Outwitting the Twitterers – Predicting Information Cascades in Microblogs

Wojciech Galuba, Karl Aberer
EPFL, Switzerland

Dipanjan Chakraborty
IBM Research India

Zoran Despotovic, Wolfgang Kellerer
Docomo Euro-Labs, Munich, Germany
Why study information flows in OSNs?

- casual link sharing
- breaking news
- activism
- viral marketing
- emergencies
- PR campaigns

**Modeling**

- improve how information flows
- new applications
- insights into underlying sociology
Information overload?

Median: 23 tw/h, 552 tw/day

Full-time job (reading tweets 40h a week at 150WPM)

(Sep 2009 data)
OSN information spread modeling

Related work:
- generative models
  - reproduce statistical properties of info spread
- predict coarse-grained aggregates
  - # of nodes reached by spread etc.

Our approach:
- Look at URL diffusion on Twitter
- Can we predict which **user** will mention which **URL** with what **probability**?
Why predict URL tweets?

- Protect from information overload
  - Sort incoming URLs by probability of retweeting
- Viral marketing
  - Select a subset of users that ensure successful URL propagation
- Spam detection
  - Mispredictions are a sign of anomalous activity
Realtime results for http

**taksilover** RT @taksilover HAAHAHA OF FENNE STUURT JE FF BABYFOTO VAN TREY SONGZ HAAHAHA < http://bit.ly/bkoOFY
less than 10 seconds ago from web

**its_shauny_yo** So beautiful imy !RT @MrsPinkylvory: http://twitpic.com/1o1fnm
less than 10 seconds ago from UberTwitter

**dominos_JP** やや重くてすみません。充実の動画でして・・・。RT @mitsuyamarines http://tl.gd/1af1mc
less than 10 seconds ago from TwitBird iPhone

**soro09** @NMANUELX_x 1235 firmas por la libertad de los presos políticos venezolanos Necesitamos tu apoyo http://bit.ly/cxRjjH
less than 10 seconds ago from web

**CisaOficial** RT @MaiteOficial: Para que se den una idea este fue mi postre ayer... Mmmmmmm buenísimo http://twitpic.com/1o1gnp
less than 10 seconds ago from Twitpic

**Taigenz** Q:Girl or boy? A:Lol http://formspring.me/TaigenzB/q/549093468
less than 10 seconds ago from formspring.me
Data

- 300 hour window in Sep’09
- 22M tweets
- 2.7M unique users
- 15M unique URLs
- 700M connections in the follower graph
- Approx. 1/15th of the Twitter traffic
Follower graph*

* active users only: that have sent at least one URL in 300h
Follower graph*

Mean (directed): 3.61

* active users only: that have sent at least one URL in 300h
User activity

![Graph showing the distribution of user activity with a log-log scale, indicating a power-law relationship between the number of users and the count of tweets and unique URLs.](image-url)
Per-URL activity

![Graph showing the distribution of #tweets and #unique users across different counts.]
Information cascades

Nodes: users that mentioned a given URL
Arcs: information flow
Re-tweeting

Space Shuttle Atlantis lifts off for final scheduled mission. http://on.cnn.com/cBDQEk
about 23 hours ago via web

chaunce322: RT @cnnbrk Space Shuttle Atlantis lifts off for final scheduled mission. http://on.cnn.com/cBDQEk
about 18 hours ago from Twitterrific · Reply · View Tweet
RT-cascade

@alice: http://url.com

@bob: RT @alice
http://url.com

@charlie: http://url.com

- Arcs: who retweets whom
  - Irrespective of whether users follow one another

- Single parent
  - Only the user name immediately after „RT” taken into account
F-cascade

@alice: http://url.com

@bob: http://url.com

@charlie: http://url.com

Arc @a→@b exists if:

- user @a mentioned URL before user @b
- user @b follows user @a
RT-cascades vs. F-cascades

- RT-cascades are trees
- F-cascades are DAGs
- 33% of the retweets credit a source that the user does not directly follow
Subcascade size
Cascade fragmentation

![Graph showing cascade fragmentation](image-url)
Cascade depth

![Graph showing cascade depth distribution with different markers indicating average (RT-cascades), maximum (RT-cascades), average (F-cascades), and maximum (F-cascades)]
Influence of the root

![Graph showing the median subcascade size vs. the number of followers of subcascade root for RT-cascades (red) and F-cascades (blue).]
Information diffusion rate

Median: 50mins
URL tweeting prediction

- Based on the past URL retweets by users, predict the future ones
- Find probability that user $i$ mentions URL $u$

$$p_i^u = ?$$
Influence
URL virality

\( \gamma_u \)

http://cnn.com/
Per-user diffusion delay

\[ \mu_i, \sigma_i^2 \]
Model

\[ \alpha_{ij} \]

\[ \beta_i \]

\[ \gamma_u \]

http://cnn.com/

\[ \mu_i, \sigma_i^2 \]
At-Least-One (ALO) model

\[ p_{i}^{u} = \alpha_{ij} \gamma_{u} p_{j}^{u} + \beta_{i} \gamma_{u} \]

\[ P(\text{at least one event happens}) \]

* Temporal component \( \mu_{i}, \sigma_{i}^{2} \)
Linear threshold (LT) model

\[ p_{i}^{u} = \sum_{j} \alpha_{ij} \gamma_{u} p_{j}^{u} + \beta_{i} \gamma_{u} \]

Thresholding function (sigmoid)

Temporal component \( \mu_{i}, \sigma_{i}^{2} \)
Performance metrics

- **Recall**: fraction of tweets predicted
  - out of all tweets that happened
- **Precision**: fraction of true positives
  - out of all tweets predicted
- **F-score**: harmonic mean of recall and precision
- **F-score is the optimization goal**
Learning

- Input: a time window of tweets
- Computation: gradient ascent method
  - Parameter space: $\alpha_{ji}, \beta_i, \gamma_u, \mu_i, \sigma_i^2$
  - Goal: maximize F-score
- Output: $p_i^u$
Lineup

- **LT** – Linear Threshold model
- **LTr** – Linear Threshold model with $\alpha_j$ instead of $\alpha_{ji}$
- **ALO** – At-Least-One model
- **RND** – baseline, makes random guesses about $p_i^u$
* training data: first 150 h, test data: next 150h, results for 100 random URLs
Summary

- Log-normal degree distribution
- Small-world: 3.6 hops from user to user
- Power-laws in the user activity and URL mentions
- Cascades are shallow: exponential depth falloff
- Log-normally distributed diffusion delay
- The LT model:
  - predicts more than half of the URL tweets
  - with less than 15% false positive rate
Ongoing work

- Investigating mispredictions
  - URLs
  - users
- Scaling up the real-time data mining
  - continuous MapReduce
  - crawler farm
- Website: personalized URL rankings for Twitter users
- Apply to other systems