“Ordinary Influencers” on Twitter

Eytan Bakshy\textsuperscript{1} Jake M. Hofman\textsuperscript{2}
Winter A. Mason\textsuperscript{2} Duncan J. Watts\textsuperscript{2}
\textsuperscript{1}University of Michigan
\textsuperscript{2}Yahoo! Research
“Influentials” and Word-Of-Mouth Marketing

- 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)
Twitter Well Suited For Identifying Influencers

- Well-defined, fully-observable network of individuals who explicitly opt-in to follow each other
- Twitter users 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”)
**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
DATA

- Crawl of Twitter follower graph:
  - 56M unique twitter users
  - 1.7B edges

- Twitter “firehose” tweet stream:
  - ~1B tweets

- Focus on bit.ly URLs
  - 87M tweets
  - 1.6M “seed” users
  - 74M diffusion events
TAKE-HOME POINTS

In general, nothing goes viral

Attributes of the user & content are related to larger cascades
- Number of followers, size of average past cascade
- Interestingness & positive feelings

But these are not sufficient conditions for large cascades

Depending on the cost function of targeting users, casting a wide net may be more efficient than targeting “influencers”
CASCADES ON TWITTER

- 1.6M distinct “seeds”
- Each seed posts average of 46.3 bit.ly URL’s
- 74M cascades total
- Mean cascade size 1.14
  - Median cascade size 1
- Mean influence score 0.14
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
RESULTS

- Model is well calibrated
  - average predicted close to average actual within partitions
- But fit is poor ($R^2 = 0.34$)
  - Reflects individual scatter
WHO ARE THE INFLUENCERS?

Circles represent individual seeds (sized by influence)
Sampled 1000 URLs, had workers on AMT classify URLs
- Spam / Not spam (795 good URLs)
- Type of URL
- General Category
- Interestingness
- Positive feeling towards URL
THE ROLE OF CONTENT

URLs rated as more interesting or evoking more positive emotions have larger cascades

(Akin to Berger & Milkman, 2010)
**The Role of Content**

- Surprisingly, content does not matter relative to user features.

- Even with content, fit is poor ($R^2 = 0.31$)
  - Much smaller subset
NECESSARY BUT NOT SUFFICIENT

- 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
COST-EFFECTIVENESS OF TARGETING STRATEGIES

 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

 Assume the following cost function
  • \( c_i = c_a + f_i \cdot c_f \), where \( c_a \) = acquisition cost; \( c_f \) = per-follower cost
  • Also \( c_a = \alpha \cdot c_f \), where \( \alpha \) 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
- When $c_a \leq $100, ($a = 10,000$), most efficient choice are low-influence users
**Broader Implications**

- **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
  - Difficult 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 could be very powerful.
THANK YOU!

Questions?