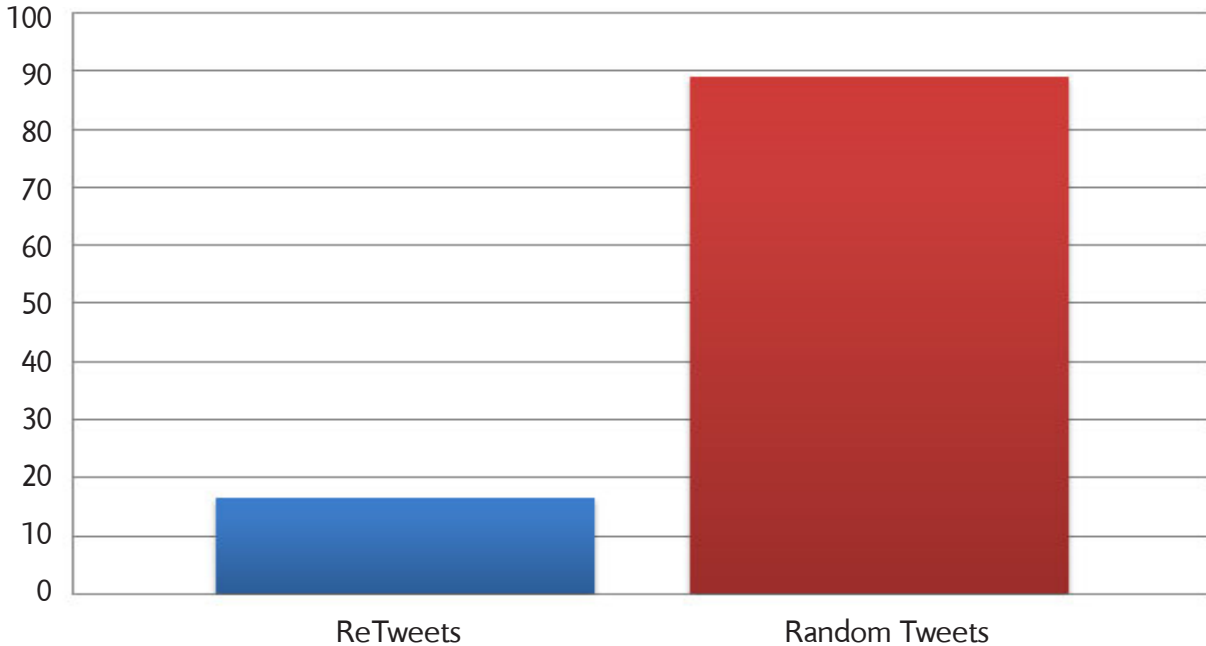


I then compared two different types of reading grade level analysis metrics and they revealed that ReTweets, in general, are less “readable” and require a higher level of education to understand.

A Flesch-Kincaid test gave ReTweets a reading grade level of 6.47 years of education, while random Tweets only required 6.04 years. The similar SMOG test (Simple Measure of Gobbledygook) indicated that ReTweets required 6.13 years of schooling, with random Tweets only needing 5.88 years.

Again, take care to notice the scale of the Y-axis.

# Word Occurrence/Novelty

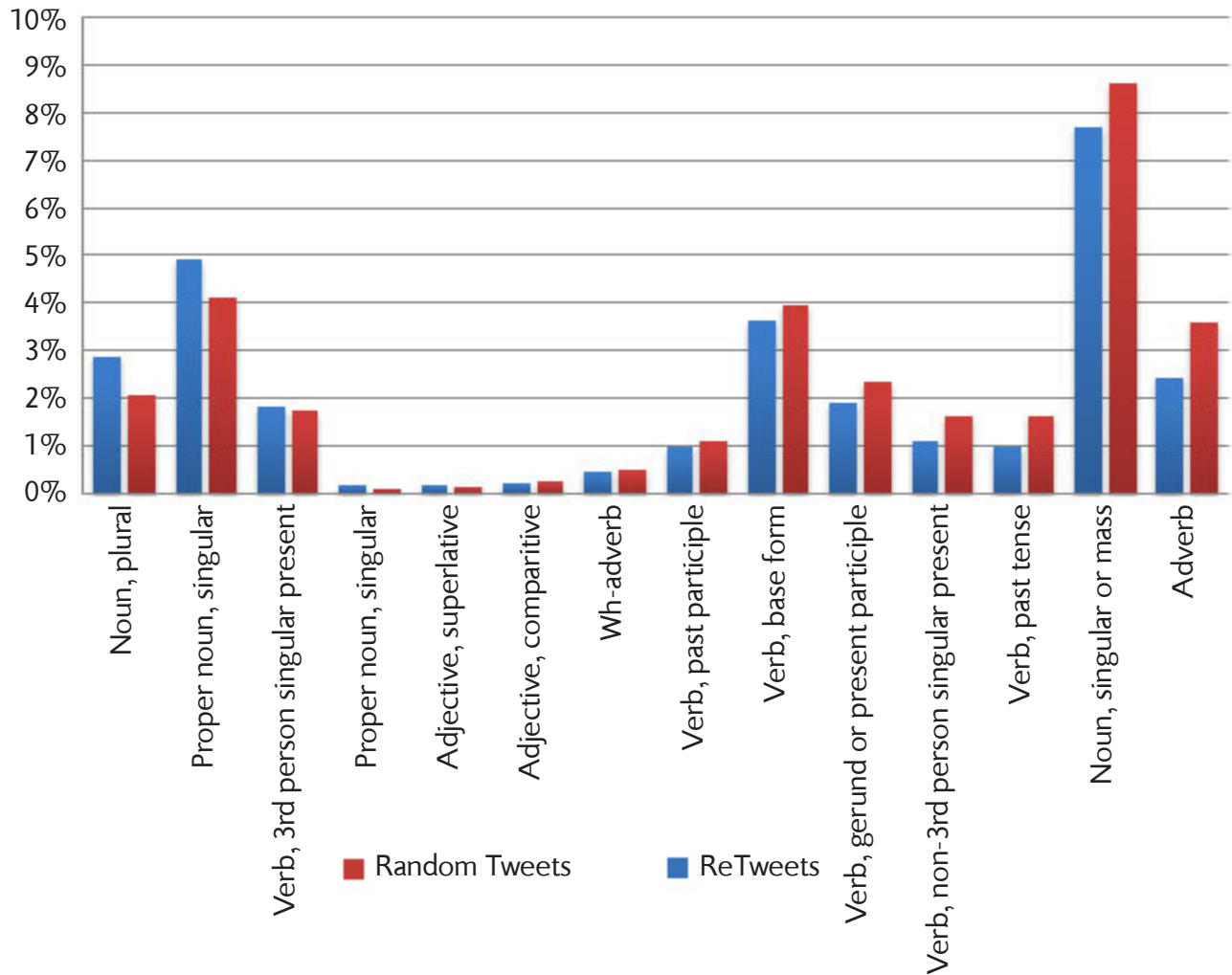


Another characteristic commonly found in viral content is novelty; that is, the “newness” of the ideas and information presented. I created a measure of novelty by counting how many other times each word in my sample sets occurred.

In the random Tweet sample, each word was found an average of 89.19 other times, while in the ReTweet sample each word was only found 16.37 other times.

This shows us that while simplicity may not be very important to ReTweetability, novelty certainly is.

# Parts of Speech

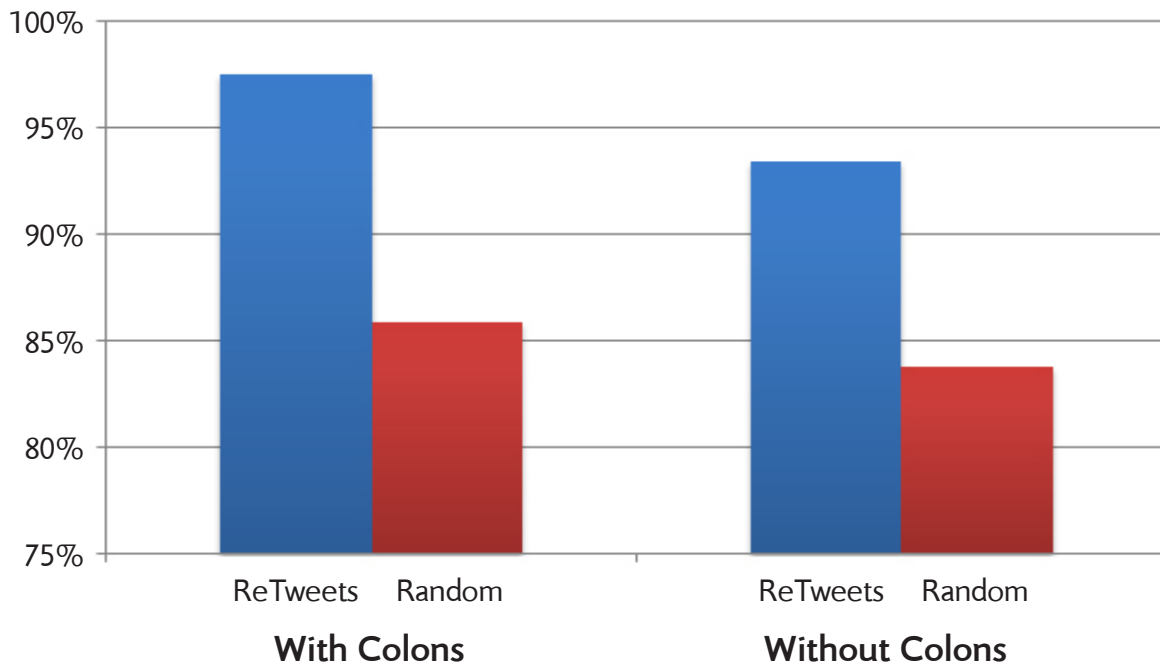


Part of speech (POS) tagging is an analysis technique in which an algorithm is used to label each word in a piece of content as a specific part-of-speech—noun, verb, adjective, etc.

The graph above shows what percentages of words in each sample were labeled as a specific part-of-speech. It lists only the most interesting parts from the much larger list of POS tags.

Interesting points from this data include the noun and 3rd-person heaviness of ReTweets, indicating a subject matter and headline type nature.

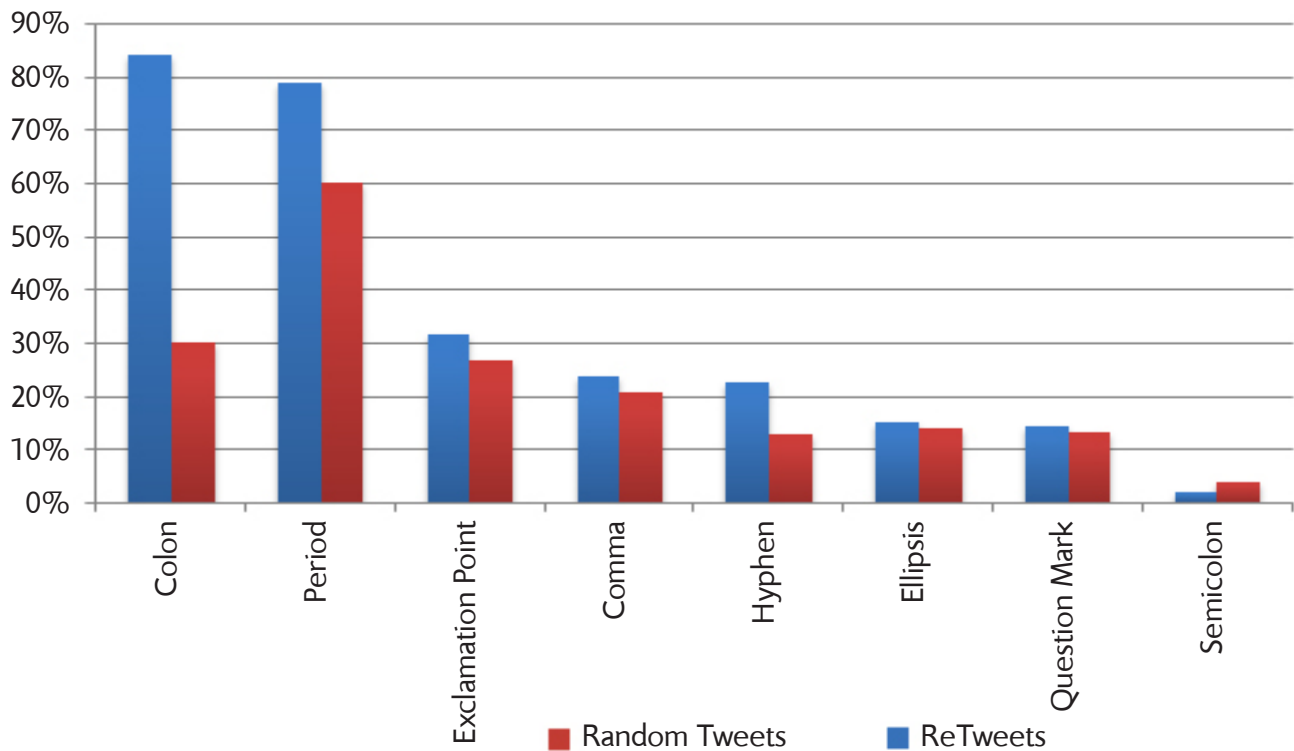
# Punctuation Occurrence



I compared a random sample of “normal” Tweets to a sample of ReTweets and found that 85.86% of Tweets contain some form of punctuation, and an overwhelming 97.55% of ReTweets do as well.

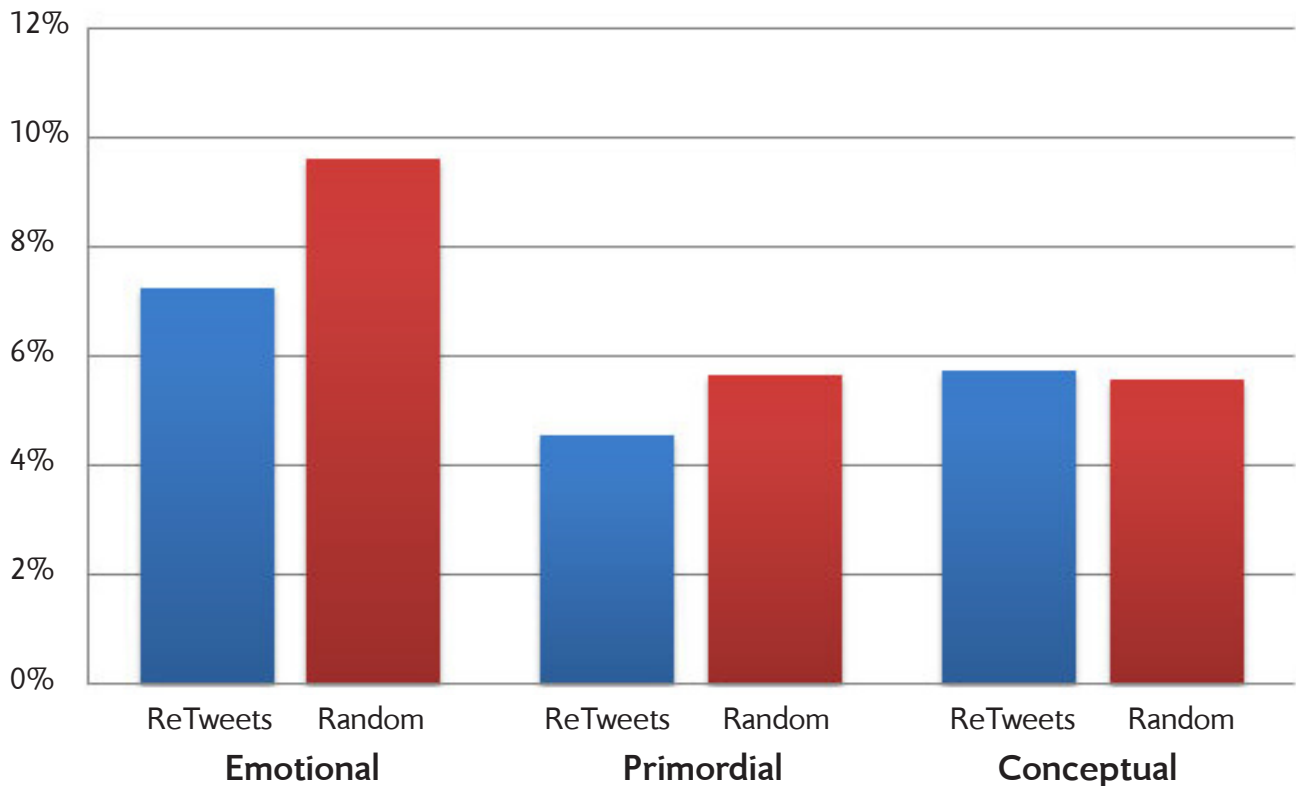
Of course, the prevailing ReTweet format includes a colon to better display the original Tweet, but even when ignoring this form of punctuation, ReTweets still contain more punctuation than non-ReTweets (93.42% to 83.78%).

# Punctuation Types



I then analyzed the frequency of specific types of punctuation and found that hyphens, periods and colons are the most ReTweetable punctuation, occurring far more commonly in ReTweets than in regular Tweets, while the rarest mark, the semicolon, is the only unReTweetable punctuation mark.

# RID Content Types

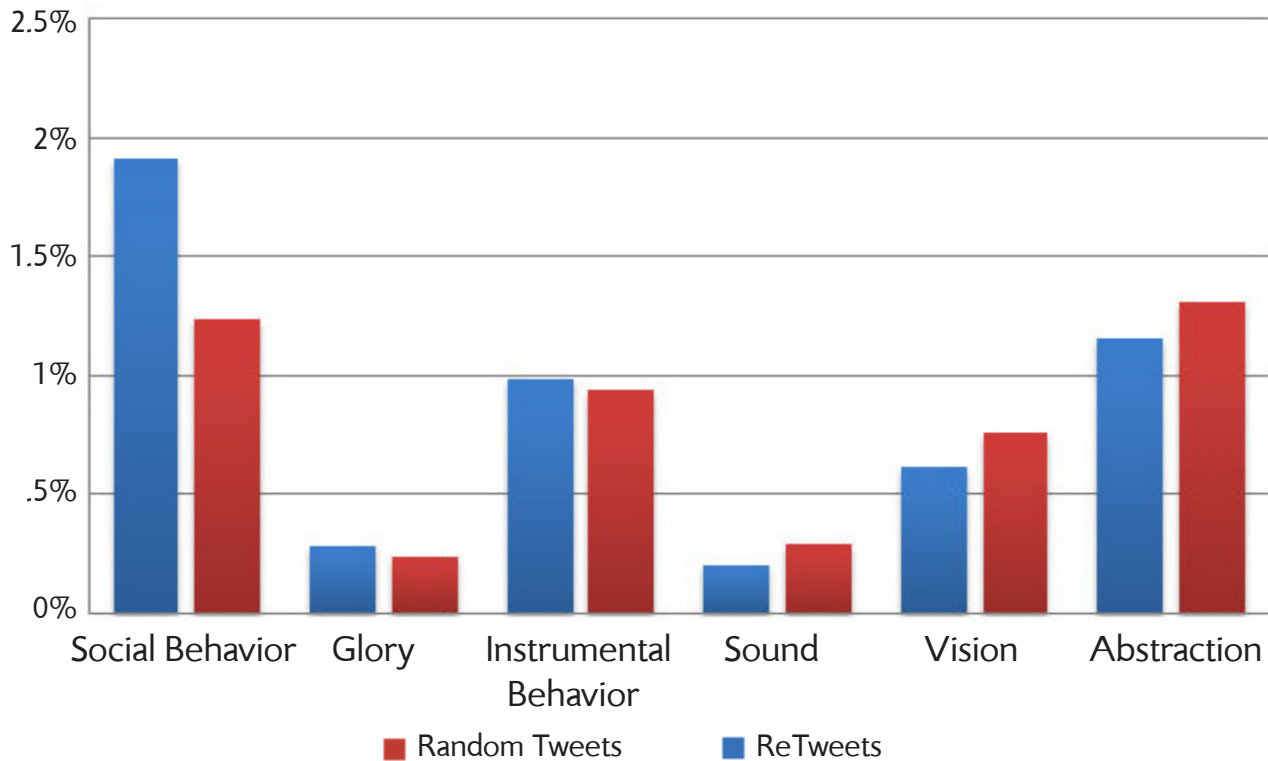


I used the two linguistic lexicons to analyze ReTweet content: RID and LIWC.

First is the more “Freudian” Regressive Imagery Dictionary (RID). This coding scheme is designed to measure the amount and type of three categories of content: primordial (the unconscious way you think, like in dreams); conceptual (logical and rational thought); and emotional.

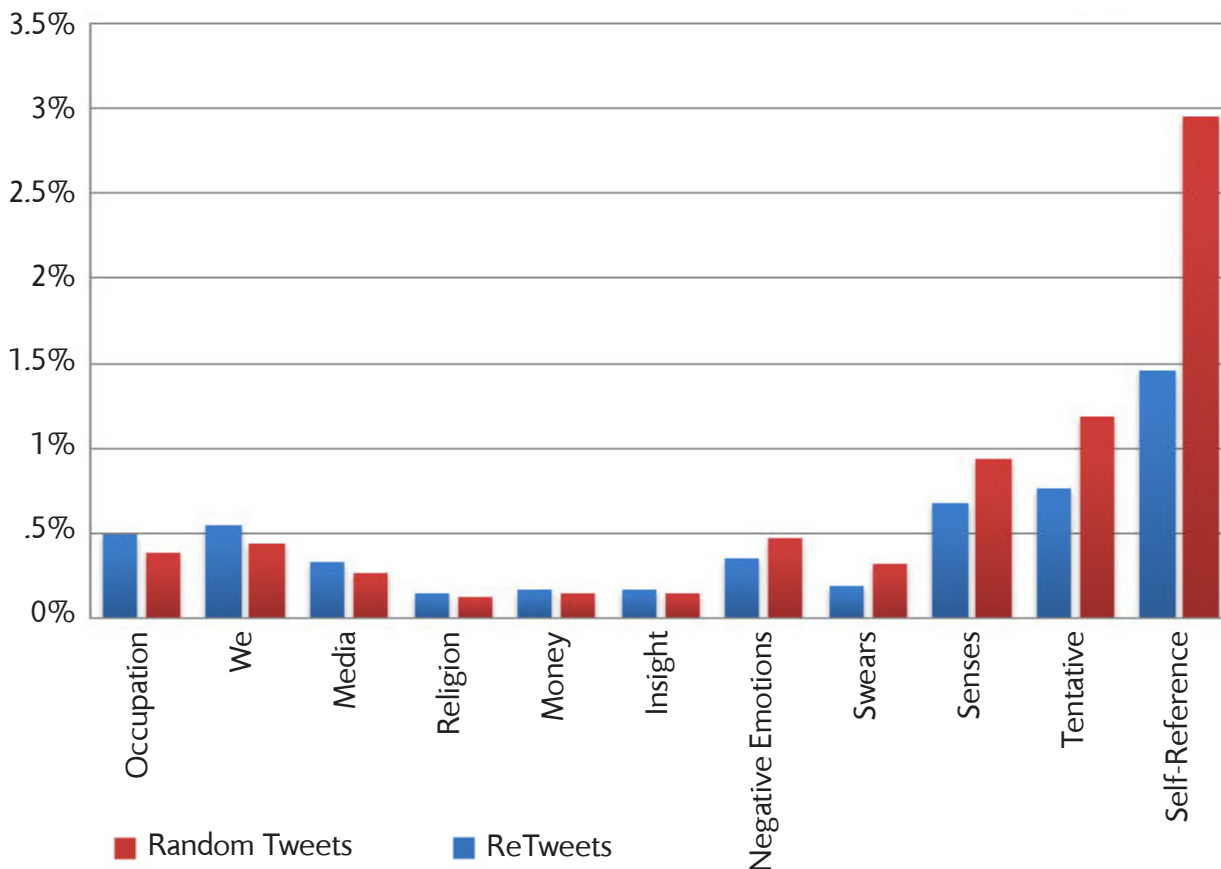
The graph above shows that ReTweets contain less primordial and emotional content than random Tweets and more conceptual content.

# RID Attributes



Looking at specific RID attributes, I saw that social and instrumental (constructive words like build and create) behaviors are ReTweetable, while abstract thought and sensation-based words are not.

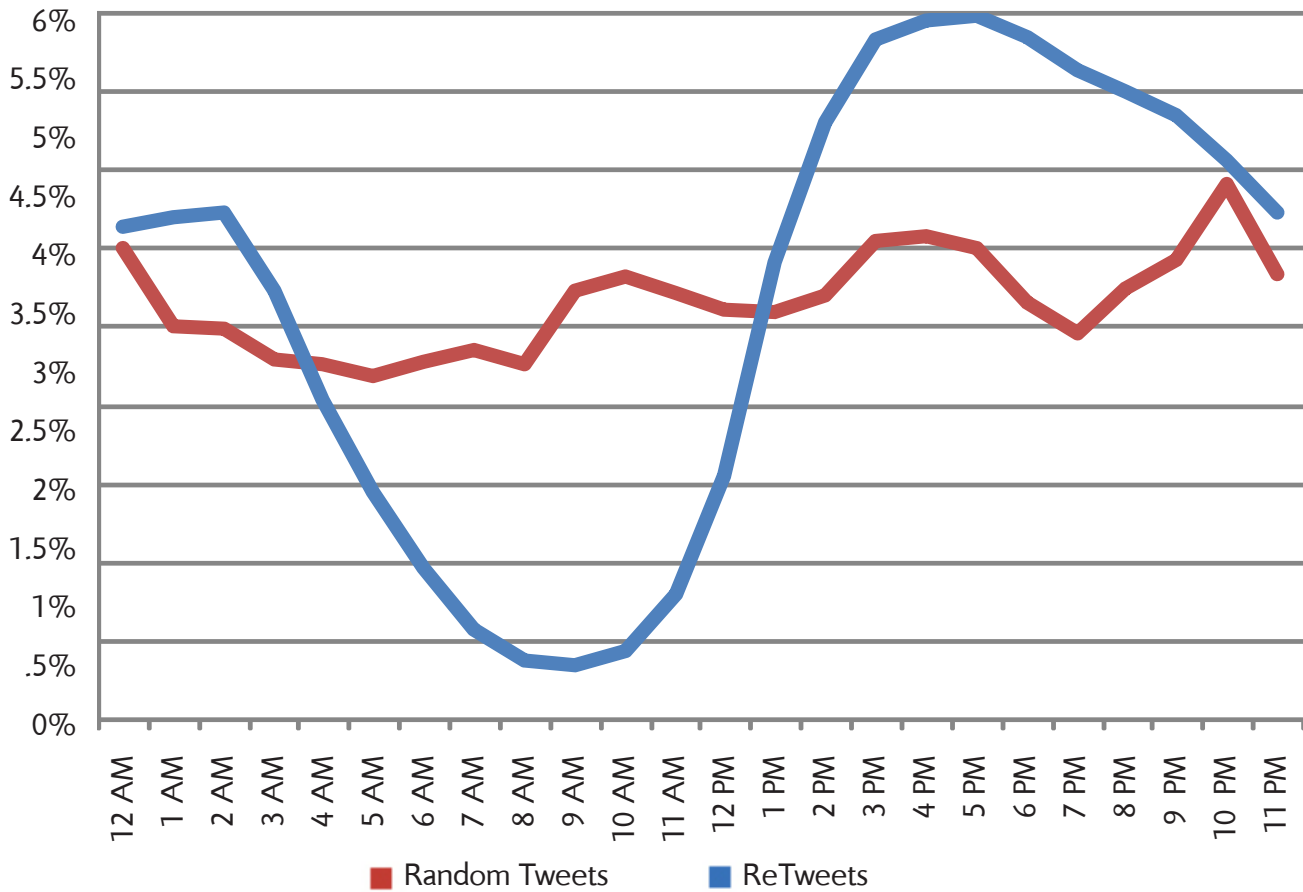
# LIWC Attributes



The next analysis I performed used LIWC (pronounced “Luke”). This is a lexicon similar to RID, but based in more reviewed and accepted research and refined over 15 years. LIWC measures the cognitive and emotional properties of people based on the words they use.

LIWC analysis shows that Tweets about work, religion, money and media/celebrities are more ReTweetable than Tweets about negative emotions, sensations, swear words and self-reference.

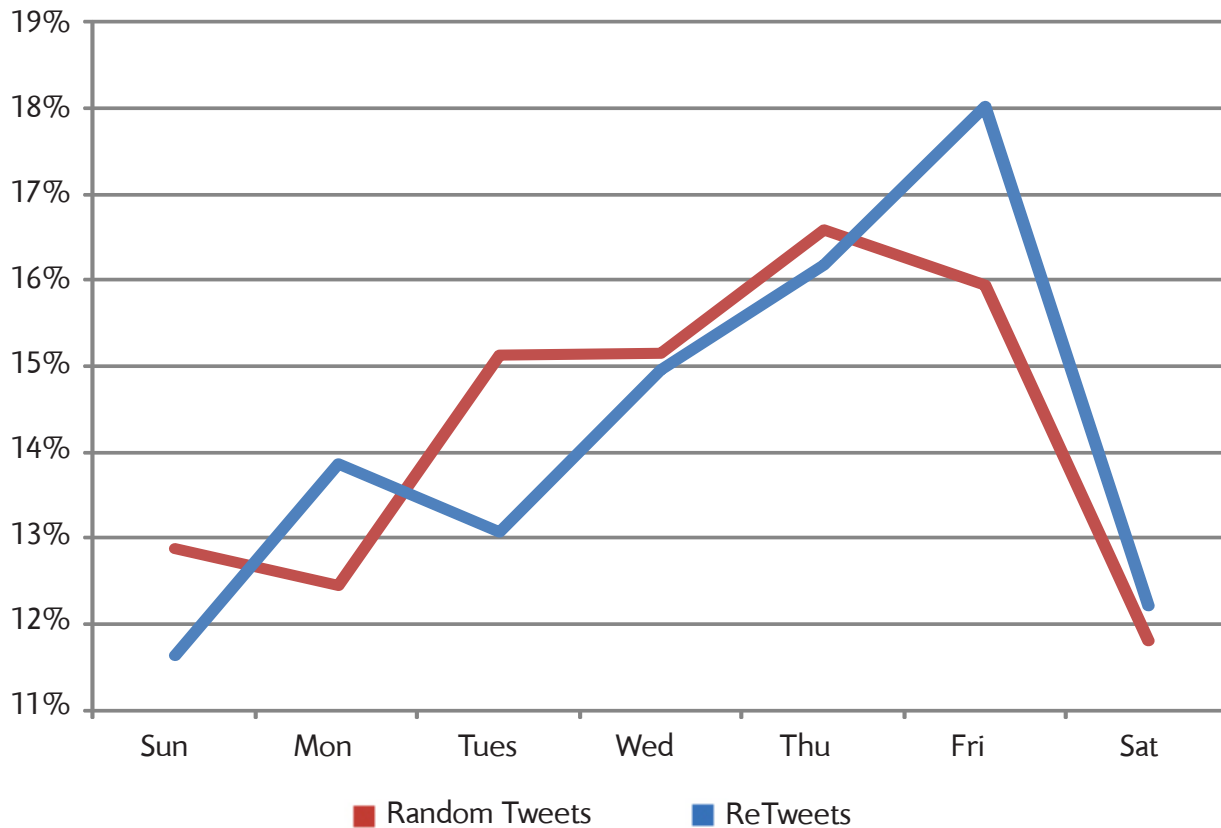
# Time of Day (EST)



I compared the volume of ReTweeting that occurs during the day to the volume of regular Tweets to find that ReTweeting is much more diurnal. While overall Tweet volume peaks during business hours and evening, ReTweeting occurs much more frequently between 3 PM and midnight.

If you want to get ReTweets, it makes sense to post your content during these hours.

# Day of Week



I also compared the volume of activity that occurs on various days of the week. Overall Tweeting activity peaks during the business week, as does ReTweeting activity.

Monday and Friday are both ReTweetable days, in that a higher percentage of ReTweeting activity for the week occurs than does regular Tweeting. Friday, however, shows the highest volume of ReTweeting, and Thursday the highest volume of standard Tweeting.

# About the Author



Dan Zarrella is an award-winning social, search, and viral marketing scientist and author of the upcoming O'Reilly media book "[The Social Media Marketing Book](#)."

Dan has written extensively about the science of viral marketing, memetics and social communications on his own blog and for a variety of popular industry blogs, including Mashable, CopyBlogger, ReadWriteWeb, Plagiarism Today, ProBlogger, Social Desire, CenterNetworks, Newsourcing, and SEOScoop.

He has been featured in The Twitter Book, The Financial Times, NYPost, The Boston Globe, Forbes, Wired, The Wall Street Journal, Mashable and TechCrunch. He was recently awarded Shorty and Semmy awards for social media & viral marketing.

A frequent guest speaker and panelist, Dan has spoken at PubCon, Search Engine Strategies, Convergence '09, 140 The Twitter Conference, WordCamp Mid Atlantic, Social Media Camp, Inbound Marketing Bootcamp, and The Texas Domains and Developers Conference. He currently works as an inbound marketing manager at HubSpot.

# About the Data

Over the course of 9 months, beginning in December of 2008, I've collected over 40 million ReTweets, including Tweets that contain variations of "RT," "ReTweet," and "Via." I've also collected a random sampling of over 10 million "regular" Tweets that may or may not be ReTweets. The Tweets come from Twitter's search API and its streaming API, both of which I have whitelisted access to. I use these 2 data sources and my own PHP scripts to analyze and compare the characteristics of both.