Infectious Knowledge in a Collaborative News Site

Eliana Feasley
Department of Computer Science
University of Texas at Austin
elie@cs.utexas.edu

Wesley Tansey
Department of Computer Science
University of Texas at Austin
tansey@cs.utexas.edu

Abstract

In recent years, the attention paid to cascading information in soical networks has been increasing in a fashion itself comparable to a social cascade. This makes sense - the way that information infects different spaces of ideas has applications to basic graph theory **Cite** , epidemiology **Cite** , future prediction **Cite** , and marketing **Cite** . In this paper, we examine the dynamics of information spread across subcommunities with overlapping networks in both the social news collaborative filering site reddit[1] and in several data mining conferences. Both of these are structured such that it is possible to track how ideas spread over time, and to discover semi-explicit communities and the connections between them.

1 Introduction

Information online travels across networks in a variety of configurations, and it is easy to see information spreading across them. Past work such as **Cite** **Cite** and **Cite** all illustrate examples of phrases or terms spreading rapidly across networks. This is an extension of the work in [2], in which the authors learn the structure of a graph from observing cascades. We assume the graph is fully connected, and predict the state of a cascade given the inferred weights on the graph.

In this paper, we explore how modeling this machanism as a timeseries can help us both to predict future topics and to discover the strengths of the connections between communities. Section 3 explains the structure of the data and the domains from which they are drawn. Section 2 delves into past work on cascades in networks. Sections 4 and 5 explain the experiments and results respectively, and Section 6 explores the broader implications of our results and some future work.

2 Social Cascades

In [4] the authors use the explicit struture of networks - observing following and friendship relationships in order to explore the effect of actual, instead of inferred, network structure on information cascades. They observe that the popularity of stories peaks with an age of about one day, and then subsides.

A unique aspect of open networks like reddit, digg, publications, &c is that it is possible for information to latently travel quicly, as opposed toin closed, action oriented networks like the one described in [5], where information, which must individually be spread from one email to another, peters out quickly.

In [6], the authors identify *bursty* keywords that suddenly appear, and attempt to align them with trends - entire topics that are becoming more popular. They do this by analyzing new bursts in the queue. One thing we can do in this paper is look at each sub as a queue and see if bursts in one are followed by bursts in another.

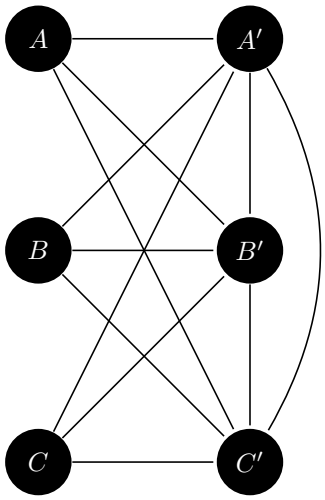
[1] www.reddit.com

2.1 Problem Statement

Given nodes $\mathcal{D}$ with overlapping connections $\mathcal{C}$ via shared users, predict topic vector $v_{d,t+1}$ in document collection d at time $t+1$ given the topics V in all documents at time t?

Following the model in [3], we refer to a node as *contagious* for a given phrase if it has had that phrase trend internally within the last timestep. A node that contains a previously trending phrase can be viewed as having become *infected.*

2.2 Problem Structure



2.3 Detecting Infections

To detect infections, we use Pointwise Mutual Information [1] to identify salient bigrams in each site in each timestep. Whenever these occur multiple times across the entire dataset, they may be infections. We examine the occurences of each ngram to see if it appears in bursts, and if it does, we designate it an infection.

2.4 Learning Parameters

We used the method for finding the MLE estimate in triangulated graphs described in [7] to learn the weights of our edges.

3 Datasets

Our algorithm is evaluated on three datasets: a reddit dataset scraped from reddit.com, a stackexchange dataset, and a synthetic dataset.

3.1 Reddit

Reddit is divided into thousands of *subreddits*, each of which is targeted towards specialty interests. There is a many-to-many relationship between users and subreddits, with most users active in many subreddits and most subreddits populated with many users.

We looked at the top 20 subreddits in terms of popularity, and as these are so active, we set our timestep to be six hours. For each timestep, we formed a document of all of the post titles present during that time. A contagion was defined as described in 2.3.

3.2 StackExchange

3.3 Synthetic Dataset

4 Experiments

We conducted three sets of experiements, on the `reddit` dataset, the `StackExchange` dataset, and the synthetic dataset. Each of these was similar, in that we used n-fold cross-validation to predict cascades with our algorithm, and with the edge weights learned by NetInf.

5 Results

6 Discussion

References

[1] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[2] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):21, 2012.

[3] D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. *Automata, Languages and Programming*, pages 99–99, 2005.

[4] Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. *CoRR*, abs/1003.2664, 2010.

[5] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[6] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In Ahmed K. Elmagarmid and Divyakant Agrawal, editors, *SIGMOD Conference*, pages 1155–1158. ACM, 2010.

[7] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.