Identifying “Influencers” on Twitter

Jake Hofman, Winter Mason, Duncan Watts

Yahoo! Research

Eytan Bakshy

University of Michigan
Research in 1950’s emphasized importance of *personal* influence

- Trusted ties more important than media influence in determining individual opinions

Also found that not all people are equally influential

- A minority of “opinion leaders” or “influentials” are responsible for influencing everyone else

Call this the “influentials hypothesis”

- “One in ten Americans tells the other nine how to vote, where to eat, and what to buy.” (Keller and Berry, 2003)
• After the fact, can always tell a story about why X succeeded
  – Can identify some group of individuals who were involved early on
  – They will seem to have been influential
• But to make use of influencers, need to identify them in advance
• Very little evidence that marketers (or anyone else) can do this consistently
Why is Measuring Influence Hard?

• To measure total influence, need to observe at least
  – Who says what to whom (the network)
  – A says X, then B says X (note, this is not necessary influence)
  – Full diffusion path over many A-B pairs

• To identify influencers
  – Need to track many such diffusion paths
  – Need to correlate with information about potential influencers

• Also many possible kinds of influencers
  – Oprah, Paris, Al Gore, Jeff Jarvis, Hipsters…
  – Unclear even what to measure

• Gathering all this data has been historically impossible
Twitter Well Suited For Identifying Influencers

- Well-defined, fully-observable network of individuals who explicitly opt-in to follow each other
- Twitterers are expressly motivated to be heard
- Includes many types of potential influencers
  - Formal organizations (media, government, brands)
  - Celebrities (Ashton, Shaq, Oprah)
  - Public and Semi-Public Figures (bloggers, authors, journalists, public intellectuals)
  - Private Individuals
- Many “tweets” include unique URLs which
  - Can originate from multiple sources ("seeds")
  - Can be tracked over multiple hops ("cascades")
Twitter Project

- Team comprised Eytan Bakshy, Jake Hofman, Winter Mason
- Crawled 56M Twitter users
  - 1.7B edges
- 1B public twitter posts between 9/15/09-11/15/09
- 90M posts containing bit.ly URL’s from 1.6M users
  - We’ll focus on this subset
Computing Influence on Twitter

- An individual “seed” user tweets a URL (here we consider only bit.ly)
- For every follower who subsequently posts same URL (whether explicit “retweet” or not), seed accrues 1 pt
- Repeat for followers-of-followers, etc. to obtain total influence score for that “cascade”
  - Where multiple predecessors exist, credit first poster
  - Can also split credit or credit last poster (no big changes)
- Average individual influence score over all cascades
  - Highly conservative measure of influence, as it requires not only seeing but acting on a tweet
  - Click-through would be good, but not available to us
Cascades on Twitter

- 1.6M distinct “seeds”
- Each seed posts average of 46.3 bit.ly URL’s
- Hence 74M cascades total
- Average cascade size 1.14
  - Median cascade size 1
- Average influence score is 0.14
Most Tweets Don’t Spread

Almost all cascades are small and shallow
A tiny fraction are large and propagate up to 8 hops
Even large cascades only reach thousands
Predicting Influence

- Objective is to predict influence score for future cascades as function of
  - # Followers, # Friends, # Reciprocated Ties
  - # Tweets, Time of joining
  - Past influence score
- Fit data using regression tree
  - Recursively partitions feature space
  - Piecewise constant function fit to mean of training data in each partition
  - Nonlinear, non-parametric
    - Better calibrated than ordinary linear regression
  - Use five-fold cross-validation
    - For each fold, estimate model on training data, then evaluate on test data
    - Every user gets included in one test set
Results

- Only two features matter
  - Past local influence
  - # Followers
- Surprisingly, neither # tweets nor # following matter
- Model is well calibrated
  - average predicted close to average actual within partitions
- But fit is poor ($R^2 = 0.34$)
  - Reflects individual scatter
- Also surprisingly, content doesn’t help
Circles represent individual seeds (sized by influence)
Seeds of large cascades share certain features (e.g., high degree, past influence).

However, many small cascades share those features, making “success” hard to predict at individual level.

Common problem for rare events:
- School shootings, Plane crashes, etc.
- Tempting to infer causality from “events,” but causality disappears once non-events accounted for.

Lesson for marketers:
- Individual level predictions are unreliable, even given “perfect” information.

Fortunately, can target many seeds, thereby harnessing average effects.
• On average, some types of influencers are more influential than others
  – Many of them are highly visible celebrities, etc. with millions of followers
  – But these individuals may also be very expensive (i.e. Kim Kardashian)

• Assume the following cost function
  – $c_i = c_a + f_i^* c_f$, where $c_a = \text{acquisition cost}; c_f = \text{per-follower cost}$
  – Also $c_a = a^* c_f$, where $a$ expresses cost of acquiring individual users relative to sponsoring individual tweets

• Should you target:
  – A small # of highly influential seeds?
  – A large # of ordinary seeds with few followers?
  – Somewhere in between?
“Ordinary Influencers” Dominate

- Assume $c_f = \$0.01$
  - Equivalent to paying $\$10K$ per tweet for user with 1M followers
- When $c_a = \$1,000$, ($a = 100,000$) highly influential users are most cost effective
- But for lower ratios, most efficient choice can be individuals who influence at most one other
Broader Implications

• Obviously Twitter is a special case
  – So need to apply same method to other cases

• Nevertheless, result that large cascades are rare is probably general
  – “Social epidemics” are extremely rare
  – Probably impossible to predict them or how they will start
  – “Big seed” approach is more reliable

• “Ordinary Influencers” seem unexciting
  – Only influence one other person on average
  – But average influence is close to zero (0.28); so they’re still more influential than average
  – Combined with mass media (big seed) could be very powerful.

