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 PRELIMINARIES & RELATED WORK

Network diffusion processes have a long history of study in the social sciences. Some of the early applications include modeling the adoption of new products and technologies, modeling disease outbreaks and word-of-mouth events, and understanding human social dynamics. With the relatively recent advent of global social network platforms such as Facebook and Twitter, many new lines of research in computer science have emerged. Particularly, the processes of influence propagation in social networks has received a lot of attention.

In the age of Youtube, Facebook and Twitter, the appeal of viral marketing is to many the ultimate free lunch: Pick some small number of people to "seed" your idea, get it to "go viral", and watch while it relentlessly spreads to reach millions, all on a shoestring budget. In [23], the authors instead propose a new model called "Big Seed Marketing" that combines the power of traditional advertising and the extra punch provided by viral propagation. This follows from years on marketing research, see for example [8] for an early study in computing the "value" of a user in a social network.

The problem of selecting a set of seed-users that will trigger a large cascade of activity was first introduced in [12]. In this seminal work the computational problem of maximizing the expected spread of a seed set is formulated as a discrete optimization problem, proven to be NP-hard, and a greedy algorithm with provable approximation guarantees is presented for a class of diffusion models. Importantly, they find that they are able to attain significantly higher values of expected spread by solving the Influence Maximization (IM) problem, as opposed to both random and centrality-based seed-selection methods. Since the greedy algorithm can be relatively slow to run on large networks, several methods have been presented to deal more efficiently with the IM problem. These include the optimized algorithms CELF, MIA, TIM and IMM, [6], [5], [20], [19].

The methods for solving the IM problem all depend on a directed, weighted social network as input. The weight of a directed edge between two users is supposed to represent the degree of influence the one has over the other. Estimating this number is an interesting and general question in itself, and there is still plenty of room for more research on it. The problem of estimating influence probabilities was first presented in [11], where the authors present static and time-dependent models for learning influence weights from a log of events.

The related question of identifying influential spreaders in complex networks was partly answered in [13]. In it, they find that there are circumstances where the best spreaders do not correspond to the best connected people or to the most central people (high betweenness centrality), rather, they are located within the core of the network as identified by the k-shell decomposition analysis, and that when multiple spreaders are considered simultaneously, the distance between them becomes the crucial parameter that determines the extent of the spreading. More answers to the same question came in [1], where the authors investigate the attributes and relative influence of a large Twitter follower graph over a two month interval in 2009. They find that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. They also find that hashtags that were rated more interesting and/or elicited more positive feelings were more likely to spread. They also find that predictions of which particular user will generate large cascades are relatively unreliable. For an excellent survey of computational models of influence propagation, see e.g. [3], or [9].

In [22], the authors present a community-based algorithm for mining top-k influential nodes in social networks. Their method is found to lead to a decrease in runtime, while not sacrificing much in terms of spread. This is perhaps not so surprising, given the above result that distance between seed-users becomes the crucial parameter that determines the extent of the spreading.

Combining the problems of influence maximization, influence estimation, and viral event dynamics, we are faced with the problem of real-time IM on dynamic social networks. This has been addressed in [16], and more recently in [21].

Some other relevant and recent research that we have taken inspiration from include studies of differences in mechanics of diffusion across topics [17], the dynamics of protest recruitment [10], sentiment reciprocality in reply networks [2] and prediction of social-link creation times [15].

The Higgs dataset was first presented in [7], where the authors present the dataset, explore the spatio-temporal properties of the data, and demonstrate a model for the information spreading in the social network during the event.

3 THE DATA SET

The dataset we are working with consists of a social network and an action log. The social network is a contains just over 450k unique Twitter users with just over 14.8M directed edges representing one user following another. The action log is a list of 563069 events of the form

```
user1, user2, event_type, time
```

where `user1` is the user who performs the action, `user2` is the user to whom the action is directed, `event_type` is one of either RT (retweet), MT (mention) or RE (reply), and `time` is the UNIX timestamp for the event. The dataset was scraped from the web using the Twitter API by the authors of [7]. The contents of the tweets and the identity of the tweeters is, unfortunately, not available. All we do know is that all of the tweets corresponding to events in the log contained one or more of the hashtags "LHC", "Higgs", "CERN" or "boson". It would have been very interesting to have access to not only the contents of the tweets, e.g. in order to perform some sort of semantic analysis, but also to have access to standalone tweets in the period, not just user to user interactions.

Of all the events in the log, roughly 62% of them are between two users where at least one of the users follows the other. This means that roughly 38% of the events are directed between "complete strangers". Of all the events, 354930 are retweets, 171237 are mentions, and 36902 are replies.

The authors of [7] identify four different periods into which the events can be divided:

Period I Before the 2nd July, there were some rumors about the discovery of a Higgs-like boson at the Tevatron accelerator

Period II On the 2nd July at 1PM GMT, scientists at the Tevatron accelerator in Fermilabs, Illinois, announced that they had discovered the Higgs boson with a 1 in 550 likelihood.

Period III After 2nd July and before 4th July there were many rumours about the Higgs boson discovery at the LHC.

Period IV The main event was the announcement on 4th July at 8AM GMT by the scientists at CERN. After 4th July, popular media covered the event.

The distribution of events in the four periods are described in Figure 1 and Table 1.

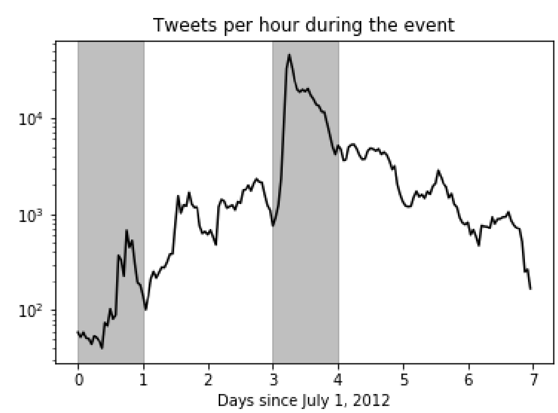

Figure 1: Activity levels through time during the event.

For completeness, the distribution of in- and out-degrees of each user (number of followers and following, respectively) can be seen in Figure 2, and the average values of the same quantities in **??**

Table 1: Distribution of activities in the different periods.

Period	#activities
I	4181
II	49604
III	356539
IV	152747

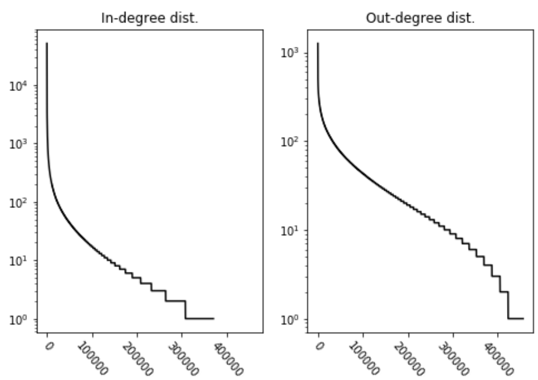

Figure 2: Distribution of in- and out- degree of users in the social network

Table 2: Average number of followers/following in the social network

	#followers	#following
mean	32.5	32.5
median	4.0	16.0

4 OUR APPROACH

In this paper, our goal is to experiment with IM on the Higgs dataset presented above in order to answer some of the questions stated in the introduction. In order to do this, we first combine the social network and the action log into a hybrid multi-digraph as follows: every node in the social network is first added to the hybrid network. Then, every link in the social network is added as a link in the hybrid network with the label "FR". Finally, all events in the action log are added as links in the hybrid network labelled as either "RT", "MT", or "RE". Later, we will also experiment with creating these graphs on a per-period level, so as to compare seeds over time.

The next step is to convert the hybrid multi-digraph into a weighted digraph, where, hopefully, the weights in the graph correspond roughly to the level of influence one user exerts over the other. This digraph can then be used to run IM experiments. Figure 3 shows the process schematically in a simple setting.

As mentioned earlier, there are many viable ways of estimating the weights in the digraph. We propose a simple additive-weights method, where all parallel edges in the network are combined using predetermined weights per type, with $\mathcal{W} = \{w_{\mathrm{FR}}, w_{\mathrm{RT}}, w_{\mathrm{MT}}, w_{\mathrm{RE}}\}$.

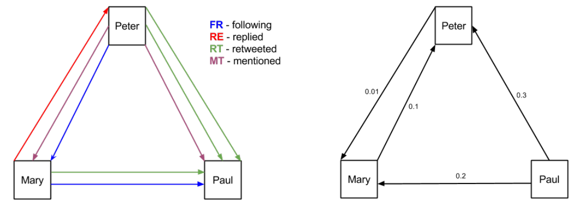

Figure 3: ***Left:*** **original hybrid unweighted multi-digraph** ***Right:*** **influence-weighted digraph**

$$p_{uv} = \min\left\{1, \sum_{e=(v,u,t)} w_t\right\}, \tag{1}$$

where the sum ranges over all edges e from v to u with type t. We find this method to be simple, cheap and effective, but it is hard to validate whether our formula actually captures the relative influence levels present in the user-set. We set the weights in $\mathcal{W}$ using a combination of trial and error and relative frequency of action-types in the event-log. An advantage of our weights-method is that we can for example choose to set $w_{\mathrm{FR}} = 0$, which means the resulting digraph has roughly 15M fewer edges, drastically decreasing the runtime of our algorithms, as we will see later.

It is an interesting question whether the weights can be set in a more principled manner, and we leave this for later work. However the weights p_{uv} were computed, it is evident that they are paramount to the later success of our IM campaign. Given a set of weights, it is hard to validate their accuracy without expensive and intrusive user-surveys, or access to extensive event-logs. Furthermore, when faced with two different sets of weights, it is hard to compare the effects of the two, given that they will greatly impact both the diffusion process and the runtime of the algorithm. For these reasons we stick to our simple formula for the remainder of this paper. The distribution of influence weights in the network can be seen in Figure 4.

4.1 Influence Maximization

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

5 RESULTS

6 DISCUSSION

6.1 Future Work

Impact of time.

Content of tweets - sentiment analysis.

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

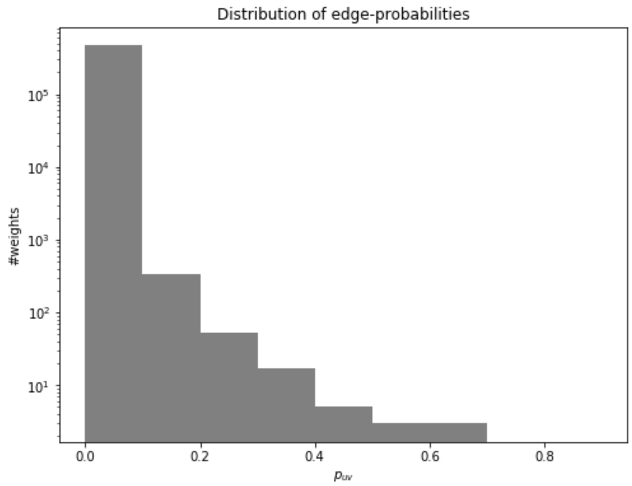

Figure 4: Distribution of edge-weights in the social/action hybrid network

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] Catherine A Bliss, Isabel M Kloumann, Kameron Decker Harris, Christopher M Danforth, and Peter Sheridan Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, 3(5):388–397, 2012.

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

[4] 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.

[5] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.

[6] 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.

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

[8] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[9] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.

[10] 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.

[11] 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.

[12] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[13] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[14] 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.

[15] 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.

[16] Manuel Gomez Rodriguez and Bernhard Schölkopf. Influence maximization in continuous time diffusion networks. *arXiv preprint arXiv:1205.1682*, 2012.

[17] 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.

[18] 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.

[19] Youze Tang, Yanchen Shi, and Xiaokui Xiao. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1539–1554. ACM, 2015.

[20] Youze Tang, Xiaokui Xiao, and Yanchen Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 75–86. ACM, 2014.

[21] Yanhao Wang, Qi Fan, Yuchen Li, and Kian-Lee Tan. Real-time influence maximization on dynamic social streams. *Proceedings of the VLDB Endowment*, 10(7):805–816, 2017.

[22] 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.

[23] Duncan J Watts, Jonah Peretti, and Michael Frumin. *Viral marketing for the real world*. Harvard Business School Pub., 2007.