Infectious Cascades in Knowledge-Based Communities

Eliana Feasley and Wesley Tansey

December 3, 2012

Social Cascades

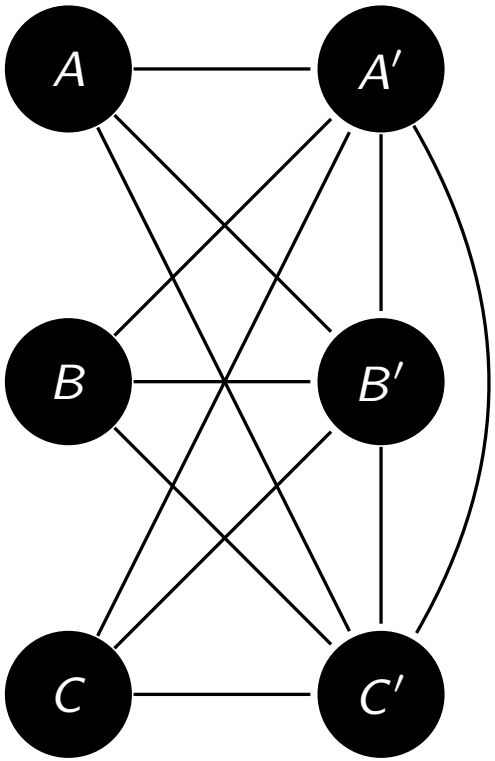
Social cascades capture the concept of new ideas, or *memes*, spreading across influence networks and infecting subcultures. Cascade theory has applications to several areas, including epidemiology, graph theory, machine learning, and marketing. In this project, we examine the dynamics of information spread across sub-communities with overlapping networks in several real-world domains: the social news site *reddit*, a collection of blogs and mainstream media [3], and the question-answer network *StackExchange*. Our goal is to capture the flow of memes by learning a graphical model for each domain. We learn a graphical model describing the transition between timesteps, effectively capturing how the memes cascade through the network.

Problem Statement

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

Following the model in [2], 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.*

We model this domain as an unfolding Markov Chain, in which at each timestep, nodes become infected or uninfected. See the below graph, in which nodes $\{A, B, C\}$ represent the states of three sites at time t, and the nodes $\{A', B', C'\}$ represent the same nodes at time $t+1$.



Datasets

Datasets

Our algorithm is evaluated on three datasets.

MemeTracker A collection of popular newssites during the campaign season of 2008. Sites are blogs and major media sources. Memes detected via the Memetracker algorithm [3] are designated infections.

StackExchange There are 28 different sites on StackExchange. Pointwise Mutual Information [1] identifies salient bigrams. Common bigrams detected by PMI that occur in bursts are infections.

Synthetic Infections are randomly initialized and spread through sampling via MCMC from generated parameters.

Parameter Learning

Learning Parameters

We used the method for finding the MLE estimate in triangulated graphs described in [4] to learn the weights of our edges. The model is trained on every pairs of steps $t, t+1$. We define indicator functions

$$\mathbb{I}[s] = \begin{cases} 1 & \text{Infection occurs in site } s \\ 0 & s \text{ is not uninfected} \end{cases}$$

and

$$\mathbb{I}[s,t;j,k] = \begin{cases} 1 & \text{if infection is present in site } s\ (j=1) \text{ and } t\ (k=1) \\ 0 & \text{otherwise} \end{cases}$$

Parameters are the logs of the expectations of these indicator functions.

Inference Problem

Our question Given a cascade at time t, what will be the state of the cascade at time $t+1$ can be restated as the following variational problem:
Given that we have learned our parameters θ_s and $\theta_s t$, and that we have observed half of the nodes $X = \{x_1, x_2, \ldots, x_n\}$ representing all of our sites at time t, what is the MLE assignment to the nodes $X^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$?
This optimization over 2^n possible assignments with n equal to the number of sites under examination is NP-complete.
Good approximations are found by searching the parameter space.

Discussion


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


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


J. Leskovec, L. Backstrom, and J. Kleinberg.
Meme-tracking and the dynamics of the news cycle.
In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.


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.