Modeling Mass Protest Adoption in Social Network Communities using Geometric Brownian Motion

Fang Jin[*], Rupinder Paul Khandpur[*], Nathan Self[*], Edward Dougherty[†], Sheng Guo[‡],
Feng Chen[§], B. Aditya Prakash[*], Naren Ramakrishnan[*]

[*]Discovery Analytics Center, Department of Computer Science, Virginia Tech.
[†]Genetics, Bioinformatics, and Computational Biology Program, Virginia Tech.
[‡]LinkedIn Inc., [§]Department of Computer Science, University at Albany, SUNY
[*]{jfang8, rupen, nwself, badityap, naren}@cs.vt.edu
[†]edougherty@vt.edu, [‡]sguo@linkedin.com, [§]fchen5@albany.edu

ABSTRACT

Modeling the movement of information within social media outlets, like Twitter, is key to understanding to how ideas spread but quantifying such movement runs into several difficulties. Two specific areas that elude a clear characterization are (i) the intrinsic random nature of individuals to potentially adopt and subsequently broadcast a Twitter topic, and (ii) the dissemination of information via non-Twitter sources, such as news outlets and word of mouth, and its impact on Twitter propagation. These distinct yet interconnected areas must be incorporated to generate a comprehensive model of information diffusion. We propose a bispace model to capture propagation in the union of (exclusively) Twitter and non-Twitter environments. To quantify the stochastic nature of Twitter topic propagation, we combine principles of geometric Brownian motion and traditional network graph theory. We apply Poisson process functions to model information diffusion outside of the Twitter mentions network. We discuss techniques to unify the two sub-models to accurately model information dissemination. We demonstrate the novel application of these techniques on real Twitter datasets related to mass protest adoption in social communities.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Information diffusion; social networks; protests; geometric Brownian motion.

1. INTRODUCTION

In recent years social networking sites such as Twitter and Facebook have provided not just a platform for communication but also a means of mobilization and strategic interaction between key players of social movements, e.g., protests. Traditionally social movements occur within a subset of the population and have spread through on-the-ground communities and unions. With the advent of leaner communication technologies like Twitter, the way such movements form and spread through modern society has changed. With Twitter, in particular, traditional slogans have transformed into hashtags which can offer a consistent way of communicating the reason and motivation of social movements like protests and uprisings.

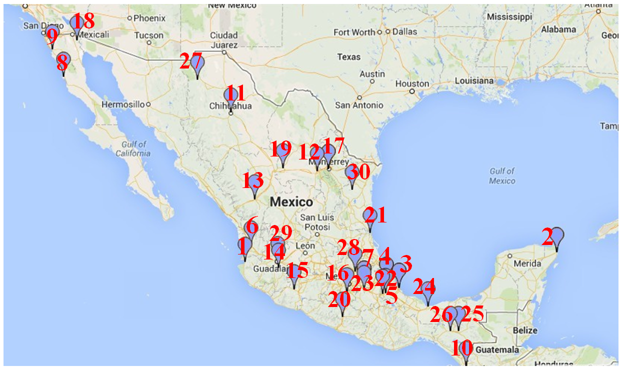


Figure 1: Mexico teacher protest events from Sep 1 to Sep 7, 2013. The blue pins denote protest cities; the numbers in red denote the sequence of protests as they spread across the country.

In this paper, we focus on Twitter's user networks during protests and similar civil unrest activities in Latin America. Our goals are to model the propagation and growth of contagion-like protest waves within a social network and to understand the social and structural dynamics underlying such phenomena. The key problem is understanding the nature of information propagation among motivated users of a social network. We have observed that such mass protests emerge very swiftly and sharply. In Twitterspeak, they would be considered trending but most such trends quickly decline on the social network even if not in the physical world. Modeling protest-related topic propagation on networks involves several challenges.

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623376.

First, social protest propagation through online media can spread over large areas more quickly than traditional methods since users are geographically distributed. For example, on September 1, 2013, the Mexican government's education reform bill drew the wrath of teachers country-wide who opposed the reform (which required regular assessments of their performance as educators). Twitter was a virtual loudspeaker, providing a platform for organization and strategization for teachers to put forth their arguments against the bill. A series of mass teacher protests erupted and spread from city to city. As shown in Fig. 1, we see the movement spreading over time to different locations with no obvious visual mobilization pattern. The second challenge is that Twitter's user network embodies many subgraphs based on social ties which might afford different propagation rates due to subgraph-specific structures.

Thus identifying how the cause of a protest is adopted by Twitter users and how mobilization happens in the underlying network is a difficult task. To address this problem we present an integrated framework with new theoretical models as well as empirical validation on real Twitter data for actual protests witnessed in the recent past. Our key contributions are:

- We model the inherent heterogeneity in propagation using a bispace model, comprised of the Twitter mentions network (where both globally and locally influential neighbors contribute to a user's recruitment) and a latent space (where external exposure to protest-related information is captured).
- We focus on the role of community-driven information propagation over the bispace model. We use geometric Brownian motion (GBM) over the mentions network and Poisson processes over the latent space to model information propagation during mass social movements.
- We illustrate the effectiveness of our approach in modeling several key mass protest adoption scenarios in multiple countries of Latin America, viz. Argentina, Brazil, Colombia, Mexico, Uruguay, and Venezuela.

The rest of this paper is organized as follows. Section 2 covers related work in the areas of social movements, information diffusion in networks, external influences, and Brownian motion. Section 3 proposes the geometric Brownian motion propagation mechanism. Section 4 introduces the bispace propagation model, especially the model of propagation in latent space. In Section 5, we present our dataset and experimental setup, followed by initial experimental findings. Section 6 discusses the evaluation results for our approach followed by a brief discussion in Section 7.

2. RELATED WORK

We briefly review related work next, which comes from multiple areas.

Social movements: Oliver and Myers [18] develop a foundation for theoretical insights of social movements and describe the limitations of simplified models. The Arab Spring of 2010 served as a context for many researchers [6, 2, 24, 4, 20] to study the role social networking sites play in the spread and recruitment of participants in protests. A detailed anatomy of modern social protests is described by Saad-Filho [20] with the June 2013 anti-government protests in Brazil as a context. In this work, we study the processes and sociological impacts of protests in the modern era, fortified by online social networks and the communities in and around them.

Information diffusion in networks: Previous studies have approached the modeling of information propagation and diffusion in social networks through several means, e.g., contagion models (SIR [3] SISa [10]), diffusion based threshold and cascade models [12], rise-and-fall patterns [13], coverage models [22], and survival theory [19]. A good survey of different models of information diffusion is presented in [7].

External influences: We believe that the effects of influences that originate external to the observed diffusion network, such as mass media and offline spread of information, can impact the way in which information flows within the online network. Myers et al. [14] study the emergence of URLs on Twitter with a probabilistic generative process using both internal and external exposure curves in a contagion-like model. Similar attention to the role of external factors is paid by Crane and Sornette [5] for tracking the popularity of YouTube videos using a diffusion model. Iwata et al. [11] use a shared cascade Poisson process model to discover latent influences in social activities such as item adoption. Using shared parameters among multiple Poisson processes, they were able to simulate sequences of item adoption events.

Brownian motion: Zhou and colleagues (e.g., [26, 8, 27]) develop the notion of Brownian motion on networks which they use to discover communities of hierarchical structure both locally and globally. We extend this approach in this paper to formulate a propagation algorithm based on geometric Brownian motion (GBM). Borrowed from statistical physics, GBM has been used heavily in finance to model stock price movements. Scale invariance and the ability to model abrupt bumps along propagation paths are the primary motivations for using GBMs to model stochastic processes [23].

Our work builds on the concepts introduced in [8, 11, 26, 27] but differs from the other diffusion models described earlier by considering both the role of communities of users and the abrupt nature of propagation of volatile information such as mass social protests. We include the notion of bispace where both latent (attributed to external influences) and observed user network influences are considered. We infer propagation rates for communities in the observed network and allow implicit recruitment of users into protest actions through a Poisson process.

3. FORMALISMS

3.1 Basics

We model Twitter activity as a network $G(V, E)$ of mentions. Here, each vertex $v \in V$ represents a Twitter user. There is a directed edge from user v_i to user v_j if v_i mentions v_j in a tweet. We define ω_{ij} to be the number of tweets in which user v_i mentions user v_j. Note that ω_{ij} is not necessarily equal to ω_{ji}. Key players such as celebrities and politicians are more likely to be mentioned by other users, rather than the other way around. As can be seen in Fig. 2, the mentions network is a directed graph. Weight w_{14} is the number of times Twitter user v_1 mentions user v_4, which is

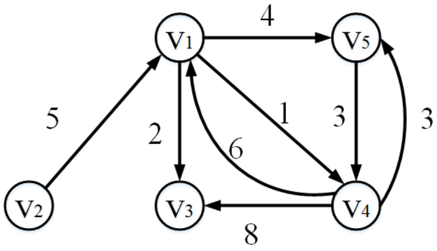


Figure 2: An example mentions network. Nodes denote Twitter users, directed edges denote direction of mentions between users, and edges are labeled with mention frequency.

1, while w_{41} is 6. Note that w_{21} is 5, while w_{12} is 0 (not shown).

We define the neighborhood $N(v_i)$ of a user v_i as the set of all users mentioned by v_i, i.e., those for whom there is a directed edge from v_i. For each user $v_j \in N(v_i)$, we define the Brownian distance from user v_i to v_j to be

$$d_{ij} = \frac{1}{(\omega_{ij}+1)(\omega_{ji}+1)^{\gamma}(\eta_{ij}+1)^{\gamma}} \tag{1}$$

Here, η_{ij} is the number of common direct neighbors shared by user v_i and user v_j [27]. In Fig. 2, node v_1 and v_4 share two common direct neighbors—v_3 and v_5—and hence η_{14} is 2.

We use the bias coefficient $\gamma \geq 1$ to heuristically weigh mentions that carry more impact. If v_i mentions v_j, meaning that $\omega_{ij} > 0$, we believe this expresses v_i's intention to propagate information to v_j. Since v_j may not know or care about v_i and consequently may seldom or never mention v_i, the return mentions, measured by ω_{ji}, are (up)weighted by γ. Furthermore, if v_i and v_j share neighbors in the mentions network, the two users may have a closer relationship than other users with no shared mentioned Twitter users, and thus this component is weighted by γ as well. A Laplacian-style (+1) correction is used when there are no counter mentions or no mutual mentions. Note that for $\gamma = 1$, d_{ij} is an unbiased Brownian distance since ω_{ij}, ω_{ji}, and η_{ij} will have the same weight.

3.2 Trust functions and GBM

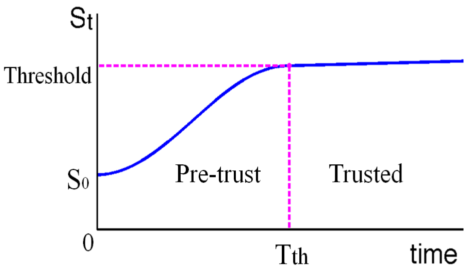


Figure 3: Trust function. A threshold defines the transition between the pre-trust and trusted period.

Next we introduce the notion of a trust function S_t which we use to model an individual user's agreement with an idea as expressed in tweets. (The trust function S_t is a function of the two entities between whom trust is modeled, but in this section we simplify the notation for ease of exposition.) We divide the trust process into a pre-trust period and a trusted period. In the pre-trust period, as a user receives new information, that user's trust, S_t, increases exponentially until S_t reaches the trust threshold at time T_{th} and enters the trusted period. In the trusted period, new information increases S_t linearly. For simplicity, an individual user cannot revoke trust once this threshold has been crossed. In our Twitter mentions network, a user's trust in a topic crosses the threshold when they have tweeted about it. During the pre-trust period, we model the trust function as follows (the coefficient μ accounts for change in the average value of this stochastic process):

$$\frac{dS_t}{S_t} = \mu dt \tag{2}$$

We then add a Wiener process W_t to account for stochasticity. According to the properties of a Wiener process [17], dW_t is essentially Gaussian white noise and contributes to our equation as:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t \tag{3}$$

In this way, we modeled the trust function S_t as a geometric Brownian motion (GBM) process which is a continuous-time stochastic process [17]. Per convention, we call μ the drift and σ the volatility. The drift represents deterministic trends while the volatility refers to the influence of unpredictable events in this model [25]. For simplicity, we consider μ and σ to be constant during the pre-trust period in this paper. (Our concern here primarily is with this period.)

According to Itō's theorem [17], given the initial value S_0, the above stochastic differential equation has the following analytic solution:

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right) \tag{4}$$

The above solution for S_t is a log-normally distributed random variable with expected value and variance given as [17]:

$$E(S_t) = S_0 e^{\mu t} \tag{5}$$

$$Var(S_t) = S_0^2 e^{2\mu t}\left(e^{\sigma^2 t} - 1\right) \tag{6}$$

S_t is a geometric Brownian motion stochastic process, which is typically denoted as $\mathcal{B}(\mu, \sigma)$. In this paper we use an initial trust of $S_0 = 1$ without loss of generality.

3.3 GBM propagation

Suppose that user v_i posts a protest-related tweet at time t_0 which indicates that v_i has been recruited or infected. Whether v_i will infect its neighbor v_j depends on v_j's trust function with v_i. For instance, if v_j is a close friend of v_i, then it is more likely that v_j will be infected in a short time because of v_j's trust in v_i. But if v_j is not a very close friend of v_i, then it might take a long time to build v_j's trust with v_i and to accept v_i's status. Only after v_j's trust with v_i crosses some threshold, v_j gets infected.

For better quantitative analysis, we consider d_{ij} to be the trust threshold. After crossing this threshold, v_j will agree with v_i's opinion. According to the properties of GBMs, the trust function S_t grows continuously over time. This implies

```
input  : mentions network $G(V,E)$, time step $\delta t$,
         propagation time $T$
output: infected users
for each infected user $v_i \in V$ do
    for each non-infected user $v_j \in N(v_i)$ do
        set $t_{ij} = 0$;
    end
    set $v_i$ as not newly infected user
end
t = 0;
for $t \leq T$ do
    for each infected user $v_i \in V$ do
        if $v_i$ is a newly infected user then
            for each non-infected user $v_j \in N(v_i)$ do
                set $t_{ij} = 0$;
            end
            set $v_i$ as not newly infected user
        end
        for each non-infected user $v_j \in N(v_i)$ do
            set $t_{ij} = t_{ij} + \delta t$
            $ln(S_t^{ij}) \sim \mathcal{N}((\mu - \frac{\sigma^2}{2})t_{ij}, \sigma^2 t_{ij})$
            if $ln(S_t^{ij}) \geq d_{ij}$ then
                set user $v_j$ as newly infected
            end
        end
    end
    $t = t + \delta t$;
end
```

Algorithm 1: GBM propagation algorithm

that, if some user is infected, all of that user's neighbors will eventually get infected given enough time for diffusion.

Since we assume a user cannot revoke trust, his or her status will never change once infected. Based on the above assumptions, we now detail our process for GBM propagation through the mentions network; see Algorithm 1. Since GBM is a time-continuous stochastic process, we discretize time using time steps of duration δt each. At the start of the simulation, all infected users are considered as newly infected users. Assume that the complete mass protest propagation duration is T. Once a user v_i becomes infected, the node is marked as a newly infected user, and the new status begins to affect the statuses of the neighbors, i.e., $N(v_i)$. For each user $v_j \in N(v_i)$, we use $t_{ij} = 0$ to initialize the time instant from which v_i begins to affect v_j. After all the time variables t_{ij} of $N(v_i)$ are so initialized, user v_i's status is updated to reflect that v_i is no longer a newly infected user, to avoid duplicate initializations.

Suppose that at current time t, v_j's trust with v_i is denoted as S_t^{ij}. According to the GBM properties, $ln(S_t^{ij})$ is a Gaussian variable given by:

$$ln(S_t^{ij}) \sim \mathcal{N}((\mu - \frac{\sigma^2}{2})t, \sigma^2 t) \quad (7)$$

If at time t, $ln(S_t^{ij}) \geq d_{ij}$, this means that v_j gets infected since v_j's trust with v_i is bigger than the distance d_{ij}. Now v_j begins to affect his or her own neighbors. Instead at time t, if $ln(S_t^{ij}) < d_{ij}$, then at the next time step, $t+\delta t$, the trust is still a Gaussian variable, but with higher expectation and variance:

$$ln(S_{t+\delta t}^{ij}) \sim \mathcal{N}((\mu - \frac{\sigma^2}{2})(t+\delta t), \sigma^2(t+\delta t)) \quad (8)$$

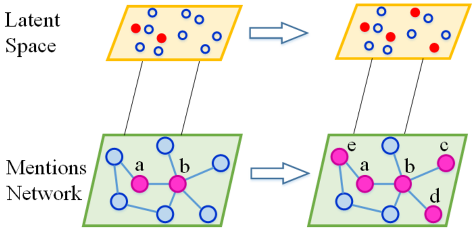


Figure 4: Bispace propagation model. In the latent space, users infections are explained by a Poisson model, and the red nodes denote the infected users from one time step to another. In the mentions network space, users are infected according to the GBM model. Here, the purple nodes (a, b, c, d, e) denote user infections explained by the GBM model.

3.4 GBM parameter estimation

We use past protest events in which Twitter played a significant role in propagation to train our GBM model parameters. For each user who gets infected we record their Brownian distance and infection time. Suppose v_j gets infected by v_i after time t_{ij}; then as per our propagation model, we claim that v_j's trust function S_t^{ij} with v_i holds:

$$ln(S_t^{ij}) \geq d_{ij} \quad (9)$$

where d_{ij} is the Brownian distance from v_i to v_j. For the convenience of parameter estimation, we can assume that $ln(S_t^{ij}) = d_{ij}$. It then follows that d_{ij} is a normally distributed random variable which can be expressed as:

$$d_{ij} \sim \mathcal{N}((\mu - \frac{\sigma^2}{2})t_{ij}, \sigma^2 t_{ij}) \quad (10)$$

Because during the parameter estimation process, for each infected user v_j, we are not interested in exactly which user gets v_j infected, we use $x_j = d_{ij}$, and $\tau_j = t_{ij}$ in the following part of this section for simplicity. The set of n users that are infected during the infection process have independent infection rates, and we get the following likelihood function:

$$\mathcal{L}(\theta, \sigma^2 \mid v_1, \dots, v_n) = \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi\tau_j}} exp(-\frac{(x_j - (\mu - \frac{\sigma^2}{2})\tau_j)^2}{2\sigma^2\tau_j})$$

The optimal estimators can be obtained by maximizing the above likelihood function. We differentiate the natural logarithm of the likelihood function above in terms of μ and σ, and set them to zeros. By solving the two equations simultaneously, we obtain the optimal estimators $\hat{\mu}$ and $\hat{\sigma^2}$.

4. BISPACE PROPAGATION MODEL

Many information diffusion models assume that propagation occurs over a single domain. However, it is hard to build a complete, exhaustive network of interactions. For instance, consider building a network based only on which Twitter users follow which other users. This network will miss interactions such as retweets and mentions and the effect of influences originating outside of Twitter. Therefore, considering only a single space will make it difficult to account for all possible factors that influence the spread of information. In this study, we propose a bispace diffusion

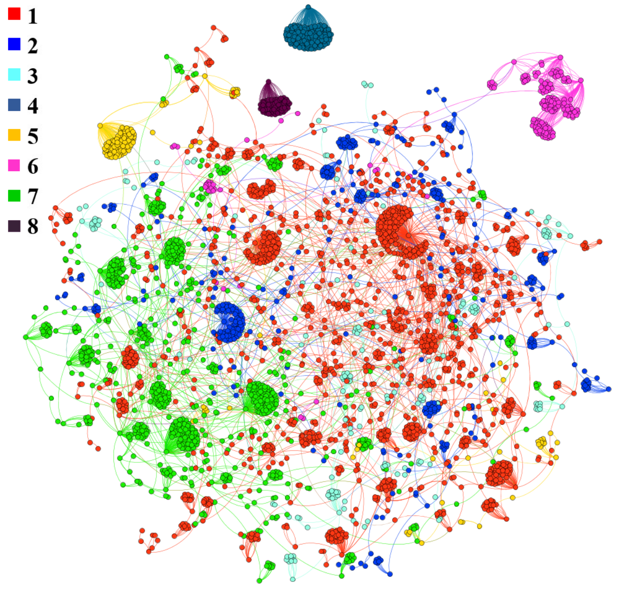


Figure 5: Major communities of teacher protest events (Sep 1 to Sep 12, 2013, Mexico).

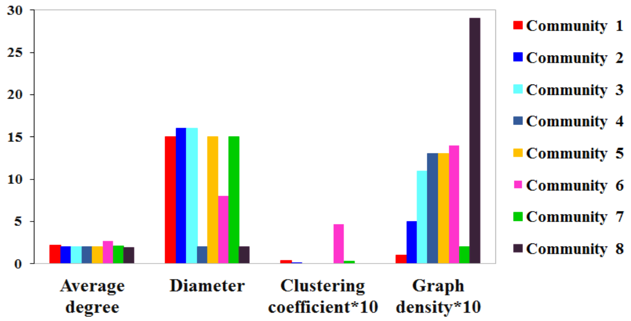
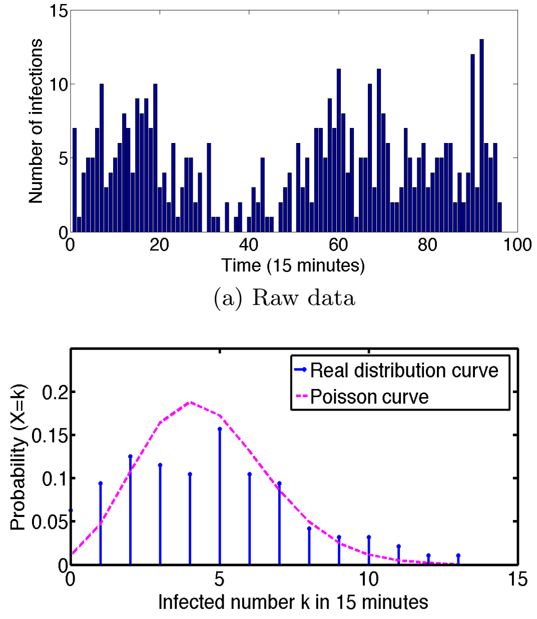


Figure 6: Key graph properties of communities underlying the Mexican teacher protest events.

model that accounts for two domains of diffusion: the observed social network and the latent space, as can be seen in Fig. 4. In our case, the observed user space is the Twitter mentions network, whereas the latent space refers to any interactions outside of this network. To account for varying diffusion dynamics, each space is intended to have its own propagation model. As described earlier, we model propagation through the Twitters mentions network as Geometric Brownian motion. We use the Poisson distribution to describe information propagation in the latent space.

4.1 GBM with Communities

Within networks, a community refers to the appearance of densely connected groups of vertices, with sparse connections between each group [16]. Instead of treating the whole network as a single propagation space, we use network structure to further split the network into communities. For our mentions network we use the Louvain method [1] for community detection to split the network into groups of users. For each community of users we can calculate classical graph features such as average degree, diameter, density, and clustering coefficient with which we can characterize them. In Fig. 6 we plot several features for each of the 8 communities found in the case study of Mexican teachers protest of 2013. Diameter $r = \max dist(v_i, v_j)$ is the length (in number of edges) of the longest geodesic path between any nodes v_i and v_j [15]. The clustering coefficient c_i is the proportion of node v_i's neighbors that are connected. Graph density is defined as $\frac{2|E|}{|V|(|V|-1)}$ where E is the number of edges and V is the number of nodes [21]. As shown in Fig. 6, diameter and graph density vary considerably.



(a) Raw data

(b) Poisson distribution model

Figure 7: Poisson distribution in latent space propagation. (a) shows the raw data outside of the mentions network of teacher protest events on Sep 3, 2013. (b) shows the probability distribution of the number of infections.

With the observed network further split into several communities, each community is intended to have its own model parameters for GBM. In GBM, $ln(S_t^{ij})$ is a Gaussian distribution $\mathcal{N}((\mu - \frac{\sigma^2}{2})t, \sigma^2 t)$. We assume that each user within a community shares the same μ and σ so that each community has its characteristic μ and σ. As information propagates through the mentions network, it may pass through different communities. For an infected user v_i and one of the non-infected neighbors $v_j \in N(v_i)$, we assume the following propagation strategy:

- If v_i and v_j are in the same community c_i, the propagation process will follow $\mathcal{B}_{c_i}(\mu_{c_i}, \sigma^2_{c_i})$.

- Propagation from one community to another happens as per the source community's model parameters. For instance, for propagation from community c_i to community c_j, we will use the source community c_i's GBM parameters.

- After information propagates into a different community, it will spread according to the new community's

Table 1: Mass protests studied in this paper.

No.	Event	Hashtags	Country	Affected cities	Event date(s)
1.	YoSoy132 student movement	#LaMarchaYoSoy132, #YoSoy132, #132, #soy132	Mexico	Nationwide	2012-05-17 to 2012-05-25
2.	Anti-government protests against tax reform and other policies pursued by President Juan Manuel Santos	#CacerolazoPaSantos, #5D	Colombia	Nationwide	2012-12-05
3.	Education reform protests by teachers	#ReformaEducativa	Mexico	Nationwide	2013-09-01
4.	Social protests against violence and crime	#UruguayosIndignados, #HartosDeLaViolencia	Uruguay	Montevideo	2012-05-14
5.	Protests against the "media law"	#LorenzettiNoMeFalles, #MediosBuitres	Argentina	Buenos Aires	2012-11-27
6.	Protests against Senate President Renan Calheiros's election	#STFjulgueRenan, #SocorroJoaquim, #ForaRenan	Brazil	Nationwide	2013-02-22 to 2013-02-26
7.	Anti-government student protests against abuse of public media for election campaign	#ConatelCareTabla	Venezuela	Caracas	2013-03-20

parameters. Once the information has entered community c_j from community c_i, subsequent infections henceforth will use community c_j's parameters.

At each time step we use the μ and σ of any given node's current community for propagation from that node.

4.2 Propagation in Latent Space

As mentioned before, in the latent space, we are modeling unobserved interactions of users. Since there are so many factors that might affect the dissemination of information, such as news outlets, word-of-mouth, it is reasonable to assume that the probability of the number of newly infected users in a given time interval satisfies the Poisson distribution [9] in the latent space.

For each node in the mentions network, it can only be infected by the GBM process. However, for those isolated users outside the mentions network, it is only possible that they get infected via the mechanics of the Poisson process. (Recall that in the GBM process, users get infected primarily via their neighbors.) We use X to represent the number of infected users with time interval δt and so the probability of the infected users is given by:

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{11}$$

To obtain an estimator of λ, we can only use information about Twitter users who are outside the mentions network as our training dataset. We count the infected users outside the mentions network with time interval of 15 minutes during the Mexican teachers protest, and plot them as shown in Fig. 7(a). Adequately modelable by a Poisson distribution, we use the average value as the estimate of λ. Fig. 7(b) depicts the Poisson distribution fit with $\hat{\lambda} = 4.18$. If there are M_0 isolated users, the probability of each of these users to get infected in time interval δt is λ/M_0. To summarize, for any user not in the mentions network, infection is only possible via the Poisson process. For a user who is already in the mentions network, infection can only happen via the GBM process over the mentions network, as described earlier.

5. EXPERIMENTS

5.1 Dataset description

The study described in this paper uses two datasets: (i) a gold standard report (GSR) of social unrest events in Latin America provided by MITRE that we use to define major mass protest events, and (ii) tweets collected over 14 months from May 2012 to September 2013 from 20 Latin American countries.

The GSR documents each civil unrest event by location, date, type of protest, and specifies the national news articles that first reported the event. For protests that were prominent on Twitter, the GSR news articles often report hashtags which were used by protestors on social media. We selected only those GSR events for which we were able to find such hashtags. This process resulted in 64 unique hashtags related to 40 different protest that occurred in Latin America since May 2012. In Table 1 we list a few of these events from our study.

Our Twitter dataset was built by querying Datasift's streaming API. Each tweet payload includes crucial metadata along with the tweet's content. Though tweets from GPS-enabled devices include geographic coordinates, the percentage of such tweets in the collected sample was too low to be useful.

For this study, we further filtered tweets by removing those that do not contain hashtags relevant to a specific protest. Since most tweets do not have location data, we estimate their location by geocoding the tweet based on each tweet's content and properties of its user. We developed our own geocoding library that uses the World Gazetteer (http://archive.is/srm8P) database to lookup location names and geographic coordinates. Tweets can be geocoded to the user's location at the time of tweeting or a location of interest about which the user is tweeting. We focused on event ge-

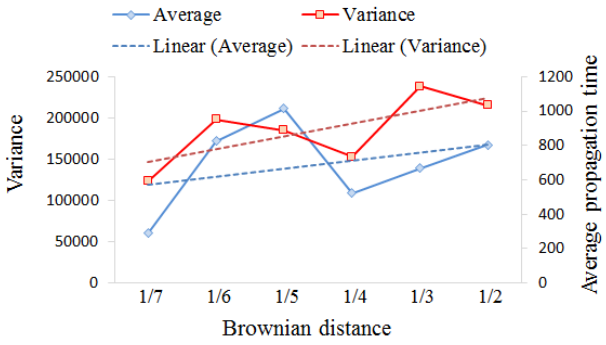


Figure 8: Brownian distance vs propagation time for teacher protest events.

olocation, which looks for location or landmark names, such as *Plaza de la Independencia* or *Quito, Ecuador*, in a tweet's text. We generated a list of 2000 landmarks by extracting place names mentioned in GSR events which had high mutual information to civil unrest. In cases where no event location was found in a tweet's text, we use geo-coordinates or self-reported location string in the tweet's metadata.

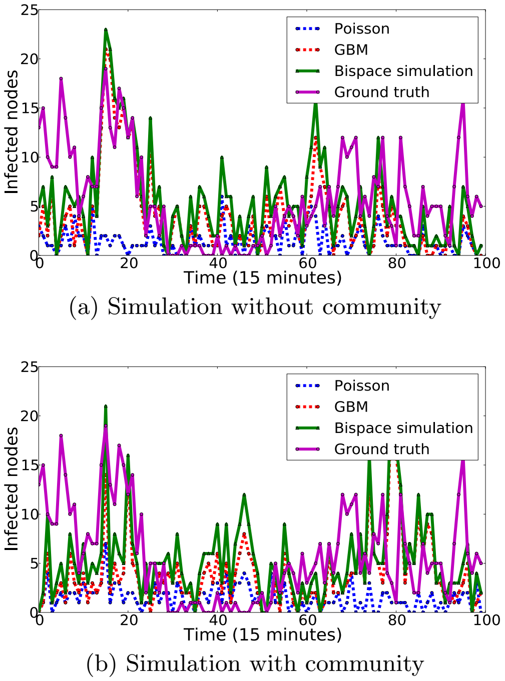
Using the above pipeline we were able to extract and geolocate 20, 227, 830 unique users to build our mentions network from the filtered tweets that were spread over daily sub-networks.

5.2 GBM Diffusion Model

For each of our mass protest events, we filter by its specific keywords (hashtags) to obtain a set of relevant tweets and construct a mentions network from those tweets. We assume that information propagates from an initial infected user to other users through the network from one node to its neighbors. We build an adjacency matrix based on the mentions network and simulate the propagation using the GBM diffusion process as follows:

1. **Brownian distance:** The Brownian distance is intended to have an inverse relationship with mention frequency. As Fig. 8 shows, users with smaller Brownian distance have greater mention frequencies resulting in shorter mean propagation times with less variance. From Fig. 8, we can see that infection time and variance generally both increase with an increase in Brownian distance. Heuristically, more frequent mentions indicate stronger ties which leads to easier adoption of information.

2. **Propagation speed:** To evaluate our dynamic GBM infection process assumptions, we estimate the GBM parameters for different protest events and depict the GBM propagation curves in Figs. 9, 10, and 11. The blue curve depicts the Poisson propagation in latent space. The red curve depicts GBM propagation through the mentions network. The green curve is the overall simulation result while the magenta curve depicts the ground truth of the protest events process. By comparing the green and magenta curves, we can evaluate the effectiveness of our bispace model in simulating the mass protest events. As shown, we find that, given a mentions network, our bispace model can simulate the propagation speed at a reasonable scale, at the right magnitudes. As seen in Figs. 9(a), 10(a) and 11(a), we find that we can capture the burst of activity at the same time point as the ground truth during protest propagation.



(a) Simulation without community

(b) Simulation with community

Figure 9: GBM and Poisson propagation simulation for Yosoy protests (Mexico) on May 19, 2012.

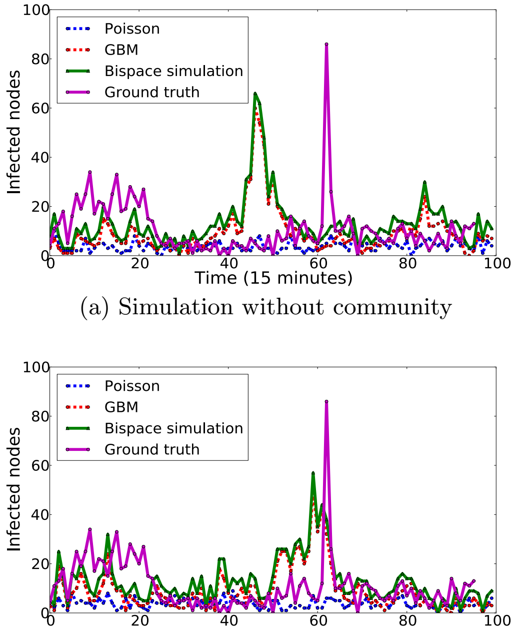
5.3 GBM Diffusion with Communities

We were also able to observe the variation in μ and σ as community structure varies. In particular, community features like graph density and diameter as shown earlier in Fig. 6 may impact GBM propagation. We experimented with two modeling approaches: (i) one set of parameters for the whole network and (ii) different parameters for each community in the network. We ran simulations for both these situations, and plotted the results of the whole network vs. community-specific approach in Figs. 9, 10, and 11. Comparing these simulation results, we find the community approach performs better, especially at capturing peak values. Taking a closer look at Fig. 12, we observe that propagation time and speed of infection are different for each community and we are able to simulate local propagation more accurately, which can be seen, e.g., from Fig. 10(b), where the GBM with community method can simulate the burst propagation effectively, while the general GBM method (see Fig. 10(a)) fails to capture the exact peak time.

5.4 Latent Space Diffusion Model

We use the following steps to calculate the properties of the latent space for each event.

1. **Latent space:** The intent is to consider all possible external influences and latent interactions in this space. We split Twitter data into unique 15 minute



(a) Simulation without community

(b) Simulation with community

Figure 10: GBM and Poisson propagation simulation for teacher protests (Mexico) on Sep 1, 2013.

intervals and count the total number of infected users in each interval.

2. **Normalize:** Twitter user activity varies based on time of day and day of week (see Fig. 13). For each 15 minute window from Step 1, we find the average number of tweets over a 4 week period and use this value to normalize the count. This baseline count of tweets over time in the latent space is close to the Poisson distribution. Fig. 7(a) shows an example of this baseline.

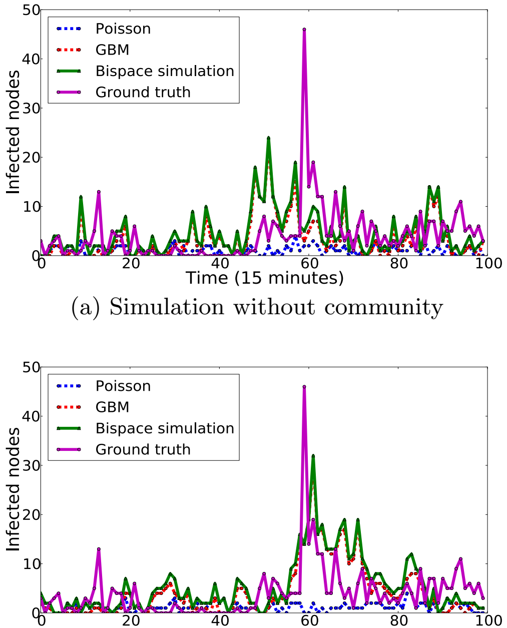
3. **Train:** Using one week's data split into 15 minute intervals, we train the Poisson distribution parameters. Fig. 7(b) shows that the training curve and ground truth curve can be matched quite well.

6. EVALUATION RESULTS

We present an exhaustive evaluation of our bispace simulation approach alongside various dimensions next:

- **How effective is the performance of the bispace model?**

Recall that the bispace modeling is comprised of two independent process: the GBM simulation in the mentions network, and the Poisson process within the latent space. Given an initial mentions network, after training the GBM parameters of μ and σ, we proceed to conduct the GBM simulation. After estimating the Poisson parameter λ, we are able to do the Poisson simulation within latent space. We see that the GBM model is capable of capturing many mass protest scenarios, to the order of magnitude. Even though it cannot simulate the propagation speed accurately at every time point, the method is effective at capturing the total number of infected nodes with an accuracy of $[0.78, 0.95]$, as shown in Fig. 14.



(a) Simulation without community

(b) Simulation with community

Figure 11: GBM and Poisson propagation simulation for Colombia protests on Dec 4, 2012.

- **How adept is the bispace model at capturing surge/burst moments? How reliable are the simulation results?**

Fig. 9 depicts the analsis of the YoSoy132 student movement, whose Twitter activity is generally tortuous, and the curve is full of surges and bursts. From Fig. 9 we see that the bispace model is capable of simulating the general surge trends. Comparing the bispace simulation results with ground truth, we can see at many time points, the bispace simulation matches the ground truth. Fig. 11 shows the second protest of people protesting against the government in Colombia; here the Twitter activity depicts a burst at a single time point which is hard to capture. We can see the bispace model did show there is a burst, but not at the precise time point, one of its current limitations.

- **Is the performance of the model better taking into account community structure?**

After numerous experiments, we plot the accuracy distribution of both approaches for all our mass protest situations in Fig. 14. Although the accuracies are sometimes interspersed, we can see that in overall the community model generally has a higher accuracy.

- **Can the bispace model simulate the propagation path?**

In addition to comparing the simulated counts of tweets over time with ground truth values, we can also compare the

Table 2: GBM simulation results for teacher protest events on Sep 2, 2013.

	Average degree	Diameter	Graph density	Connected components	Average clustering coefficient	Average path length
Simulation	1.791	11	0.002	183	0.083	4.786
Ground truth	1.726	18	0.002	204	0.008	6.261

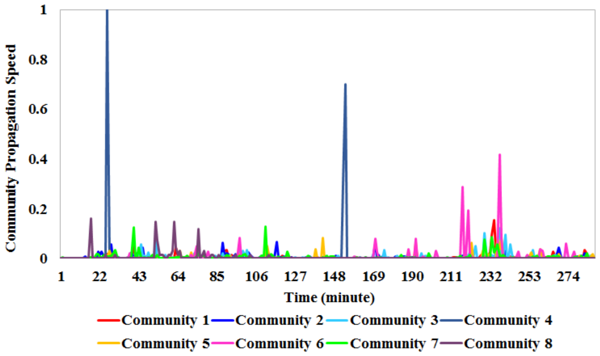


Figure 12: Normalized mass-protest propagation speed for major communities during teacher protest events.

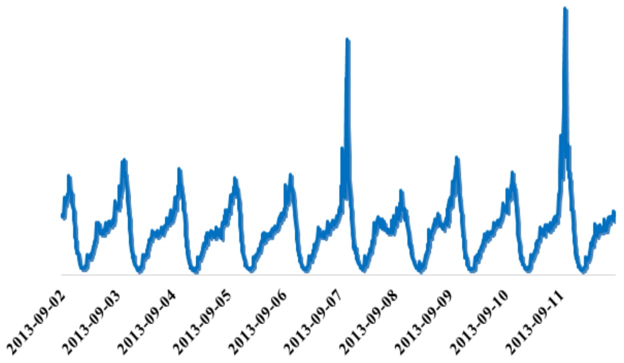


Figure 13: Total tweets over time from Sep 2 to Sep 11, 2013 (Mexico).

propagation path generated by the simulation against the actual propagation path through the mentions network. In Fig. 15 we can obtain a sense of the type of infection network bispace modeling creates as compared with the actual network. The simulation produces networks with relatively accurate paths and relevant characteristics as shown in Table 2. The component to which a user belongs is that of neighbors who can be reached from connected paths running along edges of the graph [15].

- **Between the geometric Brownian model and Poisson propagation approaches, which model is more dominant during the simulation process?**

From Figs. 9, 10, 11, by observing the blue dashed line (Poisson) and red dashed line (GBM), we can see that the Poisson process shows a mild activity, while the GBM model serves as the dominant component which can capture the moments of key surges.

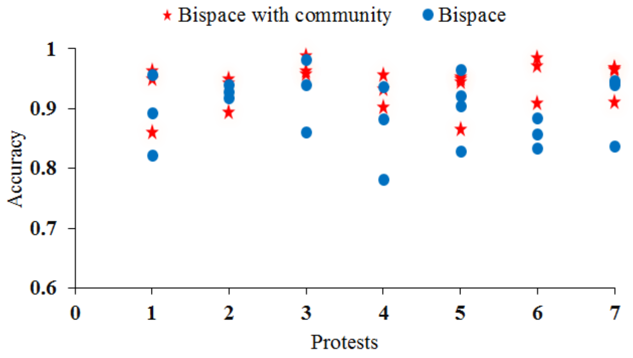
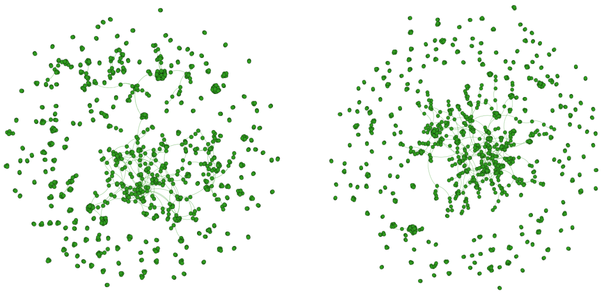


Figure 14: Performance accuracy of the bispace model for the 7 protest scenarios considered here, with and without community structure.



(a) Simulation results

(b) Truly infected nodes

Figure 15: Bispace model simulation results compared against ground truth infected nodes for the Mexican teacher protest events.

7. DISCUSSION

In this paper, we have characterized mass protest propagation using a bispace model comprising an observed mentions network space and a latent space. We have introduced a trust function to simulate propagation in observed space using a geometric Brownian motion diffusion process which can be further extended to support communities with different propagation parameters per community. We considered the latent space of all interactions outside the mentions network to be a Poisson distribution process. We have shown how the GBM diffusion model offers a new approach for modeling propagation through social networks like Twitter. Through our experiments, we find that the time required for spread of protest information through such networks is dependent on the network's substructures. Furthermore, we find that modeling the diffusion process on a community basis provides better results than the assumption that all nodes in the network spread information in the same way.

In future work, we hope to further characterize the hidden network with the goal of uncovering specific latent variables. Additionally, we envision applying the GBM model to other networks, such as the Twitter follower network, to identify those paths most susceptible to information dissemination. Finally, we desire to compare propagation of mass protest language against other themes, such as celebratory events, to aid in determining correlations between topic or sentiment and the resulting social media diffusion.

Acknowledgments

This work was partially supported by the NSF via grant IIS-1353346 and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoI/NBC, or the US Government.

8. REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[3] H. Castellini and L. Romanelli. On the propagation of social epidemics in social networks under SIR model. *eprint arXiv:nlin/0703053*, Mar. 2007.

[4] M. D. Conover, E. Ferrara, F. Menczer, and A. Flammini. The digital evolution of occupy wall street. *PloS one*, 8(5):e64679, 2013.

[5] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 105(41):15649–15653, 2008.

[6] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1, 2011.

[7] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(1):17–28, 2013.

[8] H. Zhou. Network landscape from a brownian particle's perspective. *Physical Review E*, 67(4):041908, 2003.

[9] F. A. Haight and F. A. Haight. *Handbook of the Poisson distribution*. Wiley New York, 1967.

[10] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis. Emotions as infectious diseases in a large social network: the sisa model. *Proc. Royal Society B: Biological Sciences*, 277(1701):3827–3835, 2010.

[11] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proc. KDD'13*, pages 266–274, 2013.

[12] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. KDD' 03*, pages 137–146. ACM, 2003.

[13] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proc. KDD'12*, pages 6–14. ACM, 2012.

[14] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. KDD'12*, pages 33–41, 2012.

[15] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[16] M. E. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[17] B. Øksendal. *Stochastic differential equations*. Springer, 2003.

[18] P. E. Oliver and D. J. Myers. Diffusion models of cycles of protest as a theory of social movements. *Presented at the Congress of the International Sociological Association*, 1998.

[19] M. G. Rodriguez, J. Leskovec, and B. Schölkopf. Modeling information propagation with survival theory. In *Proc. ICML'13*, pages 666–674, 2013.

[20] A. Saad-Filho. Mass protests under 'left neoliberalism': Brazil, june-july 2013. *Critical Sociology*, 39(5):657–669, 2013.

[21] J. Scott and P. J. Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[22] Y. Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proc. WSDM'12*, pages 733–742. ACM, 2012.

[23] P. Tankov. *Financial modelling with jump processes*. CRC Press, 2004.

[24] Z. Tufekci and C. Wilson. Social media and the decision to participate in political protest: Observations from Tahrir square. *Journal of Communication*, 62(2):363–379, 2012.

[25] U. F. Wiersema. *Brownian motion calculus*. Wiley. com, 2008.

[26] H. Zhou. Distance, dissimilarity index, and network community structure. *Physical Review E*, 67(6):061901, 2003.

[27] H. Zhou and R. Lipowsky. Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *Proc. ICCS'04*, pages 1062–1069. Springer, 2004.