HiggsTweet: Analyzing Influence Propagation During a Viral Event on Twitter

Kristian Flatheim Jensen
Norwegian University of Science and Technology

Gudbrand Tandberg
The University of British Colombia

In this project we analyze the influence dynamics within a certain interest-group during a viral event. The event in question is the 4. July 2012 discovery of the Higgs boson, made by researchers at CERN in Switzerland. We study a dataset consisting of a 450k-user Twitter follower network together with a 13M line event log, collected between the 1st and 7th of July 2012. We use the event log to calculate influence probabilities between pairs of users, these weights are then used to run simulations of Independent Cascade (IC) diffusion processes and compute near-optimal seed sets using a greedy algorithm. We experiment with several different preprocessing steps and heuristics for reducing the size and runtime of the simulation and optimization steps, and compare seed-sets and expected user influence across our experiments.

1 INTRODUCTION

Analyzing the spread of information through a large and complex social network is a difficult and interesting problem. In the last 10 years, the Big Data revolution has been driven, to a large extent, by innovations in understanding how humans interact and live in a mobile world. At least some of this progress has been made thanks to inter-disciplinary efforts by researchers in sociology, epidemology, psychographics, computer science. Already it has become clear the immense opportunities and dramatic shifts this revolution has brought about for marketers, news agencies, individuals and more. We will see in the near dystopian future that the study of social networks will in fact be key to developing political theory (and practice) into the 21st century.

The main goal of the broader research agenda we are following is the open-ended and general question of analyzing the power and influence dynamics of a web-community before, during, and after a major event. This study of course has a much narrower scope. Our lower level goals for this project were twofold, first, we wanted to get hands-on experience with some of the material we covered in class, notably Influence Maximization (IM), second, we wanted to perform an as-complete-as-possible scientific exposition of a new and interesting dataset. We will see how successful we were at the end!

Some questions that guided our efforts are

- *Which users in the network should we influence, and when should we influence, if we wanted to spread a rumor during a viral event?*
- *How do the power dynamics, as computed from an event log change over time during a viral event?*
- *What does the typical and the atypical user look like, in terms of event history, during a viral event?*
- *Do the seed-sets (as computed using IM) differ significantly from time to time, or is there overlap?*
- *Can the selection of seed sets be simplified, or substituted for other statistics- or feature-based heuristics?*

Course Project, December 2017, Vancouver
.

2 RELATED WORK

3 PRELIMINARIES

Present IM problem, greedy algorithm, celf

4 THE DATA SET

- very incomplete dataset

- only pairwise interactions

Social Network + Action Log

500k users; 14M edges (following/followee); 500k directed, typed actions

5 OUR APPROACH

Combining social network + "action edges" and running community based IM.

5.1 Computing Edge Probabilities

Idea: Use WC probs with different weights for different action-types.

- Hard to validate
- Hard to compare
- Effects of p_{uv} on runtime.
- Effects of p_{uv} on spread.

5.2 Influence Maximization

Greedy, CELF, or MIA? What to do..

Idea: Divide and conquer–preprocess with community detection. Wang (2012)

6 RESULTS

7 DISCUSSION

7.1 Future Work

Impact of time.

Content of tweets - sentiment analysis.

Compute p_{uv} using unsupervised learning approach. Perform evaluative analysis.

REFERENCES

[1] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[2] Francesco Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011.

[3] Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

[4] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[5] Manlio De Domenico, Antonio Lima, Paul Mougel, and Mirco Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3:2980, 2013.

[6] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1:197, 2011.

[7] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[8] Ling-ling Ma, Chuang Ma, Hai-Feng Zhang, and Bing-Hong Wang. Identifying influential spreaders in complex networks based on gravity formula. *Physica A: Statistical Mechanics and its Applications*, 451:205–212, 2016.

[9] Brendan Meeder, Brian Karrer, Amin Sayedi, R Ravi, Christian Borgs, and Jennifer Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on World wide web*, pages 517–526. ACM, 2011.

[10] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[11] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-based intelligent information and engineering systems*, pages 67–75. Springer, 2008.

[12] Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1039–1048. ACM, 2010.