Modeling Infectious Knowledge in Collaborative Communities

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

The way that information infects different spaces of ideas is a crucial topic with applications to basic graph theory, epidemiology, machine learning, and marketing. In this paper, we examine the dynamics of information spread across communities with overlapping networks in several real-world domains. Each of these collaborative communities is structured such that it is possible to track how ideas spread over time and to discover latent subcommunities. We present an approximate graph structure inference algorithm that captures the most influential connections between subsites in a collaborative community. Additionally, we provide an approach for leveraging graphical models to predict the spread of memes through a community over time. The techniques we develop provide a framework for a new class of social network analysis.

1 Introduction

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 paper, we examine the dynamics of information-spread across sub-communities with overlapping networks in several real-world domains: the social news site *reddit*; the Meme Tracker dataset [9], a collection of blogs and mainstream media; the question-answer network *StackExchange*; and the general interest forum *Something Awful*. Our goal is to capture the flow of memes by learning a graphical model for each domain. In this paper, we present approaches for inferring both the structure and weights of such models. Additionally, we model meme infections as timeseries to predict future topics and discover the structure of latent communities.

The remainder of this paper is organized as follows. Section 2 presents an overview of previous work on cascades. Section 3 presents our algorithm for approximate structure learning in collaborative communities. Section 4 details the results of using graphical models to predict social cascades. Finally, section 5 discusses potential future work.

2 Background

The rise of online social media in recent years has created numerous opportunities for social cascade researchers. In this section, we discuss some of the techniques developed for studying cascades in social media, compare their applicability to collaborative communities, and differentiate them from our own approach.

2.1 Collaborative Communities

A *collaborative community*, in this paper, is defined as any site or network of sites which has a single overall userbase and multiple subsites. Each user in the community is subscribed to at least one subsite and can optionally choose to subscribe to as many of the subsites as they like. Examples of such communities may include social news sites, question-answer sites, and discussion forums. Note that traditional friendship-based social networks are not valid collaborative communities, as users must be given permission to *friend* other users. Similarly, microblogging sites do not qualify, as users are not able to comment directly on another user's feed.

2.2 Social Cascades

Social cascades have been investigated across several kinds of social networks. In [8], following and friendship relationships are observed in order to explore the effect of explicit network structure on information cascades. In such networks, popularity of stories peaks with an age of about one day, and then subsides.

Open networks, where all information is accessible to all users at all times, are fundamentally different in their behavior than closed networks, where information is private and must be shared through an explicit user action. Closed email chain networks were analyzed in [10] and showed a short lifetime for viral information. In contrast, open networks have shown [1] rapid spreading to a large and diffuse user base.

Structureless approaches have also been explored. In Twitter[1], *"bursty"* keywords can be aligned with trends – entire topics that are becoming more popular [11]. In collaborative communities, however, we are interested in both the rise of such topics and the path they travel through the network of subsites.

2.3 Structure Learning

Networks of influence have been inferred in a variety of contexts. Contagion networks have been used [4] to model the spread of disease throughout a social network. In marketing, contagion networks have been leveraged to identify how products and ad campaigns virally reach critical mass (i.e., "go viral") [2, 6]. In a more general context, NetInf [3] was developed to infer directed network structure from a list of meme "infections" in a given network. Similarly, [12] infers graphs to visualize and better analyze flows in social networks. While all of these approaches are powerful techniques, none quite fit the requirements of collaborative community network inference. Specifically, none are able to filter a fully-connected, undirected graph down to only the most influential connections in a manner amenable to probabilistic graphical inference as an exponential family.

3 Learning Structure

In this section, we present our approach to learning the structure of influence graphs in collaborative communities. We first present an overview of the general problem, followed by our approach to learning structure. Finally, we validate our method on three real-world datasets: *reddit*, *StackExchange*, and *Something Awful*.

3.1 Challenges for Collaborative Communities

Learning graphical model structure is a particularly difficult task for collaborative communities. Each subsite forms a node in the network, and memes can potentially spread between any two subsites that share at least one user. In friendship-based social networks, we typically expect the percentage of mutual friends to be relatively low and edges exist simply as binary friend-or-not connections, resulting in a very sparse adjacency matrix. However, in collaborative communities with thousands of participants, it is reasonable to expect that most subsites share users, resulting in a nearly-complete adjacency matrix. A naive approach would thus result in a fully connected graph, which is undesirable as it yields little insight and may make certain tasks intractable.

[1] http://twitter.com

3.2 Approximate Structure Learning Algorithm

As it is possible for a meme to spread between any two subsites, any structure learning algorithm that yields a less-than-fully-connected graph is learning an approximate structure. As noted previously, an approximate structure may be desirable for gaining insights and reducing computational requirements. We next present our approximate structure learning algorithm for collaborative communities.

We first begin by defining an *adjency matrix* $\mathcal{A}$ over subsites $\mathcal{S}$, such that $\mathcal{A} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$. For each pair $(s,t) \in \mathcal{A}$ of subsites, we mine the percentage of *active* users that overlap in both s and t. We define a user as *active* if they have posted at least once on both subsites[2]:

$$a(u) = \begin{cases} 1 & u \in P_s \text{ and } u \in P_t \\ 0 & \text{otherwise} \end{cases}$$

where P_i is the set of posts in subsite i. This yields the adjacency matrix:

$$\mathcal{A}[s,t] = \sum_{\forall u \in \mathcal{U}} \frac{a(u)}{min(|\mathcal{U}^s|, |\mathcal{U}^t|)}$$

where $\mathcal{U}$ is the set of active users and $\mathcal{U}^i$ is the set of active users in subsite i.

Rather than focusing on binary co-occurrence edges, our algorithm assesses the strengths of user overlap between two subsites and removes edges that are below a user-specified threshold. We define the relative strength matrix, $\mathcal{A}^*$:

$$\mathcal{A}^*[s,t] \quad = \quad \alpha \frac{\mathcal{A}[s,t] - \mu(\mathcal{A})}{\sigma(\mathcal{A})}$$

where μ and σ are the mean and standard deviation, respectively, and α is a tuning parameter. Finally, we define the weighted edge matrix, $\mathcal{E}$:

$$\mathcal{E}[s,t] \quad = \quad \begin{cases} \mathcal{A}^*[s,t] & \text{if } \mathcal{A}^*[s,t] \geq \gamma \\ 0 & \text{otherwise} \end{cases}$$

where γ is the user-specified strength threshold.

The resulting graph, $\mathcal{G} = (\mathcal{S}, \mathcal{E})$, captures the most likely sources of influence for each node. To predict meme spreads, one could then convert $\mathcal{G}$ to a dynamic graphical model, $\mathcal{G}^*$. To do this, we introduce a set of n timesteps, $T = \{t_0, t_1, \ldots, t_n\}$, such that the state of the nodes at a given time t_i is represented by $\mathcal{S}_i$. $\mathcal{G}^*$ captures the transition from $t_i \rightarrow t_{i+1}$ as follows:

$$\begin{array}{lcl} \mathcal{S}_i^* & = & S_i \cup S_{i+1}, \\ \mathcal{E}_i^*[s,t] & = & \begin{cases} 1 & \text{if } s \in \mathcal{S}_i,\, t \in \mathcal{S}_{i+1}, \text{ and } \mathcal{E}[s,t] > 0 \\ 1 & \text{if } s \in \mathcal{S}_{i+1},\, t \in \mathcal{S}_{i+1}, \text{ and } \mathcal{E}[s,t] > 0 \\ 0 & \text{otherwise} \end{cases}, \\ \mathcal{G}_i^* & = & (\mathcal{S}_i^*, \mathcal{E}_i^*). \end{array}$$

Queries to our dynamic graph will always be performed with the nodes in $\mathcal{G}_i$ being observed, thus connecting them would not affect the MAP inference for nodes in $\mathcal{G}_{i+1}$. One important point is that while the weights inferred by our algorithm could represent final weights in a graphical model, it is not clear that doing so would produce an accurate probability distribution.

[2]Note that mining inactive users would be infeasible since most collaborative communities do not reveal the subsites to which a user subscribes.

3.3 Experiments

To validate our approach, we mined active user adjacency matrices for three real-world collaborative communities: *reddit*, *StackExchange*, and *SomethingAwful*. The results for each community are presented next.

3.3.1 reddit

reddit[3] is a social news community where each subsite[4] represents a different news topic. By default, all users are subscribed to a generic set of subsites such as *politics* and *humor*. The high degree of overlap among these default subsites produces a low signal-to-noise ratio in the adjacency matrix. Consequently, we first removed all default subsites from our dataset. To conform to our time constraints and maximize the amount of insight we could draw, we then limited the remaining non-default subsites to the top 25. The graph generated by our structure learning algorithm is shown in Figure 1.

Two interesting properties of the graph immediately stand out. First, the explicit-content subreddits, *nsfw*, *sex*, and *gonewild*, are clustered together, indicating that memes on these subreddits are not likely to spread to the rest of the network and memes within the cluster will quickly infect the remaining nodes. Conversely, the marijuana-oriented subreddit, *trees*, is a super-node that appears to exert influence on nearly every node and vice-versa. Intuitively, we may thus hypothesize that the users of *trees* are very influential on reddit and would thus expect to see memes spread from and to *trees* rapidly.

3.3.2 StackExchange

StackExchange[5] (SE) is a collaborative question-answer community where each subsite represents a different question topic. A majority of subsites focus on technical topics such as *android* and *security*, though non-technical subsites exist (e.g., *cooking*). Memes in SE are likely to take the form of rising topics of interest, such as the popularity of iPhone games or Ruby on Rails. Unlike most collaborative communities, SE user subscription data is publicly available. We therefore used the adjacency matrix of all users to build our graph, shown in Figure 2.

Perhaps counterintuitively, the *stackoverflow* subsite has two orders of magnitude more overall users than other sites, but exerts no influence on the network. This may be a result of the other subsites being relatively unknown but having a strong sub-culture of shared users. Two high-influence pseudo-clusters appear between less technical and strictly technical sites in the upper-left and lower-right corners of the graph, respectively. Thus, localized memes may be frequent among these pseudo-clusters.

3.3.3 Something Awful

Something Awful[6] (SA) is a general interest internet forum consisting of 53 subsites, where each subsite is a forum devoted to a separate topic such as games or books. SA is often regarded as a progenitor of memes, such as the infamous "All your base are belong to us" meme [5]. The graph inferred from SA is shown in Figure 3.

The SA graph reveals the important role of the main comedy subsite, *Comedy Goldmine*. In contrast, the niche-specific subsites for music, football, and automobiles form clusters among their topics and exert virtually no influence on the broader site. Thus, intuitively, we would expect memes on Something Awful to originate from the comedy subsites and quickly spread to other general interest sites, but have little affect on niche subsites.

[3] http://reddit.com

[4] reddit subsites are called subreddits.

[5] http://stackexchange.com

[6] http://forums.somethingawful.com

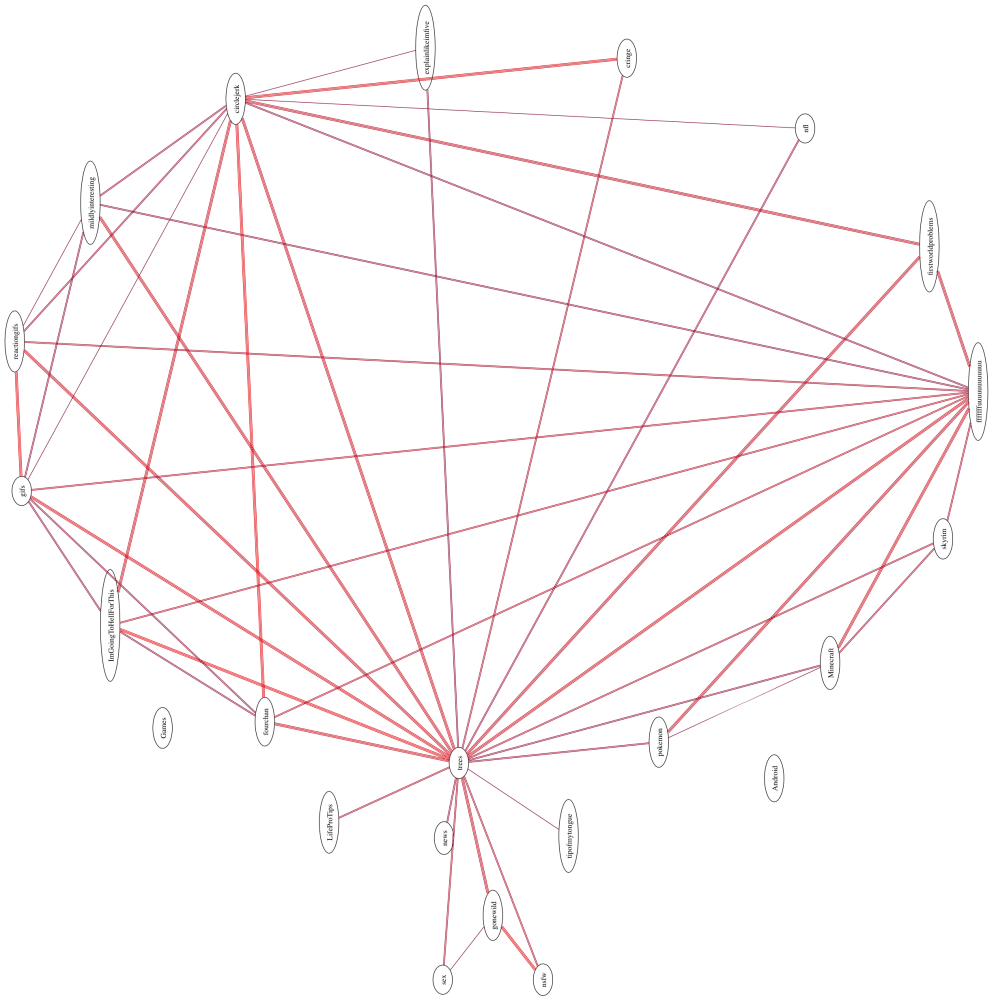

Figure 1: Graph inferred from the top 25 non-default subreddits.

4 Predicting Cascades

Given a graph structure, we are faced with a difficult inference problem: detecting cascades in networks of subsites. We leverage the exponential family techniques from [13] to learn a parameterized model. In contrast to NetInf [3], where the strength of connections between sites is used merely to visualize and explore the data, we learn a probability distribution over nodes that can predict future data. We validate our approach on the Meme Tracker [9] dataset. Due to the time constraints of the

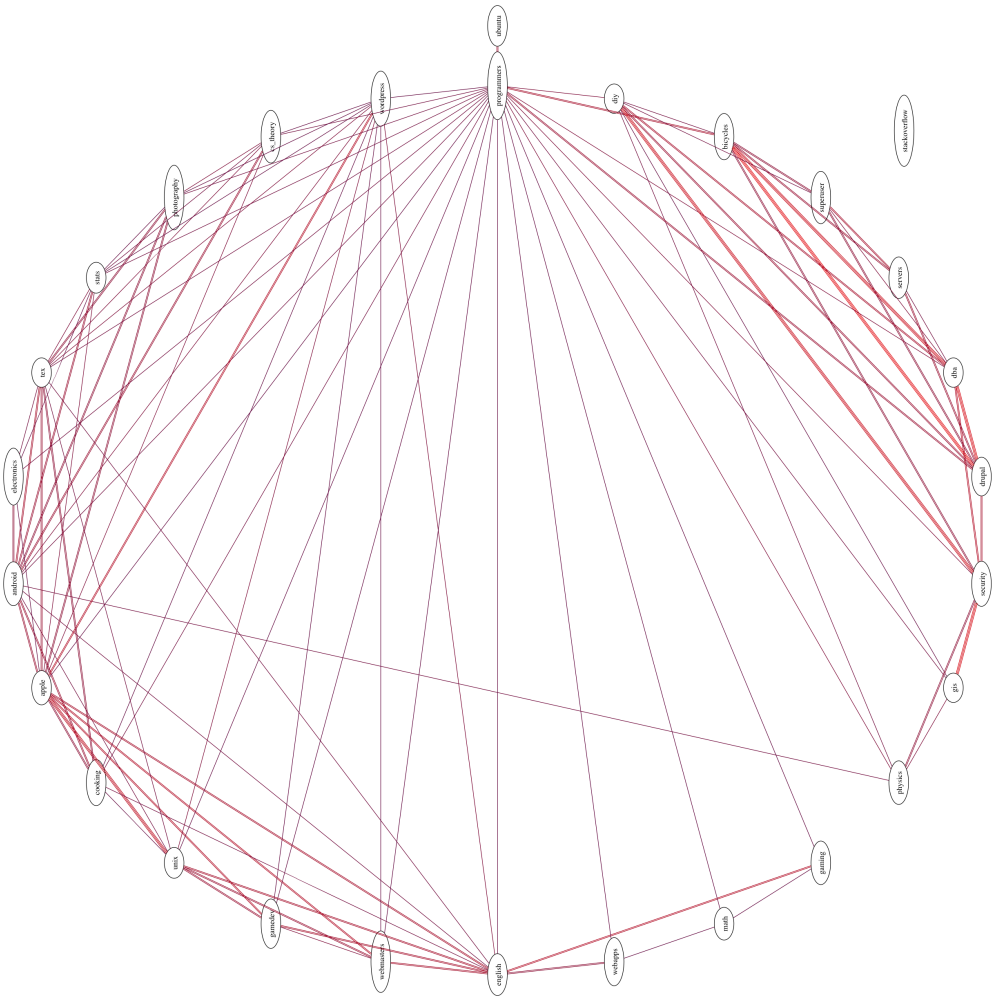

Figure 2: Graph inferred from the 28 StackExchange subsites.

project, we were unable to gather sufficient data from reddit, SE, or SA; these sites are left as future work.

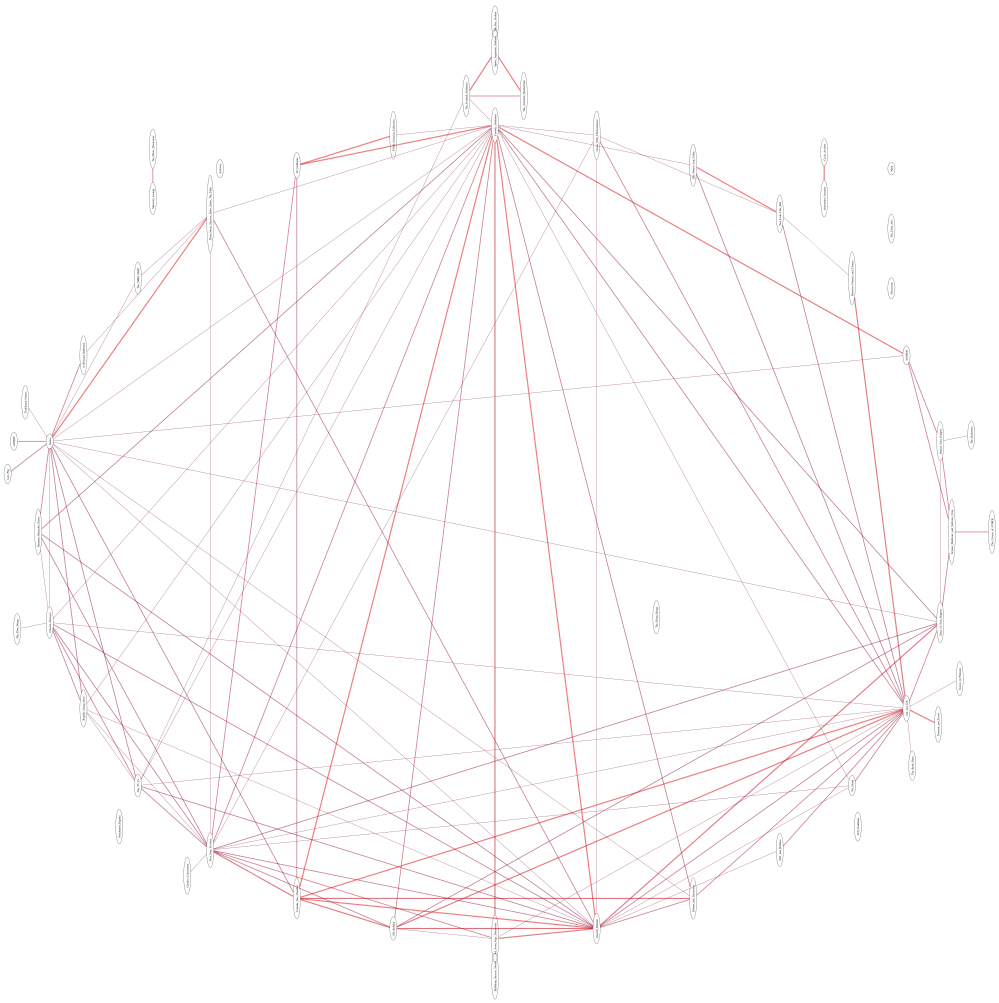

Figure 3: Graph inferred from the 53 Something Awful subsites.

4.1 Problem statement

Given a graphical model $\mathcal{G}_t^* = (\mathcal{S}_t^*, \mathcal{E}_t^*)$ created as specified in Section 3.2, with nodes $s_{i,t} \in \mathcal{S}_t$ taking values

$$s_{i,t} = \begin{cases} 1 & \text{if node } i \text{ is infected at time } t \\ 0 & \text{otherwise} \end{cases}$$

we want to predict the state of $\mathcal{S}_{t+1}$. Following the model in [7], we refer to a node s as becoming *infected* with a given phrase if that phrase trended in s during the last timestep. Once a node is infected, it is considered *contagious.*

We model this domain as an unfolding Markov chain, where at each timestep nodes may become infected. Once infected with a meme, a node stays infected for the remainder of the time series. Figure 4.1 shows an example of this structure in which nodes $\{A, B, C\}$ represent the states of three subsites at time t, and nodes $\{A', B', C'\}$ represent the same subsites at time $t + 1$.

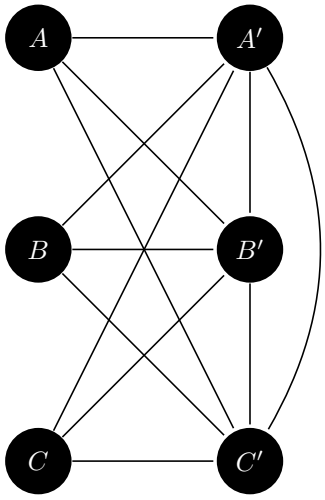

Figure 4: An example collaborative community network of three subsites.

4.2 Learning Parameters

We used the method for finding the MLE estimate in triangulated graphs described in [13] to learn the weights of our edges. The model is trained on every example graph $\mathcal{G}_t^*$ in the dataset. We define indicator functions

$$\mathbb{I}_{s;j} = \begin{cases} 1 & \text{Node } s \text{ is in state } j \\ 0 & s \text{ is not uninfected} \end{cases}, \forall s \in \mathcal{S}_t^*$$

and

$$\mathbb{I}_{r,s;j,k} = \begin{cases} 1 & \text{Node } r \text{ is in state } j \text{ and node } s \text{ is in state } k \\ 0 & \text{otherwise} \end{cases}, \forall r, s \in \mathcal{E}_t^*.$$

Parameters $\hat{\theta}$ are the logs of the expectations of these indicator functions as defined in [13].

We define the probability of an assignment to $\mathcal{S}_{t+1}$ given our graph $\mathcal{G}_t^*$ and $\mathcal{S}_t$ to be

$$p_{\hat{\theta}}(\mathcal{S}_t^*) = \exp\left\{ \sum_{s \in \mathcal{S}} \hat{\theta}_s(\mathcal{S}_t^*) + \sum_{s,t \in \mathcal{E}} \hat{\theta}_{st}(\mathcal{E}_t^*) \right\}.$$

4.3 Experiment and Results

Our algorithm is evaluated on the Memetracker dataset used in [9]. Due to time constraints, we leave the evaluation of the three collaborative communities from Section 3 as future work.

The Meme Tracker dataset is a collection of popular news sites scraped during the US presidential campaign season of 2008. Sites in the dataset are divided into two categories: blogs and major media sources. Memes detected via the Memetracker algorithm [9] are designated infections. This

Fold	Total	Complete Matches	% Matches	Avg. % Nodes Correct
0	8314	3130	0.3765	0.7512
1	9216	3194	0.3466	0.7141
2	7537	2937	0.3897	0.7910
3	8036	3503	0.4360	0.7931

Table 1: The results of our four-fold cross-validation on the Meme Tracker dataset.

experiment uses the 20 most popular webssites and the 40 most popular blogs to evaluate the performance of our learning algorithm. We validated our approach by measuring the accuracy of our model using 4-fold cross-validation, with a 1% holdout size per fold.

Accuracy in these experiments is denoted by the percentage of nodes correctly predicted by our algorithm for timestep $t+1$ given the nodes in timestep t. Our model relies on hill climbing to find a potential MAP assignment over nodes in $t+1$. The results of our experiment are shown in Table 1.

5 Discussion and Future Work

One of the major difficulties of researching infections in social networks is the degree of sparsity. If a site is only infected when a meme is present, the graph is very sparse, and phenomena like a user absorbing content at time $t-1$ and then reproducing it on another site at time $t+1$ are difficut or impossible to model. However, if a site becomes infected with the first occurance of a meme and stays infected forever, information about when a meme is first flaring or when it hasn't been active in the last several timesteps is lost.

Important future work includes reformulating this problem with real values. In every formulation of the problem thus far, an infection is either present or not present. Real-valued indicators of the presence of an infection will help with the problem of information loss and sparsity.

Finally, time constraints on the project prohibited gathering sufficient data on the reddit, SE, and SA communities from Section 3 to identify memes. We plan to continue scraping this data over the coming months and generate datasets comparable in size and scale to that of the Meme Tracker dataset.

References

[1] M. Cha, A. Mislove, and K.P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[2] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, 9(3):1–18, 2001.

[3] 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.

[4] A.L. Hill, D.G. Rand, M.A. Nowak, and N.A. Christakis. Infectious disease modeling of social contagion in networks. *PLoS computational biology*, 6(11):e1000968, 2010.

[5] Rich Johnston. All your base... *The Guardian*, February 2001.

[6] D. Kempe, J. Kleinberg, and É. 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.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[13] 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.