We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter

Brendan Meeder, Brian Karrer, Amin Sayedi, R. Ravi, Christian Borgs, Jennifer Chayes
Motivation

- Information about Twitter can be gathered from the open Twitter API.

- Twitter does not provide the time when a user starts following another.

- Crawling the social graph of Twitter is time-consuming.
Questions

● How does the rate of accumulation of followers change over time?

● What are the key factors that influence these changes?

● What is the pattern of users following celebrities in relation to their account creation times?
Users in Followers List are returned in reverse order in which they followed that user.
Inferring Edge Creation Times

- **C_u**: Account creation time
- **F_u**: Time user starts following (unknown)
  \[ C \leq F \]
- **B(u)**: set of users that appears before \( u \)
  \[ F_v \leq F_u \]
- All \( v \) in \( B(u) \) provide a lower bound for \( F_u \)
  \[ C_v \leq F_v \leq F_u \]
- Estimation for \( F_u \): greatest lower lower bound
  \[ \hat{F}_u = \max_{v \in B(u)} C_v. \]
Inferring Edge Creation Times

<table>
<thead>
<tr>
<th>Followers List</th>
<th>User X</th>
<th>&lt;=</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(c)</td>
<td>User A</td>
<td>Ca</td>
</tr>
<tr>
<td></td>
<td>User B</td>
<td>Cb</td>
</tr>
<tr>
<td></td>
<td>User C</td>
<td>Cc</td>
</tr>
<tr>
<td></td>
<td>User D</td>
<td>Cd</td>
</tr>
<tr>
<td></td>
<td>User E</td>
<td>Ce</td>
</tr>
<tr>
<td>B(c)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time

MAX

B(c)

E_c

F_a

F_b

F_c

F_d

F_e

C_c

C_d

C_e

C_b

C_a

C_c

C_d

C_e
Theoretical Analysis

Proposition 1. Let $\epsilon > 0$. If

$$
\left( \int_{\delta/2}^{\infty} \ell(t) dt \right)^{\delta/(2\lambda)} \leq \epsilon,
$$

then $P(F_u - \hat{F}_u > \delta) < \epsilon$ for any user $u$.

- If the rate of new user arrival for a celebrity is high, then the error in the inferred follow times will be small.
Empirical Validation

- Focus on users that gain followers at a high rate.

- Gathered ordered follower list from "celebrity" users.
  - Top 1000 celebrities from Twitaholic.com
  - Users on the suggested user list

- Total: 1,800 users.
Evaluating timestamp errors

- Crawl of all 1,800 celebrities:
  - Every 30 minutes
  - For 220 hours (~10 days)
  - Most recent 5,000 followers

- Total: 23,258,723 follow events.

- Evaluate the upper bound error of estimated F.
Celebrity Follow Event Errors

Error (seconds)

Number of New Followers

- Maximum error (≥ 10 RBUs)
- Mean error (≥ 10 RBUs)
- Maximum error (< 10 RBUs)
- Mean error (< 10 RBUs)
Historical Accuracy

- Evaluate the upper bound error of record-breaker difference.

- Filter
  1. All record-breaker users that are created less than 24 hours before the next record-breaker are declared accurate.
  2. Non-RB user between two RB is accurate if the later RB is accurate and created their account less than 4 hours after the earlier RB.

- Accurate celebrity = contains 95% accurate timestamps

- Total: 1508 accurate celebrities
Broad analysis of celebrity subgraph

- 74,184,348 nodes (Twitter ~ 190 million)
- 835,117,954 edges (Twitter ~ 7 billion)

- 20% of the accurate celebrities have more than a million followers.

- Peaks of following k celebrities
  - 20 (size of initial suggested user list)
  - 241 (number of users available to be suggested)
  - 461 (?)
Broad analysis of celebrity subgraph

- Accurate Celebrity Follow and Account Creation Rates.

- Adjustments to Twitter's user interface:
  1. Introduction of the suggested users list
     Feb. 13, 2009
  2. Suggested user list based on categories
     Jan. 21, 2010
  3. Introduction of "users you may be interested in" feature
     July 30, 2010
Accurate Celebrity Follow and Account Creation Rates

- Hourly follow rate
- Four day average hourly follow rate
- Daily account creation rate

Follow events (per hour), account creations (per day)

Date:
- Jan 2009
- Apr
- July
- Oct
- Jan 2010
- Apr
- July
- Oct

(1) (2) (3)
Fraction of Celebrity Follow Events Over Time

- Red: Anil Dash (@anildash)
- Green: Kim Kardashian (@kimkardashian)
- Blue: New York Times (@nytimes)

Events:
1. (1) Significant increase in followers in April 2009
2. (2) Sharp increase in followers in October 2009
3. (3) Increase in followers in July 2010
Measuring following latency

- Conditional probability that a user waits $t$ seconds to follow the celebrity given that they follow the celebrity within a month of account creation.
- 0 latency is caused by record-breaker.
- 86% follow within 24 hours.
- Fraction of followers who followed the celebrity within a month:
  - Mean: 65%
  - STD: 18%
Celebrity popularity and real-world events

\[ f_i(t) = \frac{|\text{Connections to } i \text{ within } t - \Delta \text{ and } t + \Delta|}{|\text{Edges created within } t - \Delta \text{ and } t + \Delta|} \]

- \( \Delta \): sliding window width
- \( \Delta = 1 \text{ week, } t \text{ in days} \)
- Baseline = \( 1/n(t) \), number of celebrities
Conclusion

• Simple and effective method for inferring follow times using only a single crawl and user creation times.

• Accurate to within several minutes for popular users.

• Deeper insight into the structure and evolution of a significant and large subgraph of the Twitter social network.
Thoughts

- Evaluated just "celebrities".

- Error decreases as the following rate increases.

- May work just for popular users, not being a global method for all users.