Identifying social media profession influence ranking in online social network

Presented by Jiajun Xin
Outline

- Background Introduction
- Related work
- My ideas
Background Introduction

• Directed link network

• Undirected link network
Related work

• Everyone’s an influencer: quantifying influence on twitter
• Identifying influential and susceptible members of social networks
• What is twitter, a social network or a news media?
• Twitterrank: finding topicsensitive influential twitterers
• Identifying topical authorities in microblogs
Everyone’s an influencer: quantifying influence on twitter

By Eytan Bakshy, Winter Mason

Word-of-mouth diffusion has long been regarded as an important mechanism by which information can reach large populations, possibly influencing public opinion, adoption of innovations, new product market share, or brand awareness.
Everyone’s an influencer: quantifying influence on twitter

• Dataset 1:

• Over the two-month period of September 13 2009 - November 15 2009 we recorded all 1.03B public tweets broadcast on Twitter.

• Then we extracted 87M tweets that included bit.ly URLs and which corresponded to distinct diffusion “events”.

• Finally, they identified a subset of 74M diffusion events that were initiated by seed users who were active in both the first and second months of the observation period.

• Each seed user seeded an average of 46.33 twitters.
Everyone’s an influencer: quantifying influence on twitter

• Dataset 2:
  • Crawled the follower graph of users who broadest at least one URL over the same time.
  • Contains 56M users and 1.7B edges.

![Table 1: Statistics of the Twitter follower graph and seed activity](image)
Everyone’s an influencer: quantifying influence on twitter

“Influence” means a user’s ability to seed content containing URLs that generate large cascades of reposts.

To calculate the influence score for a given URL post, we tracked the diffusion of the URL from its origin at a particular “seed” node through a series of reposts—by that user’s followers, those users’ followers, and so on—until the diffusion event, or cascade, terminated.
Everyone’s an influencer: quantifying influence on twitter

As the figure shows, the distribution of cascade sizes is approximately power-law, implying that the vast majority of posted URLs do not spread at all (the average cascade size is 1.14 and the median is 1), while a small fraction are reposted thousands of times.

Vertical axis stands for frequency distribution of cascade sizes.
Everyone’s an influencer: quantifying influence on twitter

The depth of the cascade is also right skewed, but more closely resembles an exponential distribution, where the deepest cascades can propagate as far as nine generations from their origin; but again the vast majority of URLs are not reposted at all, corresponding to cascades of size 1 and depth 0 in which the seed is the only node in the tree.
Past local influence stands for reposts by direct followers.

Vertical lines represent one standard deviation above and below.
(a) All users

(b) Top 25 users
Figure 8: (a). Average cascade size for different types of URLs (b). Average cascade size for different categories of content. Error bars are standard errors.
Everyone’s an influencer: quantifying influence on twitter

• Conclusion: Marketers, planners and other change agents interested in harnessing word-of-mouth influence could therefore benefit

• (1) by adopting more precise metrics of influence;

• (2) by collecting more and better data about potential influencers over extended intervals of time;

• (3) by potentially exploiting ordinary influencers, where the optimal tradeoff between the number of individuals targeted and their average level of influence will depend on the specifics of the cost function in question.
Identifying influential and susceptible members of social networks

By Sinan Aral, Dylan Walker

• Data:

• They conducted a randomized experiment to measure influence and susceptibility to influence in the product adoption decisions of a representative sample of 1.3 million Facebook users.

• Methodology:

• The experiment involved the random manipulation of influence-mediating messages sent from a commercial Facebook application that lets users share information and opinions about movies and so on. As users adopted and used the product, automated notifications of their activities were delivered to randomly selected peers in their local social networks. Because message recipients were randomly selected, treated and untreated peers of the application user differed only by the number of randomized messages they received.
Influence (dark gray) and susceptibility to influence (light gray) are shown with SEs (boxes) and 95% confidence intervals.
What is twitter, a social network or a news media?

By Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon

• This paper provides three ways to quantize the influence of users on microblogs and discuss the difference between the three measures.
• The first way is by PageRank algorithm which is widely used to measure the importance of a node in a graph.
• The second way is by retweets. Counting retweets is a way to measure the influence of one certain incident which implies the influence of the owner of this tweet.
• The third way is to measure the rankings of users followers, PageRank and number of retweets.
• The result suggests the rankings of users followers and rankings of PageRank returns similar result. But the ranking of retweets is different from the other two methods.
<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Name</th>
<th>Remark</th>
<th>Rank</th>
<th>ID</th>
<th>Name</th>
<th>Remark</th>
<th>Rank</th>
<th>ID</th>
<th>Name</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aplusk</td>
<td>ashton kutcher</td>
<td>actor</td>
<td>1</td>
<td>aplusk</td>
<td>ashton kutcher</td>
<td>actor</td>
<td>1</td>
<td>mashable</td>
<td>Pete Cashmore</td>
<td>news on social media</td>
</tr>
<tr>
<td>2</td>
<td>britney</td>
<td>Britney Spears</td>
<td>musician</td>
<td>2</td>
<td>Barack Obama</td>
<td>Barack Obama</td>
<td>president of U.S.</td>
<td>2</td>
<td>BreakingNews</td>
<td>ESPN News</td>
<td>news</td>
</tr>
<tr>
<td>3</td>
<td>TheEllenShow</td>
<td>Ellen DeGeneres</td>
<td>show host</td>
<td>3</td>
<td>cnnbrk</td>
<td>CNN Breaking News</td>
<td>news</td>
<td>3</td>
<td>tweetmeme</td>
<td>TweetMeme</td>
<td>news on Twitter</td>
</tr>
<tr>
<td>4</td>
<td>cnnbrk</td>
<td>CNN Breaking News</td>
<td>news</td>
<td>4</td>
<td>TheEllenShow</td>
<td>Ellen DeGeneres</td>
<td>show host</td>
<td>4</td>
<td>oxfordgirl</td>
<td>oxfordgirl</td>
<td>journalist</td>
</tr>
<tr>
<td>5</td>
<td>Oprah</td>
<td>Oprah Winfrey</td>
<td>show host</td>
<td>5</td>
<td>twitter</td>
<td>Twitter</td>
<td>subject of this paper</td>
<td>5</td>
<td>TechCrunch</td>
<td>Michael Arrington</td>
<td>news on technology</td>
</tr>
<tr>
<td>6</td>
<td>RyanSeacrest</td>
<td>Ryan Seacrest</td>
<td>show host</td>
<td>6</td>
<td>THE_REAL_SHAQ</td>
<td>THE_REAL_SHAQ</td>
<td>sports star</td>
<td>6</td>
<td>myfabulouslife</td>
<td>Fabulous</td>
<td>musician</td>
</tr>
<tr>
<td>7</td>
<td>THE_REAL_SHAQ</td>
<td>THE_REAL_SHAQ</td>
<td>sports star</td>
<td>7</td>
<td>johnmayer</td>
<td>John Mayer</td>
<td>musician</td>
<td>7</td>
<td>nytimes</td>
<td>The New York Times</td>
<td>news</td>
</tr>
<tr>
<td>8</td>
<td>KimKardashian</td>
<td>Kim Kardashian</td>
<td>model</td>
<td>8</td>
<td>lanseacrest</td>
<td>Lance Armstrong</td>
<td>sports star</td>
<td>8</td>
<td>ildoyal</td>
<td>ildoyal</td>
<td>comedian</td>
</tr>
<tr>
<td>9</td>
<td>John Mayer</td>
<td>musician</td>
<td></td>
<td>9</td>
<td>jimmyfallon</td>
<td>Jimmy Fallon</td>
<td>actor</td>
<td>9</td>
<td>iranrigged</td>
<td>IranriggedElect</td>
<td>Iran about Iran</td>
</tr>
<tr>
<td>10</td>
<td>Jimmy Fallon</td>
<td>Jimmy Fallon</td>
<td>actor</td>
<td>10</td>
<td>jimmyfallon</td>
<td>Jimmy Fallon</td>
<td>actor</td>
<td>10</td>
<td>espn</td>
<td>ESPN Sports News</td>
<td>news</td>
</tr>
<tr>
<td>11</td>
<td>Demi Moore</td>
<td>actress</td>
<td></td>
<td>11</td>
<td>jimmyfallon</td>
<td>Jimmy Fallon</td>
<td>actor</td>
<td>11</td>
<td>persiankivi</td>
<td>persiankivi</td>
<td>about Iran</td>
</tr>
<tr>
<td>12</td>
<td>lamsydey</td>
<td>lamsydey</td>
<td>musician</td>
<td>12</td>
<td>lamsydey</td>
<td>lamsydey</td>
<td>musician</td>
<td>12</td>
<td>aplusk</td>
<td>ashton kutcher</td>
<td>actor</td>
</tr>
<tr>
<td>13</td>
<td>lanseacrest</td>
<td>Lance Armstrong</td>
<td>sports star</td>
<td>13</td>
<td>lanseacrest</td>
<td>Lance Armstrong</td>
<td>sports star</td>
<td>13</td>
<td>StopAhmadi</td>
<td>Raymond Jahan</td>
<td>about Iran</td>
</tr>
<tr>
<td>14</td>
<td>PerezHilton</td>
<td>Perez Hilton</td>
<td>power blogger</td>
<td>14</td>
<td>PerezHilton</td>
<td>Perez Hilton</td>
<td>power blogger</td>
<td>14</td>
<td>AlyssaMilano</td>
<td>Alyssa Milano</td>
<td>actress</td>
</tr>
<tr>
<td>15</td>
<td>Al Gore</td>
<td>politician</td>
<td></td>
<td>15</td>
<td>nytimes</td>
<td>The New York Times</td>
<td>news</td>
<td>15</td>
<td>huffingtonpost</td>
<td>HuffingtonPost.com</td>
<td>news</td>
</tr>
<tr>
<td>16</td>
<td>Miley Cyrus</td>
<td>Miley Cyrus</td>
<td>actress /musician</td>
<td>16</td>
<td>mileycyrus</td>
<td>Miley Cyrus</td>
<td>actress /musician</td>
<td>16</td>
<td>jamielee</td>
<td>jamielee</td>
<td>musician</td>
</tr>
<tr>
<td>18</td>
<td>Coldplay</td>
<td>musician</td>
<td></td>
<td>18</td>
<td>norooms</td>
<td>NPR News</td>
<td>news</td>
<td>18</td>
<td>norooms</td>
<td>NPR News</td>
<td>news</td>
</tr>
<tr>
<td>19</td>
<td>The Onion</td>
<td>The Onion</td>
<td>news</td>
<td>19</td>
<td>PerezHilton</td>
<td>Perez Hilton</td>
<td>power blogger</td>
<td>19</td>
<td>PerezHilton</td>
<td>Perez Hilton</td>
<td>power blogger</td>
</tr>
</tbody>
</table>
Twitterrank: finding topicsensitive influential twitterers

By Jianshu Wang, Ee Peng Lim, Jing Jiang, Qi He

They believe the influence of different users is biased because of their interest, which indicates the influence can only be passed to the follower only if the topic is within the shared interest between the user and the follower.

Based on this idea, they provide an algorithm to quantize one user’s influence on a certain topic and an algorithm to aggregation of users influence on all topics.
Twitterrank: finding topicsensitive influential twitterers

Figure 3: Distribution of Tweets per Twitterer
Twitterrank: finding topic-sensitive influential twitterers

- 72.4% of the twitterers follow more than 80% of their followers,

- and 80.5% of the twitterers have 80% of their friends follow them back.

Figure 5: Number of Friends vs. Number of Followers
Definition 2. Given a topic $t$, each element of matrix $P_t$, i.e. the transition probability of the random surfer from follower $s_i$ to friend $s_j$, is defined as:

$$P_t(i, j) = \frac{|\mathcal{T}_j|}{\sum_{a: s_i \text{ follows } s_a} |\mathcal{T}_a|} \ast \text{sim}_t(i, j)$$

($3$)

$|\mathcal{T}_j|$ is the number of tweets published by $s_j$, and $\sum_{a: s_i \text{ follows } s_a} |\mathcal{T}_a|$ sums up the number of tweets published by all of $s_i$'s friends. $\text{sim}_t(i, j)$ in Eq. (3) is the similarity between $s_i$ and $s_j$ in topic $t$, which is defined as:

$$\text{sim}_t(i, j) = 1 - |DT'_{it} - DT'_{jt}|$$

($4$)

Definition 3. The teleportation vector of the random surfer in topic $t$ is defined as:

$$E_t = DT''_t$$

($5$)

$DT''_t$ is the $t$-th column of matrix $DT''$, which is the column-normalized form of matrix $DT$ such that $||DT''_t||_1 = 1$. $DT$ is one of the results obtained during the topic distillation, each entry of which contains the numbers of times words in a twitterer's tweets has been assigned to a specific topic.

With the transition probability matrix and teleportation vector defined, the topic-specific TwitterRank can be calculated.

Definition 4. The topic-specific TwitterRank of the twitterers in topic $t$, denoted as $\overrightarrow{TR}_t$, can be calculated iteratively by:

$$\overrightarrow{TR}_t = \gamma P_t \times \overrightarrow{TR}_t + (1 - \gamma)E_t$$

($6$)

$P_t$ is the transition probability matrix defined in Eq. (3), $E_t$ is the teleportation vector defined in Eq. (5). $\gamma$ is a parameter between $0$ and $1$ to control the probability of teleportation. The lower $\gamma$ is, the higher probability the random surfer will teleport to twitterers according to $E_t$, and vice versa.
Identifying topical authorities in microblogs

By Aditya Pal, Scott Counts

• They first propose a set of features to characterize users.
• They propose an algorithm to identify the most influential users within the topic.
Identifying topical authorities in microblogs

- **OT (Original tweet)**: These are the tweets produced by the author that are not *RT* or *CT*.

- **CT (Conversational tweet)**: Conversational tweet is directed at another user, as denoted by the use of the @username token preceding the text or from the meta-data available through the Twitter API.

- **RT (Repeated tweet)**: These tweets are produced by someone else but the user copies, or forwards, them in-order to spread it in her network. These tweets are preceded by “RT @username”.

<table>
<thead>
<tr>
<th>ID</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT1</td>
<td>Number of original tweets</td>
</tr>
<tr>
<td>OT2</td>
<td>Number of links shared</td>
</tr>
<tr>
<td>OT3</td>
<td>Self-similarity score that computes how similar is author's recent</td>
</tr>
<tr>
<td></td>
<td>tweet w.r.t. to her previous tweets</td>
</tr>
<tr>
<td>OT4</td>
<td>Number of keyword hashtags used</td>
</tr>
<tr>
<td>CT1</td>
<td>Number of conversational tweets</td>
</tr>
<tr>
<td>CT2</td>
<td>Number of conversational tweets where conversation is initiated by</td>
</tr>
<tr>
<td></td>
<td>the author</td>
</tr>
<tr>
<td>RT1</td>
<td>Number of retweets of other's tweet</td>
</tr>
<tr>
<td>RT2</td>
<td>Number of unique tweets (OT1) retweeted by other users</td>
</tr>
<tr>
<td>RT3</td>
<td>Number of unique users who retweeted author's tweets</td>
</tr>
<tr>
<td>M1</td>
<td>Number of mentions of other users by the author</td>
</tr>
<tr>
<td>M2</td>
<td>Number of unique users mentioned by the author</td>
</tr>
<tr>
<td>M3</td>
<td>Number of mentions by others of the author</td>
</tr>
<tr>
<td>M4</td>
<td>Number of unique users mentioning the author</td>
</tr>
<tr>
<td>G1</td>
<td>Number of topically active followers</td>
</tr>
<tr>
<td>G2</td>
<td>Number of topically active friends</td>
</tr>
<tr>
<td>G3</td>
<td>Number of followers tweeting on topic after the author</td>
</tr>
<tr>
<td>G4</td>
<td>Number of friends tweeting on topic before the author</td>
</tr>
</tbody>
</table>

Table 1: List of metrics of potential authorities. OT = Original tweets, CT = Conversational tweets, RT = Repeated tweets, M = Mentions, and G = Graph Characteristics.

Topical signal \((TS)\) = \(\frac{OT1 + CT1 + RT1}{\# \text{ tweets}}\)

Signal strength \((SS)\) = \(\frac{OT1}{OT1 + RT1}\)

Non-Chat signal \((\bar{S})\) = \(\frac{OT1}{OT1 + CT1 + \lambda \frac{CT1 - CT2}{CT1 + 1}}\)

Retweet impact \((RI)\) = \(RT2 \cdot \log(RT3)\)

Mention impact \((MI)\) = \(M3 \cdot \log(M4) - M1 \cdot \log(M2)\)
My idea

• Connecting the real world with the online social network by professions. Finding the influence ranking of different professions.
• Finding the influence ranking of different fields. For example, finance or computer science? Which is more popular?
Thanks for listening! Any suggestions?