An Agent-Based Model of Urgent Diffusion in Social Media

Jeffrey Herrmann[1], William Rand[2], Brandon Schein[3], and Neza Vodopivec[4]

[1] Dept. of Mechanical Engineering, University of Maryland, jwh2@umd.edu
[2] Center for Complexity in Business, University of Maryland, wrand@umd.edu
[3] Dept. of Mechanical Engineering, University of Maryland, scheinb@gmail.com
[4] App. Math & Scientific Comp., University of Maryland, nvodopiv@umd.edu

Abstract. With the growth of the use of social media, information is exchanged faster than ever. Understanding this diffusion of information is critical to planning for time sensitive events and situations. However, most traditional models of diffusion do not work well in situations where outside events are occurring at almost the same rate as diffusion dynamics within the network. In this paper, we present two models that provide insight into urgent diffusion dynamics using agent-based modeling. We fit these models to data drawn from four major urgent diffusion events including: (1) the capture of Osama Bin Laden, (2) Hurricane Irene, (3) Hurricane Sandy, and (4) Election Night 2012. We show that in some cases social networks play a large role in the diffusion of information and in other cases they do not, and discuss the robustness of our models to a wider variety of urgent diffusion situations.

Keywords: urgent diffusion, diffusion of information, news, hurricanes, social networks, twitter

1 Motivation

In many contexts, time plays a critical role in the diffusion of information. These events can be anything including: (1) man-made crises, such as biological and terrorist attacks, (2) natural crises, such as hurricanes, (3) critical news events, such as political elections or sporting events, and (4) corporate crises, such as brand reputation issues. We refer to these situations as *urgent diffusion* events, which we define to be events where outside dynamics and news are entering the system at the same rate or even faster than information is diffusing across the system.

Understanding how information diffuses in these situations is critical for a number of different reasons. If the goal is to construct an optimal policy response to a crisis, then we need to understand how information will diffuse in these scenarios. For policymakers understanding urgent diffusion will help them get the right word out to the right people to optimally respond to man-made or natural crises. For news agencies, understanding urgent diffusion will allow them to understand what news pieces are grabbing the most attention quickly. For

brand managers, understanding urgent diffusion will facilitate effective responses to brand crises.

In this paper, we will first discuss the relevant background literature and then move on to discussing the models and the data we used to evaluate the models. We will then compare model results with actual data, and discuss what this means providing suggestions for future work in this area. Moreover, we have attached an ODD protocol writeup of the models that we use.

2 Relevant Literature

There is an entire thread of research focused on understanding the diffusion of information in large-scale networks, but most of it has focused on non-urgent diffusion events. For instance, the original Bass model popular in the field of marketing was originally constructed for understanding the diffusion of consumer durables [1]. Though it has been applied recently to understanding the diffusion of more ephemeral objects, such as social network apps, that was not what it was originally intended for, and whether it works to model urgent diffusion is an open question [2]. Another model that is frequently used to understand the diffusion of information is the cascade model [3]. Recently the cascade model has been applied to understanding Twitter networks, but the application was still in a non-urgent situation [4] The cascade model was built to model the diffusion of information, but it was created primarily to handle non-urgent situations.

There has also been some work examining the qualitative nature of urgent diffusion and how individuals are using modern communication methods, such as social media to address urgent situations. These events include natural disasters and man-made disasters, but unfortunately not much of this work has discussed quantitative models of urgent diffusion. For instance, recent work has focused on how to process vast amounts of social media data that are diffusing urgently, but this work does not discuss how to model the diffusion of this information, and is built as a reactive tool rather than a planning tool [5] [6]. Other work has focused on how to better use social media for disaster response, but again this work does not entail a quantitative or computational model of information diffusion in these urgent situations [7] [8].

Our paper differs from this previous work in that our goal is to build a model that can be calibrated to different disaster scenarios and primarily used to understand how to respond to disasters assuming that information is diffused in the manner described by our model. As a result, our work differs from previous agent-based models (ABMs) of information diffusion because we are concerned with understanding urgent events, and it differs from previous work on the use of social media in urgent situations because we are interested not in best practices for the use of social media, but rather in modeling the status quo use of this medium in urgent situations.

3 The Models

There has been considerable previous work understanding and modeling the diffusion of information [9] [10]. However, these models were originally built for non-urgent scenarios and as such, may not describe the dynamics of situations where external events are happening at the same rate or faster than the diffusion process itself. Nonetheless, it make sense to start with these models and investigate how well they can be used to match the data that we observe in urgent contexts. In future work, we will explore modifying these models in order to better account for additional urgent situations. The goal of the current work is to explore what parameter spaces of these traditional models are the best match to our urgent diffusion data. As such, in this paper we will examine two agent-based implementations of two of the more prominent information diffusion models, namely the Bass Model and the Independent Cascade Model. It should be noted that there is one other information diffusion model, the Linear Threshold Model [11][12] that we plan to explore in future work.

3.1 Bass Model

The original Bass model was developed to model the adopting of durable consumer appliances [1], but it can be applied more generally to the diffusion of information. The model is based on the assumption that people get their information from two sources, advertising and word of mouth. The Bass model describes the fractional change in a populations awareness of a piece of information by the equation:

$$\frac{F'(t)}{1 - F(t)} = p + qF(t) \tag{1}$$

$$F(0) = 0 \tag{2}$$

where $F(t)$ is the aware fraction of the population as a function of time, p is the advertising or innovation coefficient, and q is the imitation or word-of-mouth coefficient. Traditionally, q is an order of magnitude greater than p, representing the fact that social communication has a greater effect on adoption decisions than advertising effects. The equation can be interpreted as describing a hazard rate, that is, the conditional probability that a person will become aware of information at time t given that they are not yet aware. In this case, the hazard rate $F'(t)/(1 - F(t))$ is the sum of a constant advertising effect p and a word-of-mouth effect $qF(t)$ that scales linearly in the fraction of population aware.

As is clear, in its current form this is not an agent-based model. However, the model description is easily translated to an agent-based framework, and this has been done before [13]. First, we discretize the problem, giving unaware agents an opportunity to become aware of the information at each time step. Then, instead of determining a deterministic translation of some portion of the population, we update each agents state probabilistically. If every agent observes the actions of

every other agent in the model, then this becomes equivalent to the hazard rate Bass model limited by discretization. However, it is more realistic to consider how information diffuses across a network: instead of allowing each agent to be influenced by the entire population, it is influenced only by its direct neighbors in some underlying social network.

The agent-based Bass model is a discrete-time model in which each agent has one of two states at each time step t: (1) unaware or (2) aware. At the beginning of the simulation, all agents are unaware. At each time step, an unaware agent has an opportunity to become aware. Its state changes with a probability that reflects advertising and word-of-mouth effects. The probability that an agent becomes aware due to word of mouth increases as a function of the fraction of its neighbors who became aware in previous time steps. Once an agent becomes aware, it remains aware for the rest of the simulation.

At each time step, an unaware agent becomes aware due to one of two circumstances:

1. *Innovation* - With probability $\hat{p}$, an unaware agent becomes aware due to outside effects (i.e., information from outside the network) where $\hat{p}$ is the coefficient of innovation.
2. *Imitation* - With probability $f\hat{q}$, an unaware agent becomes aware due to observing the awareness of its neighbors (i.e., information from inside the network) where f is the fraction of neighbors who have adopted and $\hat{q}$ is the coefficient of imitation.[5]

The model then repeats until either all agents have become aware or a fixed number of time steps has been reached.

3.2 Independent Cascade Model

The second diffusion model that we will be considering is the Independent Cascade Model. Examined by Goldenberg et al. (2001), this model was created to understand how information diffuses in a network, and so is very appropriate for the context we are examining. The basic idea behind the Cascade model is that an individual has probability $\hat{q}$ of becoming aware at any time step when m of their neighbors have become aware. There is also a small a probability $\hat{p}$ that the individual becomes aware due to advertising or external news events.[6] The basic intuition behind the cascade model is that information and adoption decisions ripple through a social network in cascades, rather than in long-term exposures such as the Bass model denotes.

For the agent-based model, a population of agents on a network is created. All of the agents are initially unaware then at each time step each agent becomes aware due to two circumstances that parallel the Bass rules:

[5] Since $\hat{p}$ and $\hat{q}$ are not equivalent to the hazard rates, p and q, above, we use the "hat" notation to separate them.

[6] Though clearly the exact values of the Independent Cascade model $\hat{p}$ and $\hat{q}$ will differ, their meaning is very similar. For the sake of brevity we use the same notation for both models.

1. *Innovation* - With probability $\hat{p}$, an unaware agent becomes aware due to outside effects, i.e., information from outside the network, where $\hat{p}$ is the coefficient of innovation.
2. *Imitation* - With probability $\hat{q}$), an unaware agent becomes aware due to observing the awareness of its neighbors, where $\hat{q}$ is the coefficient of imitation.

Again, the model repeats until either all agents have become aware or a fixed number of time steps has been reached.

4 The Data

We have collected four major datasets that we plan to use to examine the effectiveness of these models in understanding the diffusion of information in urgent situations: (1) Osama Bin Laden's capture and death, (2) Hurricane Irene, (3) Hurricane Sandy, and (4) the US 2012 Presidential Election. All of our data was collected from Twitter. Twitter provides two APIs for the collection of data: (1) a Streaming API, which enables the collection of all tweets on a particular topic, or user, going forward, and (2) a RESTful API which enables the collection of a very limited amount of past data, and more importantly network information. Since we need the network information for our models, we first decided to identify a subsample of Twitter users that we would collect full network information about, which would give us the ability to track information diffusion patterns across these networks. To do this, we collected a snowball network sample of 15,000 active, non-celebrity users, including all of the connections between the users. In order to contain noise, we focused on active users that were discovered during our snowball sample. Active users that form part of our dataset issued an average one tweet per day in their latest 100 tweets, and had at least one retweet in their latest 100 tweets.

Once we had established this "15K network", we could then track any number of topics diffusing across the network in one of two ways: (1) *user-focused* - we could simply collect all the tweets that those users issued over a time period, or (2) *topic-focused* - we could collect all tweets on a particular topic, and then filter out the tweets not belonging to our 15K network. Of the four datasets, only the Osama Bin Laden data was collected using the user-focused method, and all other datasets were collected using the topic-focused method. All of this data was collected from Twitter using the streaming API and a tool called TwEater that was developed in-house to collect Twitter data and dump it into a MySQL database or CSV file. TwEater is freely available on Github[7].

Once we had collected all of this data, we then post-processed it to identify the first time any user of the 15K network tweeted about the topic at hand. The time of a user's first tweet about a topic is our estimate of when the user became aware of the event (topic). For quick visualization, we then built network figures showing the spread of information through the 15K network. In these figures each node represents a Twitter user and the edges represent relationships between

[7] http://www.github.com/dmonner/tweater/

the Twitter users. These figures show the flow of new information disseminated through the network to each user, so every edge entering each node except for the earliest one was deleted. To generate these figures, a list of relationships between the users in the collections for each of the events were sorted chronologically. They were then filtered to only include the first tweet received by each user about the event. This list was then imported into Gephi 0.8.2 [14] to generate the figures. The default Force Atlas settings were used, and the program was left to run for several days until changes in the node locations, over the course of 4 hours, were barely noticeable. In the figures, several nodes can be seen floating without any relationship lines attached to them. These floating clusters are two nodes connected by one edge representing a tweeter whose tweet informed only one new person of the event.

Of course, not every member of the 15K network tweeted about every event, and we also made a distinction between people who initially tweeted about an event and those who tweeted only after someone they followed tweeted about the event. In the figures, the announcement about the death of Osama Bin Laden dataset contains 13,842 edges, which come from an initial 1,231 tweeters (See Figure 1). The Hurricane Irene dataset contains 14,373 edges, which come from an initial 814 tweeters (See Figure 2). The Hurricane Sandy dataset contains 14,508 edges, which come from an initial 839 tweeters (See Figure 3). The election dataset contains 13,408 edges, which come from an initial 832 tweeters (See Figure 4).

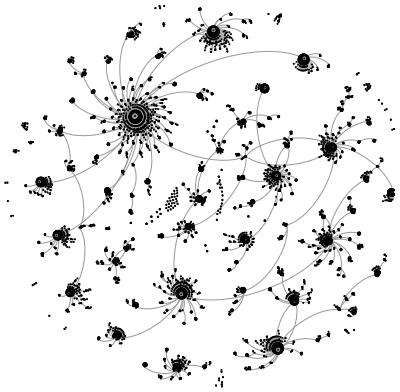

Fig. 1: Visualization of Osama Bin Laden Diffusion.

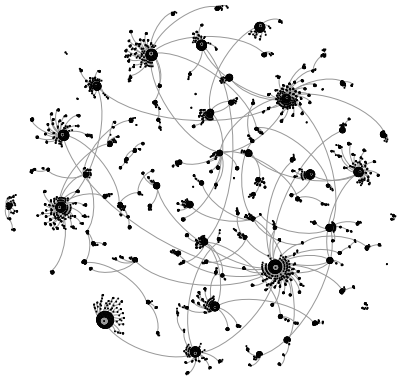

Fig. 2: Visualization of Hurricane Irene Diffusion.

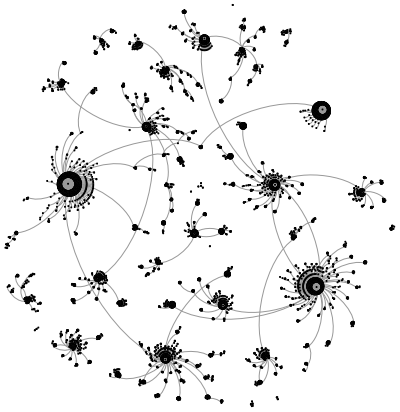

Fig. 3: Visualization of Hurricane Sandy Diffusion.

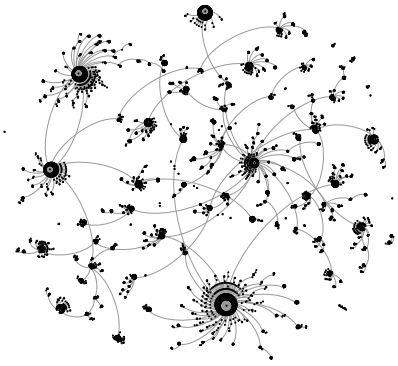

Fig. 4: Visualization of US 2012 Presidential Election Diffusion.

Finally, we cleaned up the data to make it easily comparable to the model results. Using the same data that we used to create the graphs, we placed all of the data into bins of 1 hour intervals. This allowed us to then build a standard adoption curve, where for every hour of the dataset we were interested in there was a single number associated that was the number of cumulative adopters of the information at that point.

5 Results of the Comparison

The next step was to compare the model to the data. Since there was no obvious way to set the $\hat{p}$ and $\hat{q}$ parameters in the models to achieve the best comparison to data. Thus, we conducted a grid search to find parameter values that best matched the underlying data. Initially we searched the space coarsely at 0.01 increments, and then looked at a more detailed level at 0.001 levels. For each model, and for each dataset we ran the model 10 times across this space. We then compared the simulated data to the real data, using Mean Absolute Percentage Error (MAPE), which is the difference between the true vale at time t and the mean value of time t, divided by the true value at time t and averaged over all values:

$$MAPE = \frac{1}{n} \sum_{t=0}^{n} \frac{|true_t - simulated_t|}{true_t} \tag{3}$$

Table 1 gives the range that we investigated for each of the parameters, and the values $(\hat{p}^*, \hat{q}^*)$ which minimized the average MAPE across all ten runs.

model	dataset	$\hat{p}$ range		$\hat{q}$ range		$\hat{p}^*$	$\hat{q}^*$
cascade	bin laden	0.045	0.080	0.040	0.080	0.065	0.060
cascade	irene	0.001	0.034	0.014	0.054	0.014	0.034
cascade	sandy	0.001	0.027	0	0.020	0.007	0
cascade	election	0.014	0.054	0	0.028	0.034	0.008
bass	bin laden	0.079	0.119	0	0.021	0.099	0.001
bass	irene	0.005	0.045	0	0.020	0.025	0
bass	sandy	0.001	0.024	0	0.029	0.004	0.009
bass	election	0.015	0.055	0	0.023	0.035	0.003

Table 1: Range of parameter values and optimum values as determined by lowest MAPE.

The comparison between the underlying data and the model results is visible in the following graphs. Figures 5 and 6 illustrate the two fits to the Bin Laden data. Figures 7 and 8 illustrate the two fits to the Hurricane Irene data. Figures 9 and 10 illustrate the two fits to the Hurricane Sandy data. Figures 11 and 12 illustrate the two fits to the US 2012 Presidential Election data.

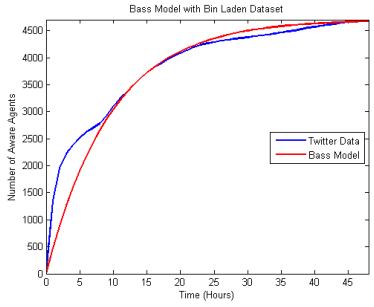

Fig. 5: Comparison of best matching Bass model to Bin Laden data.

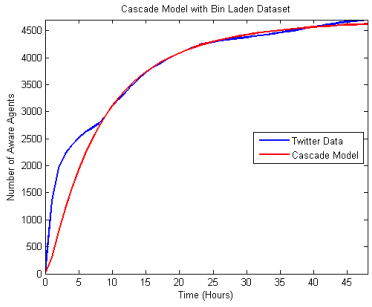

Fig. 6: Comparison of best matching cascade model to Bin Laden data.

To further explore the sensitivity of the model to the parameters that we were exploring, we also explored the full MAPE values for all of the $\hat{p}$ and $\hat{q}$ values around the identified optimal values. Heatmaps of these results are presented in

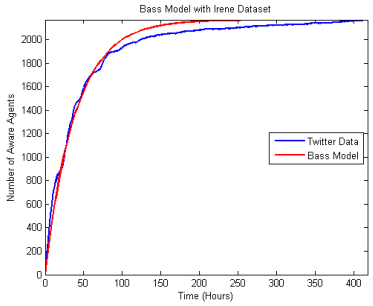

Fig. 7: Comparison of best matching Bass model to the Hurricane Irene data.

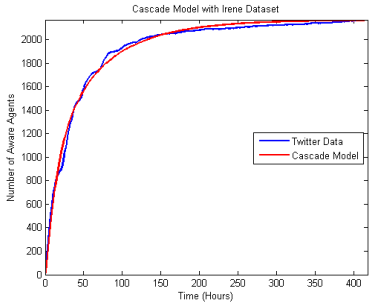

Fig. 8: Comparison of best matching cascade model to the Hurricane Irene data.

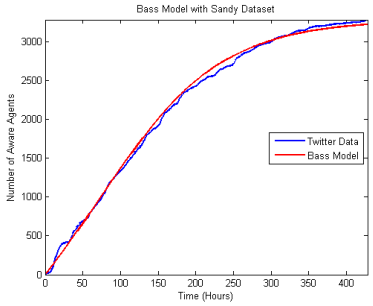

Fig. 9: Comparison of best matching Bass model to the Hurricane Sandy data.

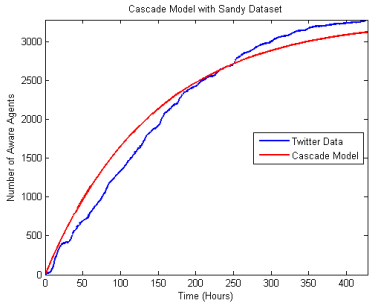

Fig. 10: Comparison of best matching cascade model to the Hurricane Sandy data.

the following figures. Figures 13 and 14 illustrate sensitivity of the Bin Laden data. Figures 15 and 16 illustrate the sensitivity of the Hurricane Irene data. Figures 17 and 18 illustrate the sensitivity of the Hurricane Sandy data. Figures 19 and 20 illustrate the two fits to the US 2012 Presidential Election data. In these heat maps, lower values are darker blues, and higher values are brighter reds, so areas of dark blue indicate areas with minimal errors and, thus, closest fits.

6 Discussion and Future Work

Based on the comparison between the model and real data, It appears that the models can fit the data better for the hurricane cases, which have longer time horizons and, we hypothesize, numerous subevents. In some ways these events are closer to the type of diffusion events that these models were originally created for. The Bin Laden and Election cases have shorter time horizons, and the models do not fit as well. However, despite the fact that these models were

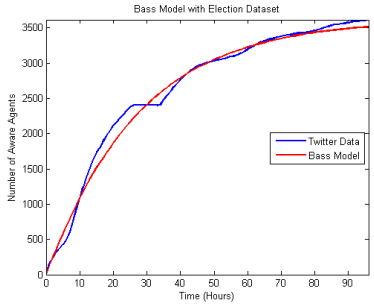

Fig. 11: Comparison of best matching Bass model to the Election data.

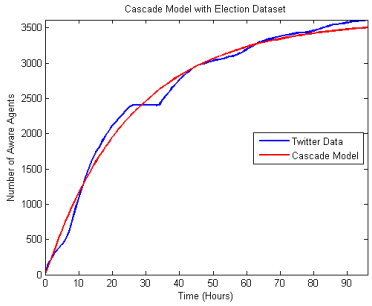

Fig. 12: Comparison of best matching cascade model to the Election data.

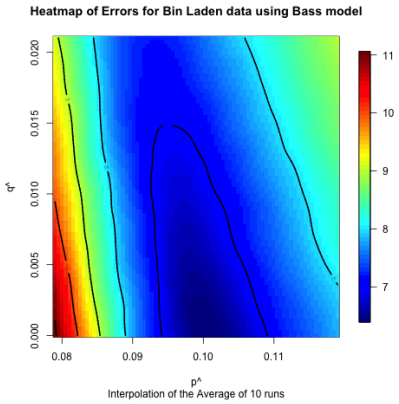

Fig. 13: Heat map illustrating the sensitivity of the Bass model on the Bin Laden data.

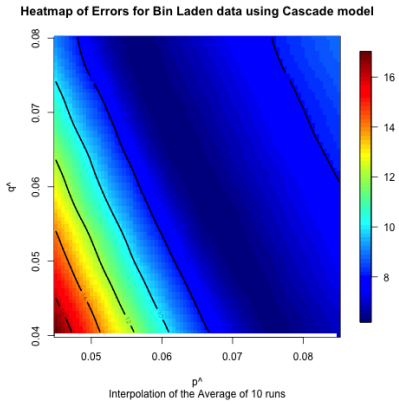

Fig. 14: Heat map illustrating the sensitivity of the Cascade model on the Bin Laden data.

not originally created to examine the diffusion of urgent information, they fit fairly well.

The heat maps also indicate that for most of the models and datasets that we examined there is a range of values that produce similar results. This seems to indicate that though we have identified "optimal" values for each of the network and model combinations, there is a range of values for which the match will be fairly good. This indicates that one does not have to have the values exactly right to get decent predictive value. Moreover, the differences between the Bass and Cascade models, both in terms of the robustness of the results and the best fits, were not very significant, which indicates that of these two models there is not a clear "best" model, so either is appropriate for future exploration. However, it does appear that the event has a large impact on the fit of the results, which seems to indicate that having the ability to classify events ahead of time in order

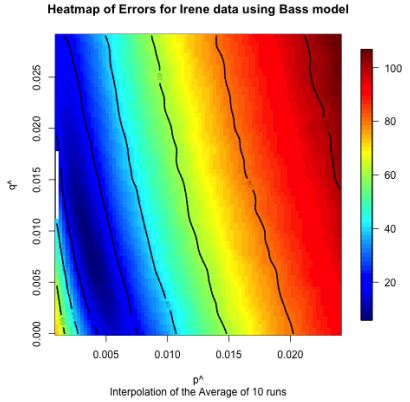

Fig. 15: Heat map illustrating the sensitivity of the Bass model on the Hurricane Irene data.

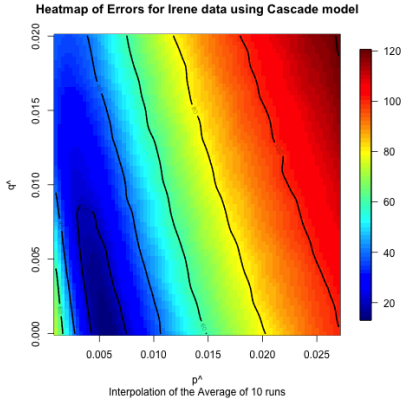

Fig. 16: Heat map illustrating the sensitivity of the Cascade model on the Hurricane Irene data.

to identify appropriate ranges of these values is useful, and is something we will explore in future work.

We hypothesize that we can improve these models in a number of different ways to improve their accuracy. The basic Bass model and the Independent Cascade model were originally created for longer time frames, and urgent diffusion events happen on a much faster pace, especially given the current 24-hour news cycle. In previous work, we have built an equation-based model that takes into account the diurnal cycle of Twitter activity, and achieves much better fits. This is clearly illustrated in the Osama bin Laden case, where the diurnal cycle is readily apparent in Figures 5 and 6. Second, the basic idea behind urgent diffusion is that there are external news events entering the system at a time rate that is faster than the speed of diffusion within the network. If we can model the entrance of this external information, we could also improve the model. We have previously done this in the math-based model, by simply estimating a daily "shock" to the system, and this method could easily be adapted to the agent-based model, though it would introduce a number of new parameters to the system. In future work, we hope to tie the external shocks directly to another data source such as stories in mainstream media.

The overall goal for this project is to provide recommendations for policy makers, brand managers, and other interested parties about how information will diffuse in urgent situations. Eventually, we would like to make these models predictive, by providing a set of guidelines for how to run a model for an event that is in the process of occurring, by tying the type of event to the parameter space. For instance, it is possible that hurricanes always generate "Type 1" events which would be different than major news stories, such as the Bin Laden or Election data, which could be called "Type 2" events. This would then provide interested stakeholders with a set of parameters to feed into the system, giving

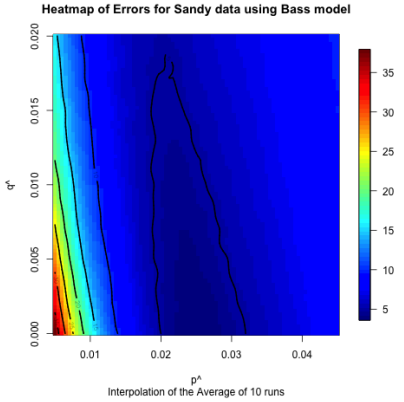

Fig. 17: Heat map illustrating the sensitivity of the Bass model on the Hurricane Sandy data.

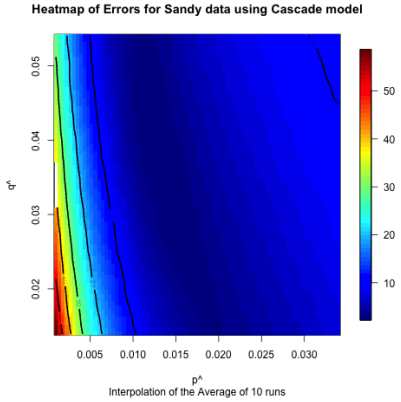

Fig. 18: Heat map illustrating the sensitivity of the Cascade model on the Hurricane Sandy data.

them the ability to predict how fast and to whom information would diffuse. This, in turn, would allow them to make decisions about how to reach out with accurate information and what to expect in terms of overall adoption of a given information topic.

7 Acknowledgments

We thank the National Science Foundation (Award #1018361), DARPA (Award #N66001-12-1-4245) and the Center for Complexity in Business (http://www.rhsmith.umd.edu/ccb/) for supporting this research.

8 ODD Protocol

In this section, we will describe the model in more detail, using a description based on ODD (Overview, Design concepts, Details) protocol for describing individual- and agent-based models [15].

8.1 Purpose

The purpose of this model is to better understand the diffusion of information in urgent situations. The model is intended to be used by any number of stakeholders to understand how a particularly urgent topic might diffuse through a large group of individuals on social media. With this understanding, users might be able to better engage with social media, and better understand real-world reactions to conversations on social media.

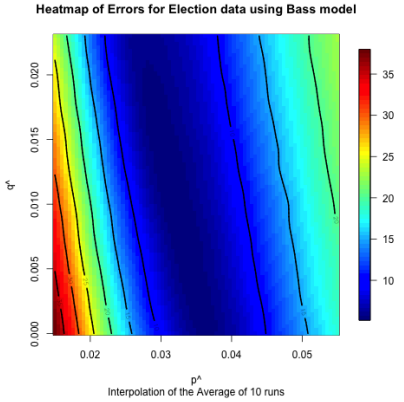

Fig. 19: Heat map illustrating the sensitivity of the Bass model on the Election data.

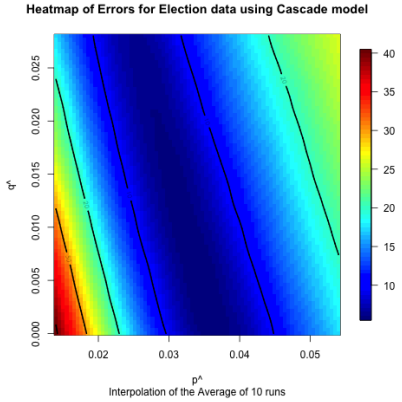

Fig. 20: Heat map illustrating the sensitivity of the Cascade model on the Election data.

8.2 Entities, state variables, and scales

One basic entity in the model is a user of social media who is interesting in consuming and transmitting information. All told these entities exist within the general scope of a social network, via a social media platform. In this paper, we examined individuals involved in Twitter conversations. As a result, another basic entity of the model is the link or relationship between two individuals. These links mean that one user is connected to another user in a way that enables the transmission of information. In the case of Twitter, we use the following relationship as indication of a social connection.

The social media user agents in the model have a number of different state variables. First, they have a property which specifies whether or not they have adopted the new piece of information. In addition, they have both a $\hat{p}$,which is the coefficient of innovation, and $\hat{q}$ which is the coefficient of imitation. These determine how individuals make a decision to adopt or not adopt a new piece of information. All agents in the model also possess a set of links which are the social relationships of that user to other users. In the current version of the model, the links themselves have no properties, but are directed links indicating that one user "follows" another user on Twitter.

As to the scales of the properties and entities, there is a one-to-one mapping between social media user agents and real Twitter users as identified by our 15K collection. This collection is a snowball network sample of 15,000 active, non-celebrity users, including all of the connections between those users. In order to contain noise, we focused on active users that were discovered during our snowball sample. Active users that form part of our dataset issued an average one tweet per day in latest 100 tweets, and had at least one retweet in the latest 100 tweets. Also with respect to scale, the $\hat{p}$ and $\hat{q}$ properties can be considered to be probabilities of an event occurring. Finally, the time step of this model is

one hour. All simulations were run for the length of time that it took to match the underlying data.

8.3 Process overview and scheduling

The underlying process model is fairly simple. The basic idea is that all agents are initialized to an unadopted state, and at the same time the social network between the agents is constructed. Then at each time step any agents which still have not adopted the new information, determine (probabilistically) if they should adopt the new information based on $\hat{p}$ and $\hat{q}$ and the state of their neighbors. This basic overview is true regardless of which model of adoption (Bass or Cascade) is being used; the only thing that changes is the decision rule.

The submodels will be discussed below but the basic processes are:

1. Initialize Network and Agents
2. Repeat for Number of Hours in Event
   (a) For each Social Media User Agent that has not adopted:
      i. Decide to Adopt based on $\hat{p}$ and $\hat{q}$
      ii. Update State of Agent based on Decision
   (b) Update Statistics

8.4 Design Concepts

In this section we will explore the design concepts of the model.

Basic Principles The basic principle of the model is to explore the best way to match urgent diffusion of information in social media networks. The goal is to use this model to help practitioners to make better decisions about how to interact with users of social media in a crisis situation.

Emergence The key aspect of this model that is emergent is the overall adoption rate of the new information. This is modeled by assuming that a small percentage of the population finds out about the new information every time step and that they spread that information via their social networks.

Adaptation Agents in the model are not really adaptive since they do not change as a result of their past experience. Instead they simply react to the presence of information among their neighbors and then decide whether or not to adopt the information.

Objectives The agents do not have particular objectives, but instead simply make decisions about whether or not to adopt novel information.

Learning The agents in the model do not learn.

Prediction Agents do not make any predictions about future states of the world.

Sensing The social media user agents can sense the state of their immediate neighbors that they are following. For instance, they know what fraction of their neighbors have adopted the information, and whether or not it was adopted in the previous time step. $\hat{p}$ can also be thought of as the sensing of an external source of information that informs the agent about the new information.

Interaction The social media user agents interact with each other by exchanging new information. Each agents determines whether or not to adopt the new information based on the submodels described below.

Stochasticity At each time step of the model, each agent who has not adopted the information draws random numbers on the uniform interval [0,1] to determine if they should adopt the information or not. These random numbers are compared to both $\hat{p}$ and $\hat{q}$ to determine if they should adopt the information or not. Because of this each run of the model can result in very different diffusion patterns.

Collectives There is one major collective in the model, which is the social network connecting all of the users. Of course, in the standard social network sense, cliques also exist within the networks which can be thought of as smaller collectives.

8.5 Observation

For the purposes of this paper, the observation we are most concerned with is the overall adoption of the information at each time step of the model run. We are then comparing this information to empirical data to determine which values of $\hat{p}$ and $\hat{q}$ best match the observed empirical data.

8.6 Initialization

The following steps are the major parts of the initialization:

1. Create a number of social media user agents equivalent to the final number that tweet in this dataset
2. Initialize all of the social media user agents to the unadopted state
3. Set the $\hat{p}$ and $\hat{q}$ of all agents to the values currently being explored
4. Connect the social media agents together using the social network data from the 15K network

8.7 Input Data

The only major source of input data is the 15K network data about who is collected to whom in Twitter, based on the snowball sample described in Entities section of the ODD protocol. However, there were also four datasets collected from Twitter by filtering out tweets that matched certain keywords. These datasets were used to establish the validity and examine differences in the underlying model:

1. *Bin Laden* - this dataset examines the 48 hours after Osama Bin Laden's capture and death, and contain all of the tweets that had the words related to this event in them. The dataset contains 13,842 edges, which come from an initial 1231 tweeters
2. *Irene* - This dataset examines roughly a week around the landfall and aftermath of Hurricane Irene. This dataset contains 14,373 edges, which come from an initial 814 tweeters.
3. *Sandy* - This dataset examines roughly a week around the landfall and aftermath of Hurricane Sandy. The dataset contains 14,508 edges, which come from an initial 839 tweeters.
4. *Election* - This dataset examines the 25 hour period from midnight the night before the US 2012 Presidential National election to 1 AM the next morning. The election dataset contains 13,408 edges, which come from an initial 832 tweeters.

Since these datasets describe the total space of people who discussed these topics within the 15K dataset they were used to determine how many agents would be created in each model.

8.8 Submodels

There are two major submodels that are involved in the decision of users to adopt: (1) the Bass model, and (2) the Cascade model.

Bass Model Based on a hazard-rate model that was originally developed to understand the adoption of consumer durables [1], the agent-based Bass model is a discrete-time model in which each agent has one of two states at each time step t: (1) unaware or (2) aware. At the beginning of the simulation, all agents are unaware. At each time step, an unaware agent has an opportunity to become aware. Its state changes with a probability that reflects advertising and word-of-mouth effects. The probability that an agent becomes aware due to word of mouth increases as a function of the fraction of its neighbors who became aware in previous time steps. Once an agent becomes aware, it remains aware for the rest of the simulation.

At each time step, an unaware agent i becomes aware due to one of two circumstances:

1. *Innovation* - With probability $\hat{p}$, an unaware agent becomes aware due to outside effects, i.e., information from outside the network, where $\hat{p}$ is the coefficient of innovation.
2. *Imitation* - With probability $f\hat{q}$, an unaware agent becomes aware due to observing the awareness of its neighbors, where f is the fraction of neighbors who have adopted and $\hat{q}$ is the coefficient of imitation.

Independent Cascade Model The second diffusion model that we examined was the Independent Cascade Model [3]. The basic idea behind the Cascade model is that an individual has $\hat{q}$ probability of becoming aware at any time step when m of their neighbors have become aware. There is also a small a probability $\hat{p}$ that the individual becomes aware due to advertising or external news events. The basic intuition behind the cascade model is that information and adoption decisions ripple through a social network in cascades, rather than in long-term exposures such as the Bass model denotes.

For the agent-based model, a population of agents on a network is created, and all of the agents are initially unaware then at each time step each agent becomes aware due to two circumstances that parallel the Bass rules:

1. *Innovation* - With probability $\hat{p}$, an unaware agent becomes aware due to outside effects, i.e., information from outside the network, where $\hat{p}$ is the coefficient of innovation.
2. *Imitation* - With probability $\hat{q}$), an unaware agent becomes aware due to observing the awareness of its neighbors, where m is the number of neighbors who have adopted in the previous time step and $\hat{q}$ is the coefficient of imitation.

References

1. Bass, F.M.: A new product growth for model consumer durables. Management Science **15**(5) (January 1969) 215–227
2. Trusov, Michael, Rand, W., Joshi, Yogesh: Product diffusion and synthetic networks: Improving pre-launch forecasts with simulated priors. Working Paper (2013)
3. Goldenberg J., Libai B., Muller E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters **12**(3) (2001) 211–223
4. Lerman, K., Ghosh, R.: Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: International Conference on Weblogs and Social Media. (2010)
5. Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., Anderson, K.M.: Natural language processing to the rescue?: Extracting'Situational awareness' tweets during mass emergency. Proc. ICWSM (2011)
6. Yin, J., Lampert, A., Cameron, M., Robinson, B., Power, R.: Using social media to enhance emergency situation awareness. IEEE Intelligent Systems **27**(6) (2012) 52–59
7. Abbasi, M.A., Kumar, S., Filho, J.A.A., Liu, H.: Lessons learned in using social media for disaster relief - ASU crisis response game. In Yang, S.J., Greenberg, A.M., Endsley, M., eds.: Social Computing, Behavioral - Cultural Modeling and

Prediction. Number 7227 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (January 2012) 282–289
8. Shklovski, I., Palen, L., Sutton, J.: Finding community through information and communication technology in disaster response. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work. CSCW '08, New York, NY, USA, ACM (2008) 127136
9. Rogers, E.M.: Diffusion of innovations. Simon and Schuster (2010)
10. Valente, T.W.: Network Models of the Diffusion of Innovations (Quantitative Methods in Communication Series). Hampton Press (NJ)(January 10, 1995) (1995)
11. Granovetter, M.: Threshold models of collective behavior. American journal of sociology (1978) 1420–1443
12. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences **99**(9) (2002) 5766–5771
13. Rand, W., Rust, R.T.: Agent-based modeling in marketing: Guidelines for rigor. International Journal of Research in Marketing **28**(3) (2011) 181–193
14. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: ICWSM. (2009)
15. Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F.: The odd protocol: a review and first update. Ecological Modelling **221**(23) (2010) 2760–2768