Finding Trendsetters in Information Networks

Diego Saez-Trumper\textsuperscript{1} \quad Giovanni Comarela\textsuperscript{2}

Virgilio Almeida\textsuperscript{2} \quad Ricardo Baeza-Yates\textsuperscript{3} \quad Fabricio Benevenuto\textsuperscript{4}

\textsuperscript{1}Universitat Pompeu Fabra, Barcelona
\textsuperscript{2}Universidade Federal Minas Gerais, Brazil
\textsuperscript{3}Yahoo! Research, Barcelona
\textsuperscript{4}Universidade Federal de Ouro Preto, Brazil

Beijing, August, 2012
What is a Trendsetter?
What is a Trendsetter?

Trendsetters are people:

- Adopt and spread new trends before these trends become popular.
- Propagate these trends over the network.

<table>
<thead>
<tr>
<th></th>
<th>node indegree</th>
<th>threshold for adoption of new ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>follower hubs</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>innovation hubs</td>
<td>lower</td>
<td>low</td>
</tr>
</tbody>
</table>
Finding trendsetters in a graph
TIME!
Time

Am I Trendsetter?

$A_{t_A} = 3$

I'm Late Adopter

$E_{t_E} = 8$

Am I Trendsetter?

$B_{t_B} = 5$

$I'm$ just an Early Adopter

$G_{t_G} = 1$

$C_{t_C} = 4$

$D_{t_D} = 6$

$F_{t_F} = 7$

$H_{t_G} = 2$
How to find Trendsetters?
Weight edges and run PageRank
Topics and Influence Model
Topic: collection of trends (Urls, memes, #hashtags, quotes, etc)

We denote this set by a vector: \( \{ h_1, \ldots, h_n \} \).

For each node we store the timestamp when she adopt a trend \( h_1 \).
Weight Edges

\[ s_1(v)_i = \begin{cases} 
1, & \text{if } t_i(v) > 0, \\
0, & \text{otherwise} 
\end{cases} \quad (1) \]

\[ s_2(u, v)_i = \begin{cases} 
e^{-\frac{\Delta}{\alpha}}, & \text{if } t_i(v) > 0 \text{ and } t_i(v) < t_i(u), \\
0, & \text{otherwise} 
\end{cases} \quad (2) \]

for \( i = 1, \ldots, n_k \), where \( \Delta = t_i(u) - t_i(v) \) and \( \alpha > 0 \).
Weight Edges

\[ v_t = [1, -8] \quad u_t = [2, 5, 3] \]

\[ s_1(v) = [1, 0, 1] \quad s_2(u, v) = [e^{-\frac{4-1}{\alpha}}, 0, 0] \]

Next,

\[ I_k^*(u, v) = \left( \frac{s_1(v) \cdot s_2(u, v)}{\|s_1(v)\| \times \|s_2(u, v)\|} \right) \times \left( \frac{L(s_2(u, v))}{n_k} \right) \]
The trendsetters (TS) rank of node $v$ in a network $G_k(N_k, E_k)$, denoted by $TS_k(v)$, is given by:

$$TS_k(v) = d \cdot D_k(v) + (1 - d) \sum_{w \in \text{In}_{G_k}(v)} TS_k(w)l_k(w, v),$$

(4)
Evaluation
- In-degree ranking
- PageRank
Dataset

- Twitter until August 2009.
- Over 50 Millions users with all their *followers* and *followees*.
- 1.6 Billions tweets
- We use #tags as trends.
Example:
Iran Elections on Twitter
例：伊朗选举：#{iran, iranelections, tehran}

- TS：@Lara（“报道来自中东”）
- PR：@cnnbr（“CNN新闻”）
We use the #tag classification made by Romero et al.

<table>
<thead>
<tr>
<th>Category</th>
<th>#Topics</th>
<th>Example of Hashtags</th>
<th>#Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celebrity</td>
<td>16</td>
<td>#michaeljackson, #niley</td>
<td>1,036,101</td>
</tr>
<tr>
<td>Games</td>
<td>13</td>
<td># mafiawars, #ps3 #</td>
<td>2,556,437</td>
</tr>
<tr>
<td>Idioms</td>
<td>35</td>
<td>#musicmonday, #followfriday</td>
<td>7,882,209</td>
</tr>
<tr>
<td>Movies</td>
<td>29</td>
<td>#heroes, #tv</td>
<td>1,769,945</td>
</tr>
<tr>
<td>Music</td>
<td>33</td>
<td>#lastfm, #musicmonday</td>
<td>2,785,522</td>
</tr>
<tr>
<td>None</td>
<td>153</td>
<td>#quotes, #sale</td>
<td>2,227,971</td>
</tr>
<tr>
<td>Political</td>
<td>39</td>
<td>#honduras, #Iranelection,</td>
<td>8,156,786</td>
</tr>
<tr>
<td>Sports</td>
<td>27</td>
<td>#soccer, #rugby</td>
<td>1,914,061</td>
</tr>
<tr>
<td>Technology</td>
<td>41</td>
<td>#twitter, #android</td>
<td>7,459,471</td>
</tr>
</tbody>
</table>
Trendsetters: early adopters?
Experiments I

---

**Diagram Description:**

The diagram illustrates the percentage of Top-100 users before the peak across different categories. The x-axis represents the categories, including 'CELEBRITY', 'MUSIC', 'IDIOMS', 'GAMES', 'POLITICAL', 'NONE', 'MOVIES', 'TECHNO.', and 'SPORTS', in that order. The y-axis shows the percentage of Top-100 users.

Three different metrics are plotted: InDegree (black squares), PageRank (red triangles), and TrendSetters (green circles). Each metric shows a trend indicating the percentage of users before the peak for each category.

- **InDegree** appears to have the highest percentage across most categories, except for 'MOVIES' and 'TECHNO.'.
- **PageRank** shows a lower percentage than InDegree, indicating a smaller contribution of Top-100 users.
- **TrendSetters** have the lowest percentage, suggesting the least influence.

**Graph Analysis:**

- The 'CELEBRITY' category shows a peak around the 50% mark, indicating a significant contribution of Top-100 users.
- 'MUSIC' and 'IDIOMS' categories follow with a similar trend, hovering slightly below the peak of 'CELEBRITY'.
- 'GAMES' and 'POLITICAL' categories show a steady increase but remain below the peak.
- 'NONE', 'MOVIES', and 'TECHNO.' categories have a more gradual increase, with 'TECHNO.' showing the least increase.
- 'SPORTS' category shows a moderate increase, indicating a moderate contribution of Top-100 users.

---

**Table:**

<table>
<thead>
<tr>
<th>Category</th>
<th>InDegree</th>
<th>PageRank</th>
<th>TrendSetters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEBRITY</td>
<td>100</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>MUSIC</td>
<td>90</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>IDIOMS</td>
<td>80</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>GAMES</td>
<td>70</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>POLITICAL</td>
<td>60</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>NONE</td>
<td>50</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>MOVIES</td>
<td>40</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>TECHNO.</td>
<td>30</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>SPORTS</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Note:**

The exact percentage values are approximations based on the graph's visual representation. The table provides a summary of the trends observed in the diagram.
In-degree vs adoption time
Influenced Followers Ratio
Influenced Followers Ratio

$IF_k(v)$ is the fraction of followers of $v$ that adopted at least one trend of the topic $k$ after $v$.

<table>
<thead>
<tr>
<th>Category</th>
<th>(%)ID</th>
<th>(%)PR</th>
<th>(%)TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLITICAL</td>
<td>0.013</td>
<td>0.084</td>
<td>0.174</td>
</tr>
<tr>
<td>CELEBRITY</td>
<td>0.015</td>
<td>0.089</td>
<td>0.148</td>
</tr>
<tr>
<td>MUSIC</td>
<td>0.013</td>
<td>0.096</td>
<td>0.160</td>
</tr>
<tr>
<td>GAMES</td>
<td>0.022</td>
<td>0.058</td>
<td>0.115</td>
</tr>
<tr>
<td>SPORTS</td>
<td>0.004</td>
<td>0.054</td>
<td>0.098</td>
</tr>
<tr>
<td>IDIOMS</td>
<td>0.001</td>
<td>0.034</td>
<td>0.088</td>
</tr>
<tr>
<td>NONE</td>
<td>0.011</td>
<td>0.001</td>
<td>0.085</td>
</tr>
<tr>
<td>TECHNOLOGY</td>
<td>0.006</td>
<td>0.054</td>
<td>0.078</td>
</tr>
<tr>
<td>MOVIES</td>
<td>0.006</td>
<td>0.043</td>
<td>0.067</td>
</tr>
</tbody>
</table>
Ranking with Partial Information

**Graph Description:**
- The graph shows the number of Top-100 users found against the ratio of users considered (sorted by time).
- Different lines represent different topics or sets:
  - musicmonday TS
  - musicmonday PR
  - iranelection TS
  - iranelection PR
  - swineflue TS
  - swineflue PR
  - followfriday TS
  - followfriday PR
  - mw2 TS
  - mw2 PR
  - fb TS
  - fb PR
  - f1 TS
  - f1 PR
  - michaeljackson TS
  - michaeljackson PR

**Axes:**
- **X-axis:** Ratio of users considered (sorted by time)
- **Y-axis:** Number of Top-100 users found
Final Remarks

- Usually, follower hubs (celebrities) are late adopters.
- Trendsetters have lower in-degree, but they spread new ideas.
- Code available (diego.saez [at] upf.edu)
- Implementation in python + C
Questions
We also group the topics by shape, using the KSC algorithm.