PLOS ONE

RESEARCH ARTICLE

Forecasting Social Unrest Using Activity Cascades

Jose Cadena[1,2]*, Gizem Korkmaz[1], Chris J. Kuhlman[1], Achla Marathe[1,3], Naren Ramakrishnan[2], Anil Vullikanti[1,2]

1 Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA, **2** Department of Computer Science, Virginia Tech, Blacksburg, VA, USA, **3** Department of Agricultural and Applied Economics, Virginia Tech, Blacksburg, VA, USA

* jcadena@vbi.vt.edu

CrossMark
click for updates

OPEN ACCESS

Citation: Cadena J, Korkmaz G, Kuhlman CJ, Marathe A, Ramakrishnan N, Vullikanti A (2015) Forecasting Social Unrest Using Activity Cascades. PLoS ONE 10(6): e0128879. doi:10.1371/journal.pone.0128879

Academic Editor: Tobias Preis, University of Warwick, UNITED KINGDOM

Received: May 25, 2014

Accepted: May 4, 2015

Published: June 19, 2015

Copyright: © 2015 Cadena et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Relevant data are within the Supporting Information files.

Funding: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (Do(/NBC) contract number D12PC000337, National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM109718-01, Defense Threat Reduction Agency grant number HDTRA1-11-1-0016, Defense Threat Reduction Agency Comprehensive National Incident Management System Contract contract number HDTRA1-11-D-0016-0001, National Science

Abstract

Social unrest is endemic in many societies, and recent news has drawn attention to happenings in Latin America, the Middle East, and Eastern Europe. Civilian populations mobilize, sometimes spontaneously and sometimes in an organized manner, to raise awareness of key issues or to demand changes in governing or other organizational structures. It is of key interest to social scientists and policy makers to forecast civil unrest using indicators observed on media such as Twitter, news, and blogs. We present an event forecasting model using a notion of activity cascades in Twitter (proposed by Gonzalez-Bailon et al., 2011) to predict the occurrence of protests in three countries of Latin America: Brazil, Mexico, and Venezuela. The basic assumption is that the emergence of a suitably detected activity cascade is a precursor or a surrogate to a real protest event that will happen "on the ground." Our model supports the theoretical characterization of large cascades using spectral properties and uses properties of detected cascades to forecast events. Experimental results on many datasets, including the recent June 2013 protests in Brazil, demonstrate the effectiveness of our approach.

1 Introduction

Social media has become a window into happenings on the ground, from earthquakes [1] to specific news stories [2]. A key population-level event is civil unrest, i.e., protests, strikes, and occupy events, wherein civilian populations mobilize to raise awareness of key issues. As is evident from recent protests in many countries (e.g., Egypt, Turkey, Brazil), social media plays an important role in documenting, triggering, mobilizing, or even quelling such events, e.g., see [3, 4]. However, it is not clear when preliminary chatter observed on social media becomes a true precursor for a protest, and thus understanding the structure of the tweet behavior is a relevant problem.

While Twitter is used for many purposes (e.g., discontent expression, event reporting, planned protest recruitment), all of which can be used in a predictive model for civil unrest, we focus on modeling activity cascades (using a formulation of [5–7]) as a uniform precursor for

Foundation (NSF) under grant number CCF-1216000 and NSF grant number CNS-1011769. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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 IARPA, DoI/NBC, or the US Government. None of the funders or grants represent commercial funding sources, since they are all US federal agencies.

Competing Interests: The authors have declared that no competing interests exist.

civil unrest forecasting. Our goal is to design a model that forecasts the date of the event. Cascades, as is well known, help formalize the spread of influence and information, e.g., see [8–12]. Informally, cascades are subgraphs (often trees) that capture the spread of influence from a node to its newly activated/influenced neighbors and descendants.

There are many notions of cascades, which afford varying levels of formal characterization and utility. For instance, in Bakshy et al. [8], an edge $A \rightarrow B$ is included in a cascade only if it can be argued that some action (e.g., a posting or use of a URL) by B can be directly attributed to A, which makes the analysis intensive in terms of data and computation; also, the mathematical analysis under their model becomes challenging. Another common approach is to define cascades in terms of the (random) subgraph over which diffusion processes like linear threshold and independent cascades models spread, e.g. [12, 13]. Here, we use a simpler notion of cascades (referred to as "activity cascades"), introduced by [5–7]. Informally, such a cascade consists of a tweet emitted by a user u, and the tweets of the users who see/mention u's message (for instance, her direct followers or users who mention her) given that those tweets are sent within a small time interval (denoted by Δ), and so on (see Section 3.1 for a precise definition). This turns out to be a special case of Hawkes processes [14–16], which are based on mixtures of mutually exciting point processes. Although simpler to define, we demonstrate that this notion of cascades has good predictive power for modeling civil unrest, and is also amenable to rigorous analysis. Our key contributions are focused on the following three questions:

1. When do large activity cascades happen? A common empirical observation is that cascades seldom become very large. We rigorously prove necessary and sufficient conditions for large cascades in terms of spectral properties of the underlying graph (Section 3.2); these also imply a similar characterization for a class of Hawkes processes. We find that this characterization closely matches our empirical observations for synthetic traces. Specifically, our analyses show if the spectral radius of a cascade graph (defined in Section 3.2) is below a particular constant, then large cascades are not possible. Our techniques build on approaches for analyzing the spread of epidemics [17–19], and are the **first such results** for cascades of this kind.

2. Are there critical subsets of users that contribute to cascades? We study the questions of identifying critical subsets of users responsible for formation and survival of cascades, and formalize these as two complementary problems: CriticalSetFormation (CSFP) and CriticalSetShattering (CSSP). We show both to be NP-complete, and we evaluate different greedy heuristics to approximate them empirically by studying large, monthly cascades for all (country, month) combinations for Brazil, Mexico, and Venezuela over a 15-month period. Our results for CSSP show that a very small set of users are critical for a cascade to exist—their non-participation causes the cascade to shatter. We also prove that a high degree strategy gives a constant factor approximation for CSSP in random power law graphs. On the other hand, the results for CSFP suggest that unless a large fraction of users participate a cascade cannot exist. Thus, one needs a reasonable fraction of users, plus some critical users for a cascade to exist. These are validated through empirical observations in Section 4.2.

3. Can we forecast protests using activity cascades? Since large cascades are not very common, their occurrence signals a big event. We analyze over 353 million tweets from three Latin American countries over a 1.5-year period, and consider activity cascades formed by a filtered set of tweets. These tweets contain at least 3 keywords from a dictionary that has over 900 words in English, Spanish and Portuguese, related to civil unrest activities. This ensures that the resulting activity cascades will be relevant to the topic of civil unrest. Next, we build a feature set based on the structural properties of the cascades, to be used as predictors of social unrest. Statistical models are then used to remove redundancies among features and make predictions of events from the reduced feature set. The model is tested on multiple countries for robustness and compared against a baseline model. We show that our approach can 'beat

the news,' i.e., contribute a lead time of one to two days over the reporting of a protest in major news media, with an accuracy of over 0.75. It can even predict black swan events like the Brazilian Spring with an accuracy of 0.83. (Section 4.6).

Our paper helps explain the model and observations of [5–7] rigorously, especially conditions for occurrence of large cascades. Since their frequency is relatively low, their occurrence is a signal of significant events—-this corroborates with the observations of [5], and is the basis of our approach for forecasting protest events.

1.1 Related work

Analyzing traffic trends in twitter and other social media sites is a very active topic of research. Some of the specific applications include identifying specific news stories [2, 20], tracking natural disasters [1], predicting stock market moves [21, 22] and understanding political or cultural events [5, 6, 23, 24]. Yang et al. [25] predict temporal patterns in the usage of specific hashtags in social media data. Hutto et al. [26] show that increases in followers on Twitter are predicated on social behavior, message content, and social network structure variables in roughly equal proportions. Hsieh et. al. [27] demonstrated that experts could not match the crowd in identifying future interesting news stories. Most of these works have focused on counts of keywords and hashtags, and do not capture peer influence in the use of such terms. Peer influence is often modeled by diffusion processes, such as linear threshold and SI/SIS/SIR epidemic models, e.g., [12, 13]; in this context, cascades are used to refer to the (random) subgraph on which the diffusion spreads. There has also been a lot of work on using semantic information for attributing influence more carefully, e.g., [8–11], as discussed in Section 1.

A simpler notion of cascades is studied by [5, 6] in Twitter follower graphs and by [7] in the mentions graphs. Large cascades involving protest-related hashtags are found to occur infrequently. Our formulations extend point process models, which have been studied extensively. Two closely related approaches are by [15, 16]. Simma et al. [15] consider a model in which a Poisson process triggers other Poisson processes. They develop an EM algorithm to infer the random forest of events, which captures the cause of each event. Zhou et al. [16] use a multi-dimensional Hawkes process.

There are other works that utilize models for characterization and prediction. Linear regression models using average tweet rates, and tweet rate time series (per-day tweet rates over a 7-day period), have been used to predict box-office revenues from movies [28]. A classifier and hidden Markov model have been used with tweet content to establish the onset and end of identified events (versus event prediction) [29]. Natural language processing and LDA have been used to identify topics that capture collections of events identified in tweets; a linear regression model is then used to predict crimes [30].

With respect to forecasting social unrest, [31] provides empirical data showing that increases in food prices correlate with protests in 2008 and 2011. Rather than predicting specific unrest events, [32] uses a 2-parameter dynamics model to predict the distributions of numbers of unrest events per year, for many regions of the world. Disease outbreaks, deaths, and riots are forecasted with topic detection and tracking using news articles, and a Bayes scheme to compute the probability of some event, given other events occurring beforehand [33]. A tension parameter, based on hashtag usage, was shown to correlate well with clashes in Egypt between secularists and Islamists [34]. A generalized least squares model of political *violence* [35] is used to predict the overall level of violent activity in a country, by year. By contrast, we are interested in violent and non-violent protests.

Network characteristics and spectral bounds have been used for analyzing epidemic spread in networks. Ganesh et al. [17] develop necessary and sufficient conditions for the duration of

an SIS process; our analysis strongly builds on this approach, but our model requires the use of a variant of the node expansion, instead of edge expansion. Similar spectral radius bounds are also considered in [18, 19] for the SIS process.

Our work can be differentiated from the above studies in the following ways. This is the first work of which we are aware that predicts daily civil unrest events in multiple countries using a combination of different graph cascade characteristics. Further, we explain theoretically and demonstrate empirically conditions that delineate small and large cascade regimes, using spectral properties of the underlying graphs.

2 Materials

2.1 Twitter Dataset

For event prediction, we use a set of over 353 million tweets collected for Brazil, Mexico and Venezuela for the period of May 2012 through November 2013. Our dataset constitutes a 10% sample of the tweets for these countries during the above time period.

Our analysis is done separately for each country, and, as a pre-processing step, the tweets are filtered by country (using geolocation codes, place identifiers, language detection, author identification, and other enrichment processes), ignoring tweets for which a country of origin could not be determined. Next, we organized a vocabulary of 614 protest-related words (such as march, riot, strike, organize, democracia, conflicto, revolucion, criminalidade), 192 key-phrases (such as "right to work", "marcha por la paz"), and 105 country-specific key players (which include important public figures, political parties, labor unions), collectively referred to henceforth as keywords. Compiled by social scientists who are experts in the region, the keywords include English, Spanish, and Portuguese translations. We then subselect tweets which contain at least 3 keywords from our vocabulary. Tweet volumes before and after filtering are shown in Table 1.

2.2 Follower Data

We obtained the follower network for a subset of the users who appear as authors in our dataset, for each country by using Twitter API from a large number of machines. The size of the graphs for the three countries are the following: Mexico has 79,598 nodes and 1,437,687 edges, Venezuela has 312,241 nodes and 21,119,120 edges, and Brazil has 142,176 nodes and 6,854,368 edges.

2.3 Gold Standard Report (GSR) Datasets

GSR datasets are compiled by an independent group, selected by IARPA (Intelligence Advanced Research Projects Activity), which is comprised of social scientists and experts on Latin America. A small set of well-reputed newspapers for each country are used to identify the instances of civil unrest events to be included in the GSR. For each event, the GSR captures the

Table 1. Number of tweets for the period May 2012 to Nov 2013.

Country	Raw	Filtered
Mexico	97,873,616	3,524,695
Venezuela	105,938,438	6,683,834
Brazil	150,147,141	1,575,041

doi:10.1371/journal.pone.0128879.t001

when, where, who and why of the event, i.e. the date of the event, its geographic location, the population protesting (e.g. labor, medical workers, general population) and the reason for the protest (i.e., the event type, e.g., economic, political, resource). Here we are interested in forecasting primarily the if/when of the event (although text classification and geo-coding of the tweets reveals insight into where/who/why, something we do not study here further).

3 Methods

3.1 Activity Cascades

Let $G = (V, E)$ denote a directed graph, with $N_o(u)$ and $N_{in}(u)$ denoting the set of out-neighbors and in-neighbors for a node $u \in V$, respectively. The nodes represent Twitter users. We consider two kinds of graphs—a *follower* graph and a combined *mentions* and *retweet* graph. An edge (u, v) has different interpretations depending on the graph, as discussed below. We assume each user u sends at most one tweet at a time (with one second granularity), so a node-time pair (u, t) identifies a tweet. For a node $u \in V$, time t and time interval Δ, we define a cascade $C(u, t, \Delta)$ to be a set of tweets/node-time pairs in the following recursive manner, using the formulation of [5–7].

- If there is no tweet *driven by* node u at time t, then $C(u, t, \Delta) = \phi$.
- Else, $C(u, t, \Delta) = \{(u, t)\} \cup \{x \in C(v, t', \Delta) : v \in N_o(u), t' \in (t, t+\Delta]\}$

The term *driven by* will be explained below. This general notion of cascade makes no prior assumptions about the nature of the edges connecting the nodes of the graph, which gives us the flexibility to define the neighborhood of a node in different ways. Though this definition does not explicitly look for any correlation between the messages of u and v (which is used by, e.g., [8–12]), we use it on a set of tweets that are already filtered for protest related keywords, as done by [5–7], which brings in some correlation. Also, note that this notion of cascades is different from the more commonly studied notion associated with diffusion processes, e.g., [12, 13]—here, a cascade is a random subgraph on which the influence spreads.

In this paper, we study two types of activity cascades: *follower* (F) and combined *mention* plus *retweet* (MRT) cascades, defined, respectively, by the *follower* and *mention* and *retweet* graphs, each of which models a different kind of interaction between users in the Twitter network. Our methodology is the same as [36], except that we also include retweets.

In a follower graph, for every node $u \in G$, $N_o(u)$ is the set of Twitter followers of u, and $N_{in}(u)$ is the set of Twitter friends of u (i.e. users followed by u). From this definition, follower cascades in Twitter emerge in the following manner: a user u posts a tweet at time t starting a cascade where she is the only participant. For each follower v of u who posts a tweet at some time $t' \in (t, t+\Delta]$, (v, t') is added to this cascade, and so on, as illustrated in Fig 1. This process is repeated until no more users can be added to the cascade.

We combine mentions and retweets to form an MRT graph because both types of tweets indicate influence between pairs of users. Suppose a user w with name W composes a tweet at time t_1 that mentions another user u with name U, where W and U are sequences of characters of the form `[a-zA-z0-9_]+`. In the following, we specify the concatenation of two sequences of characters, $A = (a_1, a_2, \ldots, a_q)$ and $B = (b_1, b_2, \ldots, b_r)$, using the operator $\oplus$, as $A \oplus B = (a_1, a_2, \ldots, a_q, b_1, b_2, \ldots, b_r)$. Let P_W be the payload or content of the tweet of W; P_W is a sequence of characters. Because w mentions u, we have $(@) \oplus U \subseteq P_W$, which produces the directed edge (u, w) of the MRT graph. This edge has the semantics that u influences w. Analogously, if x with name X retweets a message from user w at time t_2, where $(RT\ @) \oplus W \subseteq P_X$, then we have the directed edge (w, x), where the semantics are the same as in the mentions edge: w

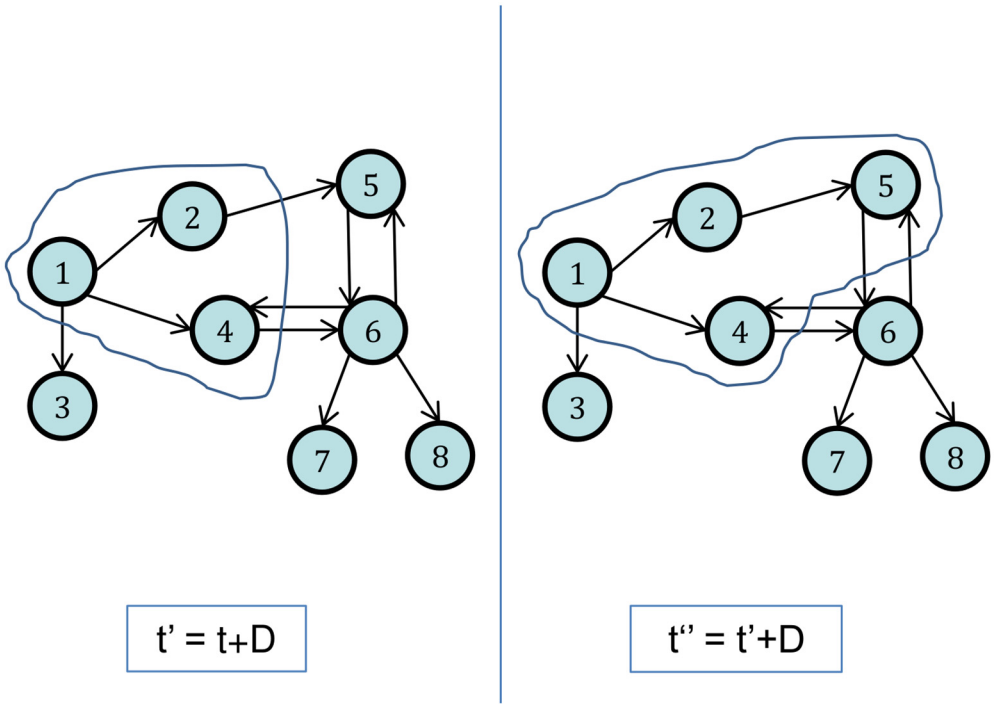

Fig 1. Formation of cascades in the Twitter follower network. At time t, node 1 posts a tweet. Nodes 2 and 4 post at times t_2 and t_4 between t and $t' = t + D$. Node 5, which follows 2, posts at some time t_5 between t' and $t'' = t' + D$. Therefore, the cascade $C(1, t, D)$ is $C(1, t, D) = \{(1, t), (2, t_2), (4, t_4), (5, t_5)\}$.

doi:10.1371/journal.pone.0128879.g001

influences x. If the two tweets occur such that $t_2 \in (t_1, t_1+\Delta]$, then the two edges link up to form a directed path of length 3, $u \rightarrow w \rightarrow x$, and the cascade $C(u, t_1, \Delta) = \{(u, t_1), (w, t_2)\}$.

We have several notes. The term *driven by* indicates the user that instigates the cascade. For a follower graph, the instigator is the user that sends the first tweet of a cascade. For an MRT graph, the instigator is the first influencer of a cascade. Second, users (nodes) in an MRT graph with zero out-degree (x in this example) are not included in the MRT cascade because there is no evidence that these nodes influence other users. Also, a single tweet can produce multiple edges in an MRT graph. Since retweets of an original tweet preserve the original tweeter, no matter how many times the original tweet is (sequentially) retweeted, a set of these retweets (without mentions) produces a star subgraph of the MRT graph. Finally, MRT cascades, unlike follower cascades, directly use tweet payloads; however, F cascades utilize the follower graph.

We provide additional definitions that will be useful in forecasting social unrest. We say that a (follower or mentions/retweet) cascade C is *active* on day d if there exists at least one message or tweet (u, t) by user u at time t, such that t is some time during day d and (u, t) is an element of C. The *size* of a cascade is the number of tweets comprising it. A *user* or *participant* is a tweeter.

3.2 Characterizing large cascades in terms of graph properties

Our empirical results suggest that large and long cascades are rare, and arise within communities of users. We now attempt to explain this behavior by relating it to the spectral properties of the graph, by considering a formulation based on a slight relaxation of the notion of cascades: We consider a cascade starting at a random initial node u_0 at time t_0; (i) $X(0)$ denotes the initial

configuration. We say that (u_0, t_0) is active in the cascade at time t_0. Following our definition, we will think of a cascade as consisting of tweets indexed by user-time pairs (u, t); (ii) The number of tweets sent by each user u is a Poisson process with parameter α_u; (iii) If user u sends a message at time t, and some other user v with $u \in N_o(v)$ is active at time t, then (u, t) becomes active in the cascade at time t; (iv) A tweet (u, t) ceases to be active after a (random) time duration $D(u, t)$ drawn from an exponential distribution with parameter $\delta = 1/\Delta$; and (v) The cascade $C(u_0, t_0)$ dies when there are no more active tweets in it.

We now model this as a Markov process $\mathbf{X(t)}$ with values in $\mathbb{N}^V$. Let $X_u(t)$ denote the number of messages by user u that are active at time t. Then, the cascade evolves in the following manner:

$$X_u: \text{ increases by } 1 \quad \text{at rate } \alpha(u) \text{ if } \{v \in N_{in}(u) : X_v > 0\} \neq \phi \tag{1}$$

$$X_u: \text{ decreases by } 1 \qquad \text{at rate } \delta X_u \tag{2}$$

Every cascade eventually becomes inactive, since $X_u = 0$ for all u is the unique absorbing state for this Markov process. The *lifetime* of the cascade, the duration for which it lasts, is precisely $T = \sup\{t : X_u(t) > 0, \text{for some } u \in V\}$. We now derive necessary and sufficient conditions for obtaining large cascades.

3.2.1 Multivariate Hawkes Processes. Our formulation above makes it a special case of the multivariate Hawkes processes, as we now discuss. A Hawkes process $\mathbf{N_t}$ is a type of self-exciting counting process characterized by a time-dependent intensity (rate) $\lambda(t)$ [14–16].

Let $\mathbf{N_d(t)}$ be a multidimensional counting process, where $d \in \{1, \ldots, D\}$ denotes a dimension (with D being the number of dimensions). Let $\lambda_d(t)$ denote the intensity of $N_d(t)$. The process is defined in the following manner:

$$\lambda_d(t) = \mu_d(t) + \sum_{d'=1}^{D} \int_{-\infty}^{t} \kappa_{d'd}(t-s) dN_{d'}(s),$$

where μ_d is a base intensity for dimension d, and $\kappa_{d'd}(\tau)$ is a kernel function describing the influence of the previous events in dimension d' on the current rate on d.

For our formulation, let each node $u \in V$ be a separate dimension, and let $\mathbf{N_u(t)}$ be the number of messages contributed to an ongoing cascade. We have $\mu_u(t) = 0$ and the kernel function as $\kappa_{vu}(\tau) = \alpha_{vu} \times \kappa(\tau)$, where: (i) $\alpha_{vu} = \alpha_u N_v(t)$ if $v \in N_{in}(u)$; otherwise, $\alpha_{vu} = 0$. Here, α_u is the (fixed) tweeting rate of u, and α_{vu} describes the fact that u's contributions to the cascade are proportional to her in-neighbors' contributions (i.e. her friends in the follower graph); and (ii) $\kappa(\tau) = 1$ if $0 < \tau \leq \Delta$; otherwise $\kappa(\tau) = 0$. As a result, we have:

$$\lambda_u(t) = \alpha(u) \sum_{v \in N_{in}(u)} (N_v(t) - N_v(t-\Delta))$$

We use the process $\mathbf{X(t)}$ below for our discussion, since it simplifies the analysis; our results hold for a class of Hawkes processes with the kind of kernel function mentioned above.

3.2.2 Conditions for Small Cascades. We now derive conditions when the maximum cascade size is $O(\log n)$, with high probability, where n is the number of nodes. The process $\mathbf{X(t)}$ is non-linear, making it quite complex to analyze; instead we consider the following relaxation $\mathbf{Y(t)}$:

$$Y_u: \text{ increases by } 1 \quad \text{at rate } \alpha(u) \textstyle\sum_{v \in N_{in}(u)} Y_v \tag{3}$$

$$Y_u: \text{ decreases by } 1 \qquad \text{at rate } \delta Y_u \tag{4}$$

Lemma 1 The process **Y(t)** stochastically dominates **X(t)** so that $X(t) \leq Y(t)$ *for all* $t \geq 0$.

Proof 1 Our proof is based on designing a coupling that ensures that $X(t) \leq Y(t)$ for all $t \geq 0$, and builds on [17]. Clearly, $X(0) \leq Y(0)$. We consider the process **Y(t)** and for each node u, we sample random variables R_u^1 and R_u^2 from exponential distributions with parameters $\alpha(u)\Sigma_{v\in N_{in}(u)}\, Y_v$ and δY_u, respectively. The first transition out of $Y(0)$ happens at time τ, which equals $\min_u\{R_u^1, R_u^2\}$. Our coupling will specify the transition for the process **X(t)** in the following manner. Suppose the transition at time τ corresponds to $Y_u(\tau) = Y_u(0) + 1$; this would have happened with rate $\alpha(u)\Sigma_{v\in N_{in}(u)}\, Y_v(0)$. For the corresponding process **X(t)**, the transition $X_u(\tau) = X_u(0)+1$ is made with probability $\frac{1}{\sum_{v\in N_{in}(u)} Y_v(0)}$, if $\{v\in N_{in}(u){:}X_v > 0\} \neq \phi$; otherwise $X_u(\tau) = X_u(0)$. This ensures that the transition $X_u(\tau) = X_u(0)+1$ happens with the correct rate. Similarly, the transition corresponding to $Y_u(\tau) = Y_u(0)-1$ can be handled to get a coupling of the first jumps in **X(t)** and **Y(t)**.

Lemma 2 Let $\rho(A)$ denote the spectral radius of A, the adjacency matrix of G. Assume that G is a bi-directed graph and let $\alpha_{max} = \max_u \alpha(u)$. If $\alpha_{max}\,\rho(A) < \delta$, the duration of the cascade T satisfies $\Pr[T > t] \leq ne^{-(\delta-\alpha_{max}\,\rho(A))t}$ and $\mathrm{E}[T] \leq \frac{\log n+1}{\delta-\alpha_{max}\rho(A)}$.

Proof 2 Our proof is an adaptation of that of [17] for the SIS model; we describe it here completely for completeness. From equations (3, eqn:y2), it follows that

$$\begin{aligned} E[Y_u(t+dt) - Y_u(t)|\mathbf{Y}(\mathbf{t})] &= \alpha(u)\sum_{v\in N_{in}(u)} Y_v(t)dt - \delta Y_u(t)dt + o(dt) \\ &\leq \alpha_{max}\sum_{v\in N_{in}(u)} Y_v(t)dt - \delta Y_u(t)dt + o(dt), \end{aligned}$$

which implies $\frac{dE[\mathbf{Y}(t)]}{dt} \leq (\alpha_{max}A - \delta I)E[\mathbf{Y}(t)]$.

This has solution $E[\mathbf{Y}(t)] \leq e^{(\alpha_{max}\,A-\delta I)}\,\mathbf{Y}(0)$. From Lemma 1, and since $\mathbf{X}(0) = \mathbf{Y}(0)$, we have $E[\mathbf{X}(t)] \leq e^{(\alpha_{max}\,A-\delta I)}\,\mathbf{X}(0)$.

Let $N_t = \Sigma_v X_v(t) = \mathbf{1}^T\,\mathbf{X}(t)$ denote the number of nodes infected at time t. Then, $N_t \leq \mathbf{1}^T\, e^{(\alpha_{max}\,A-\delta I)}\,\mathbf{X}(0)$. Since A is a symmetric matrix, $e^{(\alpha_{max}\,A-\delta I)}$ is also symmetric, and we have $\parallel e^{(\alpha_{max}A-\delta I)}\mathbf{X}(0) \parallel \leq \rho(e^{(\alpha_{max}A-\delta I)}) \parallel \mathbf{X}(0) \parallel = e^{\alpha_{max}\rho(A)-\delta}\sqrt{n}$. This implies $E[N_t] \leq ne^{\alpha_{max}\,\rho(A)-\delta} = ne^{-(\delta-\alpha_{max}\,\rho(A))}$, since $\parallel \mathbf{X}(0) \parallel \leq \sqrt{n}$. The first part of the lemma follows since $\Pr[T > t] = \Pr[N_t \geq 1] \leq E[N_t]$.

For the second part of the lemma, we have

$$\begin{aligned} E[T] &= \int_0^\infty \Pr[T > t]dt \\ &\leq \int_0^\infty \min\{1, ne^{\alpha_{max}\rho(A)-\delta}\} \\ &\leq \int_0^{\log_{n/(\delta-\alpha\rho(A))}} 1dt + \int_{\log_{n/(\delta-\alpha_{max}\rho(A))}}^\infty ne^{-(\delta-\alpha_{max}\rho(A))}dt \\ &\leq \frac{\log n + 1}{\delta - \alpha_{max}\rho(A)} \end{aligned}$$

Lemma 2 implies that when $\alpha_{max}\,\rho(A) < \delta$, any cascade has size $O(\log n)$. We are able to prove Lemma 2 only when G is symmetric, because the proof relies on all eigenvalues being real, though the statement might be true in general.

3.2.3 Conditions for Large Cascades. We now consider the conditions for having a large cascade (of size c^m, where c is a constant larger than 1, and m is a parameter). We need the following version of the isoperimetric constant, which captures node expansion.

$$\hat{\eta}(G,m) = \min_{S \subseteq V, |S| \leq m} \frac{\sum_{v \in V-S: N_{in}(v) \cap S \neq \phi} \alpha_v}{|S|}.$$

We sometimes omit the reference to the graph G in $\hat{\eta}(G,m)$, and just use $\hat{\eta}(m)$ when G is clear from the context. We now consider a Markov process $\mathbf{Z}(t)$ with state space $\{0, \ldots, m\}$, defined in the following manner:

$$\begin{aligned} Z(t) &= Z(t)+1 \text{ at rate } \hat{\eta}(m)Z, \text{ if } Z < m \\ Z(t) &= Z(t)-1 \text{ at rate } \delta Z, \text{ if } Z > 0 \end{aligned}$$

Lemma 3 $Z(t)$ is stochastically dominated by $\Sigma_u X_u(t)$, i.e., $Z(t) \leq \Sigma_u X_u(t)$ for all $t \geq 0$.

Proof 3 The proof is also by designing a coupling, as in Lemma 1. We assume that $Z(0) \leq \Sigma_u X_u(0)$, and prove the statement by induction. We consider the process $\mathbf{X}(t)$ and for each node u, we sample random variables R_u^1 and R_u^2 from exponential distributions with parameters $\alpha(u)1_{\{v \in N_{in}(u): Xv > 0\} \neq \phi}$ and $\delta X_u(0)$, respectively. The first transition out of $X(0)$ happens at time τ, which equals $\min_u\{R_u^1, R_u^2\}$. Our coupling will specify the transition for the process $\mathbf{Z(t)}$ in the following manner. Let $S = \{w: X_w(0) > 0\}$. Let $N^+(S) = \{v \in V-S: N(v) \cap S \neq \phi\}$.

Suppose the transition at time τ corresponds to a transition $X_u(\tau) = X_u(0)+1$ for some node u (which increases the number of active messages). The total rate at which such an increase happens equals $\Sigma_u \alpha(u)1_{\{v \in N_{in}(u): Xv > 0\} \neq \phi} = \Sigma_{u \in N^+(S)} \alpha(u)$. First, suppose that $|S| < m$. The transition $Z = Z+1$ is now made at time τ with probability

$$\frac{\hat{\eta}(m)Z}{\sum_{u \in N^+(S)} \alpha(u)}.$$

This fraction is in [0, 1], because $Z(0)\hat{\eta}(m) \leq \sum_{u \in N^+(S)} \alpha(u)$, by definition of $\hat{\eta}(m)$, and because $|S| < m$, so that the transition happens with the correct rate. Second, if $|S| \geq m$, Z is unchanged, which is the correct rate.

Next, we consider the case that the transition at time τ corresponds to a transition $X_u(\tau) = X_u(0)-1 = 0$ for some node u. In this case, the transition $Z(\tau) = Z(0)-1$ is made with probability $\frac{Z(0)}{\sum_u X_u(0)}$, which is well defined since this is in [0, 1]. Also, note that there is some probability that $\Sigma_u X_u(0)$ decreases by 1, but $Z(0)$ does not— this does not violate the property, because in this case $Z(0) < \Sigma_u X_u(0)$.

Therefore, in either case, we have $Z(\tau) \leq \Sigma_u X_u(\tau)$, and the lemma follows.

Lemma 4 Suppose $r = \frac{\delta}{\hat{\eta}(m)} < 1$. Then, we have $\Pr[T > \frac{r^{-m+1}}{2m}] \geq \frac{1-r}{e}(1 + O(r^m))$.

Proof 4 The proof of the above lemma follows by observing that the process $\mathbf{Z}(t)$ is a one-dimensional random walk, defined in the following manner. Consider the discrete time Markov chain associated with Z. Let $p(i, j)$ denote the probability that Z switches to value j from i.

Then, we have:

$$\begin{aligned} p(i,i+1) &= \frac{\hat{\eta}(m)}{\hat{\eta}(m)+\delta},\ i=1,\cdots,m-1, \\ p(i,i-1) &= \frac{\delta}{\hat{\eta}(m)+\delta},\ i=1,\cdots,m-1 \\ p(0,0) &= 1, \\ p(m,m-1) &= 1. \end{aligned}$$

Then, the duration of the cascade, T, is the time before the process hits 0. As in [17], this is the standard gambler's ruin probability, and the rest of the proof follows exactly as in [17].

Spectral connection. Vertex expansion is related to the graph spectrum. If G is a d-regular graph, and if its spectral gap, i.e., the difference between the smallest and second smallest eigenvalue, is μ, the vertex expansion for sets of size at most m is $\frac{1}{(1-m/n)\mu^2+m/n}$.

3.2.4 Identifying Critical Sets in a Cascade. We now consider the following questions: What is the critical subset of users whose tweets are responsible for the cascade to survive? What is the critical subset of users whose removal would cause the cascade to disintegrate? These are related and complementary problems, which can help explain the conditions for cascade formation. We consider a slightly more general notion of cascades than the one defined in Section 3.1—for a set S of nodes, we define $C(S, t, \Delta) = \cup_{u\in S} C(u, t, \Delta)$ to be the union of cascades starting at nodes in S. As a result, any directed acyclic graph can be seen as a cascade formed by its sources.

CriticalSetShattering Problem CSSP(G, $\mathcal{C}$, k):

Input: A set of cascades $\mathcal{C}$ in a graph $G = (V, E)$ and parameter k.

Goal: Determine the smallest set $S \subseteq V$ of users, such that the sub-cascades of all $C\in\mathcal{C}$ in $G[V\backslash S]$ are of size at most k.

Thus, the goal in CSSP is to find the subset S whose removal causes all cascades in $\mathcal{C}$ to be "shattered".

CriticalSetFormation Problem, CSFP(G, $\mathcal{C}$, $\boldsymbol{\alpha}$, k):

Input: A set of cascades $\mathcal{C}$ in a graph $G = (V, E)$, tweet rate $\boldsymbol{\alpha}$ and parameter k.

Goal: Determine the smallest set $S \subseteq V$ of users, such that for every $C\in\mathcal{C}$, a sub-cascade of size at least k exists in the graph $G[S]$ with tweet rate $\boldsymbol{\alpha}$.

Thus, CSFP quantifies the number of users needed to cause large cascades. While CSSP and CSFP are closely related and seem to be complementary problems, they are quite different from a computational perspective.

Complexity and algorithms for CSSP. We have the following result.

Lemma 5 CSSP (C, G, k) is NP-complete.

Proof 5 It is easy to verify that CSSP is in NP. The NP-hardness of CSSP is by a reduction from the balanced graph partitioning problem (see, e.g., [37])— this problem involves finding the smallest subset $S \subset V'$ of nodes in an undirected graph $H = (V', E')$ so that all components in $H[V'-S]$ have size at most b, which is a given parameter.

Let C be a DAG formed by orienting the edges of H arbitrarily, so that it forms a DAG. Let $B \subset V'$ whose removal splits C into weakly-connected components $H_1, \ldots, H_r$, each of size at most $k = b$; as discussed earlier, each component H_i is a cascade formed by the sources in that DAG. If we ignore the directions of the edges, we get components of size at most $k = b$. This implies that the solution to CSSP in C corresponds to a solution to the separator problem on H. The converse also holds similarly.

Whenever the condition in Lemma 2 is tight, i.e., it gives both necessary and sufficient conditions, CSSP can be solved by simply attempting to reduce the spectral radius $\rho(A)$. We consider the special case of the Chung-Lu random graph model [38]: given a weight sequence $\mathbf{w} = (w(v_1), w(v_2), \cdots, w(v_n))$ for nodes $v_i \in V$, the random graph $G \in G(\mathbf{w})$ is obtained by choosing each edge (u, v) with probability $\frac{w(u)w(v)}{\sum_{v_i \in V} w(v_i)}$. We use the following result from [39].

Lemma 6 *[39] If* $G = G(\boldsymbol{w})$ *is a random graph in the Chung-Lu model with the weight sequence being a power law with exponent* $\beta > 2$, *removal of the* $\Theta(n/T^{2(\beta-1)})$ *nodes with the highest weight ensures that the spectral radius of the residual graph is at most T, almost surely*.

Motivated by Lemma 6, we study heuristics for CSSP based on degree and the core number in the underlying graph. Since $\rho(A) \geq \sqrt{\max_v deg(v, G)}$, a natural heuristic for CSSP is to reduce the maximum degree $\max_v deg(v, G)$. Also, since $\rho(A) \geq \frac{2|E(H)|}{|V(H)|}$ for any subgraph H of G, another natural heuristic for CSSP is to reduce the density of every subgraph H. Motivated by these bounds on the spectral radius of a graph, we consider the following heuristics for CSSP: (*i*) *high degree heuristic*: remove nodes in decreasing order of degree in G; and (*ii*) *high core number heuristic*: remove nodes in decreasing order of their core-number in G.

Complexity and algorithms for CSFP. We have the following hardness result.

Lemma 7 CSFP $(C, G, \boldsymbol{\alpha}, k)$ is NP-complete.

Proof 7 It is easy to verify that CSFP is in NP. We only discuss the NP-hardness. Our proof is by a reduction from the Set Cover problem, an instance of which consists of a set B of elements, a set A of subsets of B; the goal is to select the smallest subset $A' \subset A$ such that each element in B is covered by a set in A'.

We construct an instance of CFP in the following manner. We set ϵ to be a large integer. We construct a graph $G = (\{r, r'\} \cup A \cup B, E)$, where E consists of the following edges: edges (j, i) if $j \in B$, $i \in A$ and j is contained in set i, edges (r, i), (r', i) for all $i \in A$, and edge (r, r'). We have $\alpha_r = \alpha_{r'} = \epsilon$, while $\alpha_u = 1/n$ for all $u \in A \cup B$. We note that $\eta(G, \hat{1}, \{u\}) \geq \epsilon$ for all $u \in \{r, r'\} \cup A$.

Suppose $A' \subseteq A$ is a minimum set cover. Then, increasing $\alpha_u = \epsilon$ for all $u \in A'$ will ensure that $\eta(G, \hat{1}, \{j\}) \geq \epsilon$ for all $j \in B$. Similarly, suppose S is the optimum solution to the CFP problem. Clearly, $S \subseteq A$; if $S \cap \{r, r'\} \neq \phi$, we can drop r, r' from S without affecting the feasibility of the solution. For each $j \in B$, there must be at least one neighbor in S; else, we cannot have $\eta(G, \hat{1}, \{j\}) \geq \epsilon$. This implies S is a set cover.

This completes the reduction.

We consider a greedy algorithm for CSFP: pick nodes in non-increasing order of degree until the cascade on the graph induced by these nodes has size at least k.

We note that the maximum cascade size can be estimated for a given rate assignment within a factor of $1 \pm \epsilon$, with high probability, in time $O(|E| \log n/\epsilon^2)$ by a standard Chebyshev bound.

3.3 Forecasting Social Unrest using Cascades

Social media is believed to be responsible for facilitating critical communication often required to fuel momentum preceding the events of civil unrest. Here we explore the ability of Twitter data to act as a predictive signal of future civil unrest. Specifically, we study the prediction of civil unrest events (e.g., protests, strikes) by using properties of the activity cascades in Twitter data.

We employ a regression model to predict the probability of a civil unrest event in a given day by using features based on the structural properties of the activity cascades described earlier. We hypothesize that, in general, unusually large and long cascades are likely to be indicative

of future events of interest. These could be sport events, concerts, revolutions, elections, etc. However, *given that our tweets are filtered by civil unrest related keywords, we expect the events detected will be of civil unrest type.*

Starting from May 1, 2012 to November 30, 2013, each day, we compute the total number and size of cascades, number of participants and duration (in days) of cascades, change in the number of participants and tweets, average growth rate of tweets and average growth rate of participants. These features are collected daily for each active follower cascade and MRT cascade. For each of the features described above, we also compute the minimum, maximum, median, and average of the cascade size, duration, and users, as well as the average value of the 1st, 2nd, 3rd, and 4th quartile of their distribution and add them to the feature set. Therefore, our initial feature set consists of 114 attributes: (7 cascade properties × 8 aggregate statistics × 2 types of cascades + (daily cascade count × 2 types of cascades)).

Our initial list of features is expected to be highly correlated, resulting in unstable estimates if used in regression modeling. For example, the cascade size of an MRT cascade is almost equivalent to the number of users in that cascade, since people tend to retweet a message only once. In addition, the majority of these cascades last for one day (duration = 1), which makes the average growth equivalent to the change in the cascade size on the last day, which in turn gives the change in the number of users, and so on.

In order to address the problem of multi-colinearity, we compute the correlation between every pair of features (i.e. the correlation matrix) and remove highly correlated features. Specifically, if the correlation between two variables is greater than 0.7, we only keep one of the two. This methodology reduced the size of our feature set from 114 to 13. Next, we use a regression methodology called LASSO (Least Absolute Shrinkage and Selection Operator) to further remove the redundant features and to shrink coefficients [40]. The LASSO based logistic regression model uses cascade features from both the follower and the MRT models.

3.3.1 Baseline model. We build a baseline model in order to set a benchmark and to measure the added predictability provided by the Twitter data. The baseline model uses no external input in the regression model. It uses an autoregressive logistic model of order 1, since we have a binary dependent variable. It uses lagged values of itself as the predictor. Formally, we estimate $Y_t = \alpha + \beta Y_{t-1} + \varepsilon$ where Y_t is the binary variable: 1, if there is an event on day t in the GSR; 0 otherwise. The fitted values of Y_t, which give the likelihood of future events, are compared against the actual events in the GSR for measuring the model's performance.

3.3.2 Evaluation metrics. Once the probabilities are estimated for the test days, a threshold t is used to determine whether or not the probability exceeds t for an event to occur. The optimal threshold t^* is determined by cross-validation, especially maximizing the area under the ROC (receiver operating characteristic) curve. Once learned, t^* is further used to separate events from non-events given the estimated probabilities. We evaluate our models against two settings—a lead time of 1 day and a lead time of 2 days—and compare the results against the GSR using standard measures such as precision, recall, and the misclassification rate.

3.3.3 Operational issues. All the Twitter data used in this study is in compliance with the Terms of Use and all website conditions of Twitter. We use data from May 2012 to infer the activity cascades. The GSR data is available only from Nov 2012. Hence, data from Nov 2012 to May 2013 is used for training the forecasting model and the held-out June 2013 data is used for testing in case of Brazilian Spring. This training/test split is actually a tough test for the forecasting algorithm because June 2013 depicted very significant changes in rates of occurrences of protests in Brazil (more on this below), and our approach was nevertheless able to forecast this variation in its forecasts.

4 Results and Discussion

Our experimental results are focused on answering the following questions.

1. Do our theoretical conditions for large cascades from Section 3.2 hold true in real datasets? (Section 4.1)
2. What level of increase in user tweeting initiates large cascades? Does the solution to the CFP problem using our greedy heuristic suggest that a few important users are sufficient or is large number of users necessary? (Section 4.2)
3. How adept are activity cascades at detecting precursors and surrogates for protests? (Section 4.3)
4. Which cascade features yield the best forecasting performance? Are these features consistently better across multiple countries? (Section 4.4)
5. Are cascade models for forecasting protests significantly better than baseline models? Is there value in building features from different Twitter networks (i.e. follower and MRT)? (Section 4.5)
6. Can our methods help forecast 'black swan' events like the Brazilian Spring? (Section 4.6)

4.1 Validation: Two Regimes for Cascade Sizes

We empirically verify the conditions for large cascades uncovered using our theoretical analysis. We find ten of the largest follower cascades in Mexico between June 27 and Sep 7, 2012, and the subgraphs induced by these users (also referred to as the cascade graphs). We consider synthetic twitter traffic generated using a Poisson process (as in Section 3.2) for the users in these cascade graphs with rate $\alpha_u = \alpha$, and then compute the cascades induced by this for $\Delta = 4$ hours. Fig 2 shows the maximum cascade size n_α, as a function of α for each of these cascade graphs. The y-axis is normalized by the number of nodes n in the respective cascade. For most of the cascades, we observe a clear phase transition for α somewhere in the range [0.05,0.15]. For these particular graphs, we find that $\rho(\hat{A})$ (in the notation of Lemma 2) is below δ when α is in the range [0.10,0.15]. We also observe in Fig 2 that the cascades die out when $\alpha \leq 0.05$, which is consistent with the condition in Lemma 2. Note that some of the cascades die out even for higher values of α, which is consistent with the gap between the necessary and sufficient conditions in Lemmas 2 and 4.

4.2 Identifying Critical Sets in Cascades: CSSP and CSFP

Empirical analysis of heuristics for CSSP. We start with collections of tweets from Brazil, Mexico, and Venezuela that form cascades, in monthly intervals, from May 2012 through July 2013. We use reciprocal follower graphs (i.e., two users must follow each other to form an edge in the reciprocal follower graph, which implies a stronger association between users [5]) to determine which users follow each other. We use $\Delta = 4$ hours for the maximum duration that may separate a user's and a follower's tweets in forming edges in the cascade graph. The reciprocal follower graphs for Brazil, Mexico, and Venezuela have 1.9, 0.5, and 4.9 million edges, respectively, and 123409, 69226, and 253423 nodes.

We select nodes (users) from the cascade graphs based on node properties in the *follower graph*. Specifically, we successively remove nodes (*i*) from greatest degree to least, and (*ii*) from the greatest k-shell to the least, from the follower graphs. We then remove these nodes from cascade graphs and compute the numbers of nodes and the sizes of the largest weakly

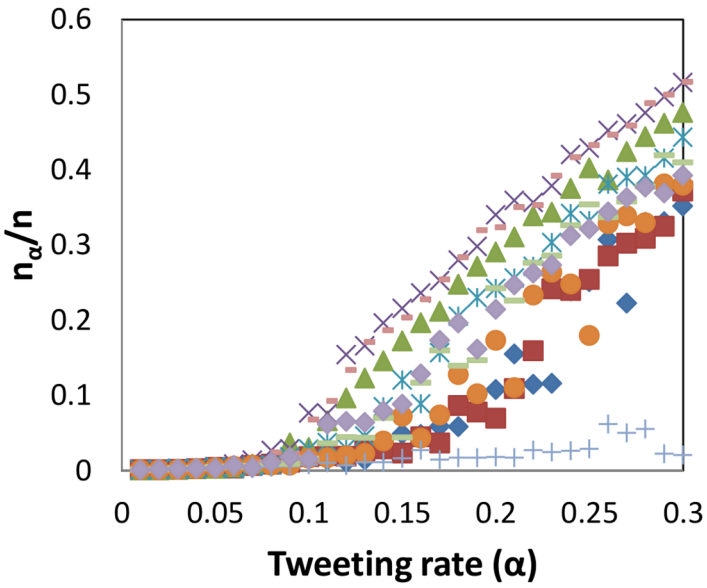

Fig 2. Follower cascade size as a function of tweeting rate for ten follower cascades in Mexico (produced between June 27, 2012 and September 7, 2012), with synthetic traffic. As the tweet rate of the users increases, we observe a sudden transition from a regime of very low user participation to a higher-activity regime.

doi:10.1371/journal.pone.0128879.g002

connected components that remain in them (cf. Section 3.2.4). Recall that nodes in a cascade graph are (user,time) pairs. Results are provided in Figs 3 and 4 for the largest cascades of Brazil and Venezuela, respectively. Results for other cascades, across countries and months, show the same behavior.

Removing relatively small fractions of high degree nodes and high k-shell nodes are both effective in reducing the sizes of cascades. For all (country, month) combinations, the high degree heuristic is more effective than the high k-shell heuristic. Differences between the methods can be significant, particularly for small numbers of removed nodes.

Empirical analysis of heuristics for CSFP. We now solve the CSFP problem for selected large cascades in Mexico, Brazil, and Venezuela using the greedy heuristic in Section 3.2.4, in order to approximate the change in the level of tweeting that caused the cascade. We examine the differences in aggregate level of tweets and user participation, as well the characteristics of cascades that might result at lower levels of participation. When we consider the largest cascades and retain either the tweets or the users involved with probability p, the resulting sub-cascade size varies quite gradually with p, instead of showing a clear phase transition (in contrast with the results in Section 4.1). It is possible that the more gradual change is due to the non-uniform rates α_u for users u in the large cascades, which cause a higher level of weighted vertex expansion, even for moderate values of p. These results are omitted because of space constraints.

We now consider the effect of a greedy choice of users from the original cascades, using variants of the greedy heuristic described for CSFP. The first (structural) heuristic selects k nodes $\{v_1, \ldots, v_k\}$ with the greatest values $\mid N'_o(v) \mid$, where $\mid N'_o(v) \mid$ is the number of out-neighbors of

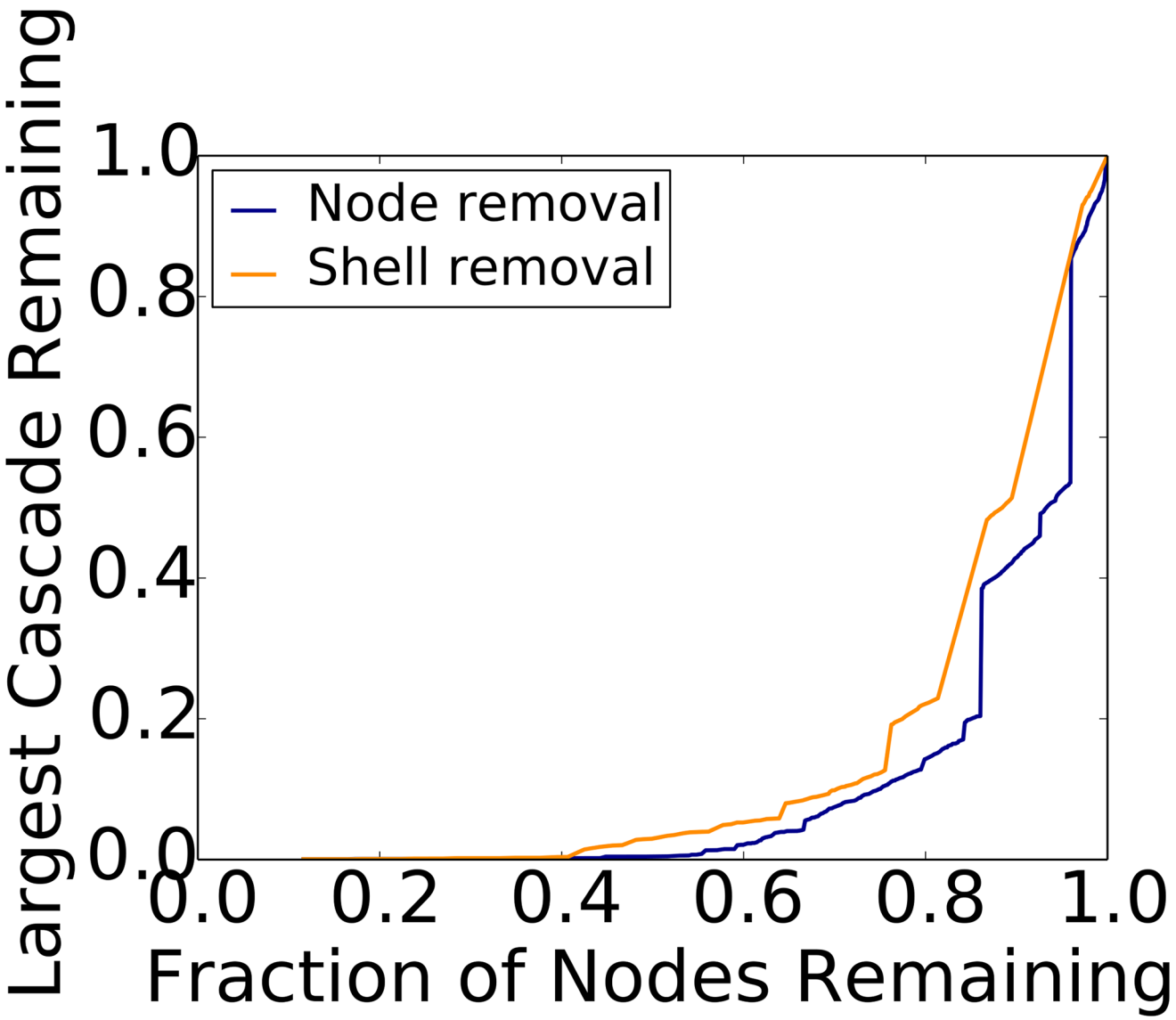

Fig 3. Node and shell removal heuristics for CSSP (Brazil). Here, we see the largest remaining sub-cascade size in terms of numbers of tweets (normalized by the original size) as a function of numbers of remaining nodes in the cascade graph (normalized by the original number of nodes). This cascade occurred in June 2013, and its original size is 15,791 tweets.

doi:10.1371/journal.pone.0128879.g003

v appearing in the maximum cascade for a (country, Δ) pair; this is a high-degree heuristic. The second (dynamical) heuristic simply chooses the k nodes with the greatest frequency of occurrence in a cascade. In both heuristics, $k = pN_c$, where p is the probability of selecting a node (cf. previous subsection) and N_c is the number of nodes in an original cascade, making it consistent with the earlier analysis. We compare these with a random selection of users with probability p. Fig 5 shows the (normalized) maximum size of a cascade for each of the above heuristics (labeled "degree", "frequency" and "random", respectively), averaged over 50 trials. Fig 6 shows the corresponding normalized maximum number of unique users in the cascades.

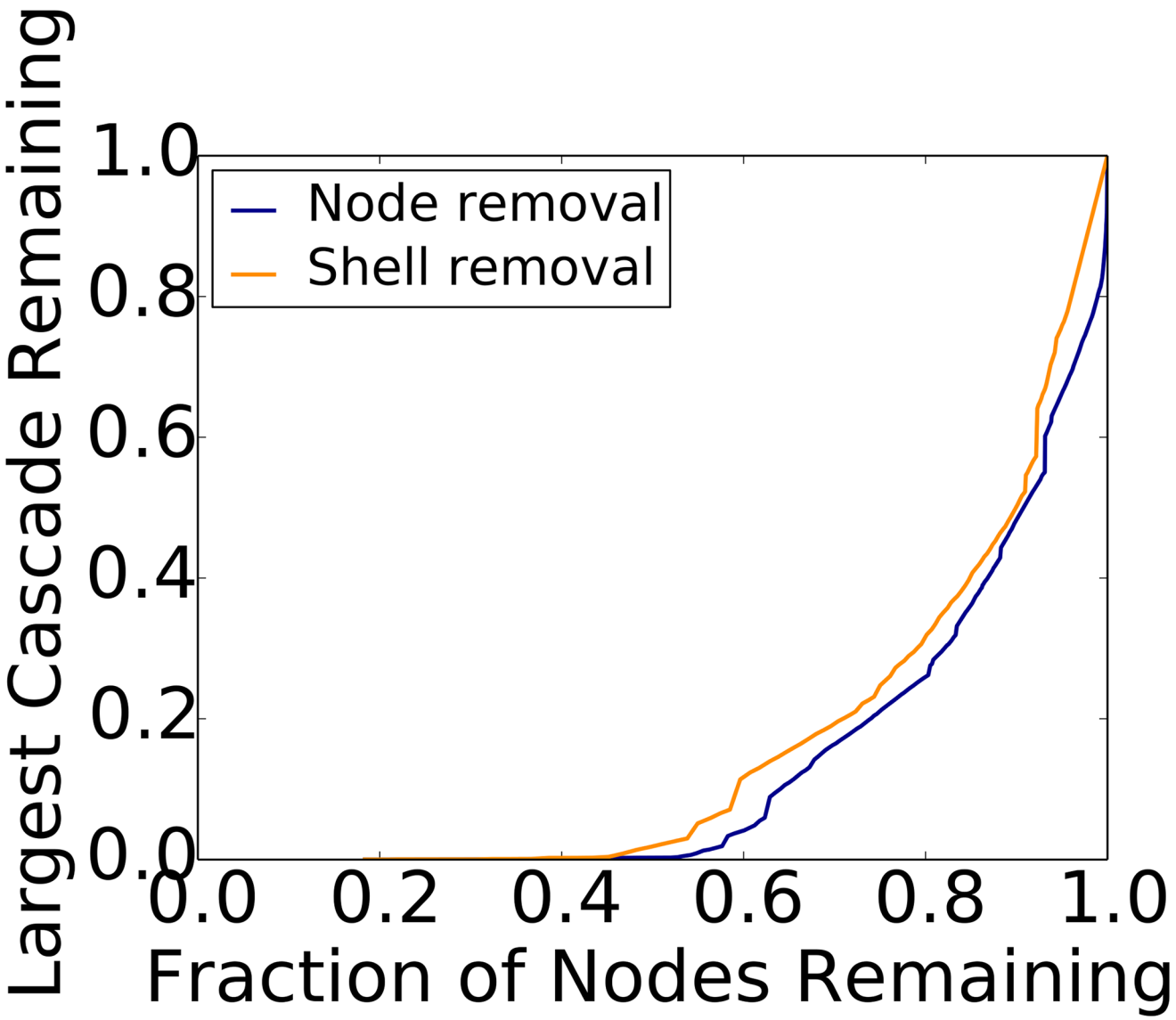

Fig 4. Node and shell removal heuristics for CSSP (Venezuela). Here, we see the largest remaining sub-cascade size in terms of numbers of tweets (normalized by the original size) as a function of numbers of remaining nodes in the cascade graph (normalized by the original number of nodes). This cascade occurred in April 2013, and its original size is 226,179 tweets.

doi:10.1371/journal.pone.0128879.g004

The normalization constant in each plot is the empirically determined maximum cascade size and maximum number of users, respectively. We find that the high degree heuristic generally produces the largest cascades in terms of tweets and users. For ordinate values in the range 0.2 to 0.4, the maximum sizes for the high degree heuristic are 2× to 10× those of the random heuristic. These data indicate that large cascades are tenuous; e.g., even with the high degree heuristic, 80% of the original users are required to produce a cascade that is 80% of the maximum measured size. Thus, it is not the case that a few users drive cascade formation. However, for CSSP, removal of a smaller fraction ($\sim$ 10–20%) of users can significantly reduce cascade size.

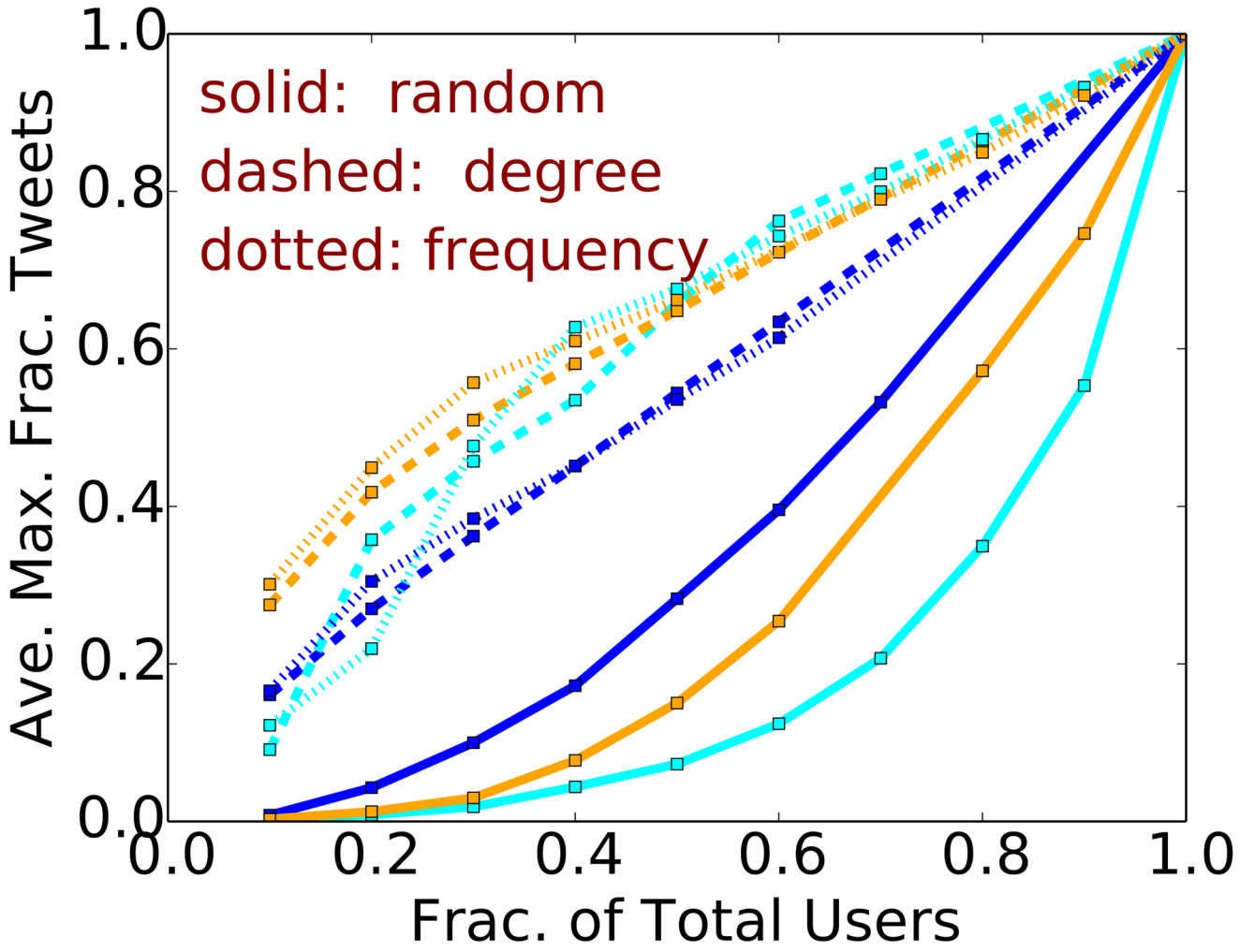

Fig 5. Greedy heuristic for CSFP. The (normalized) maximum cascade size vs. the fraction of users selected for some of the largest cascades in different countries. Data are: blue (Mexico, Δ = 1 hour); light blue (Brazil, Δ = 4 hours); and orange (Venezuela, Δ = 4 hours).

doi:10.1371/journal.pone.0128879.g005

4.3 Illustrative Results of Tweet Contents

What types of tweets form large activity cascades that are predictive of protests? There are at least two broad classes of such tweets that we highlight here. The first kind pertains to tweets as an early reporting mechanism that then go on to form activity cascades that can serve as a protest recruitment or mobilization staging ground. The second kind are tweets that explicitly call for protest action by individuals.

As an example of the first kind, we discuss two tweets with a high retweet count found in our MRT cascades for Brazil. The original tweets were sent on January 27, and they are about a past event (night club fire) which led to the deaths of 231 people in Santa Maria. These tweets eventually formed part of a cascade that corresponded to an actual demonstration that took place on January 28, 2013. According to the news articles, 35,000 people marched and held a

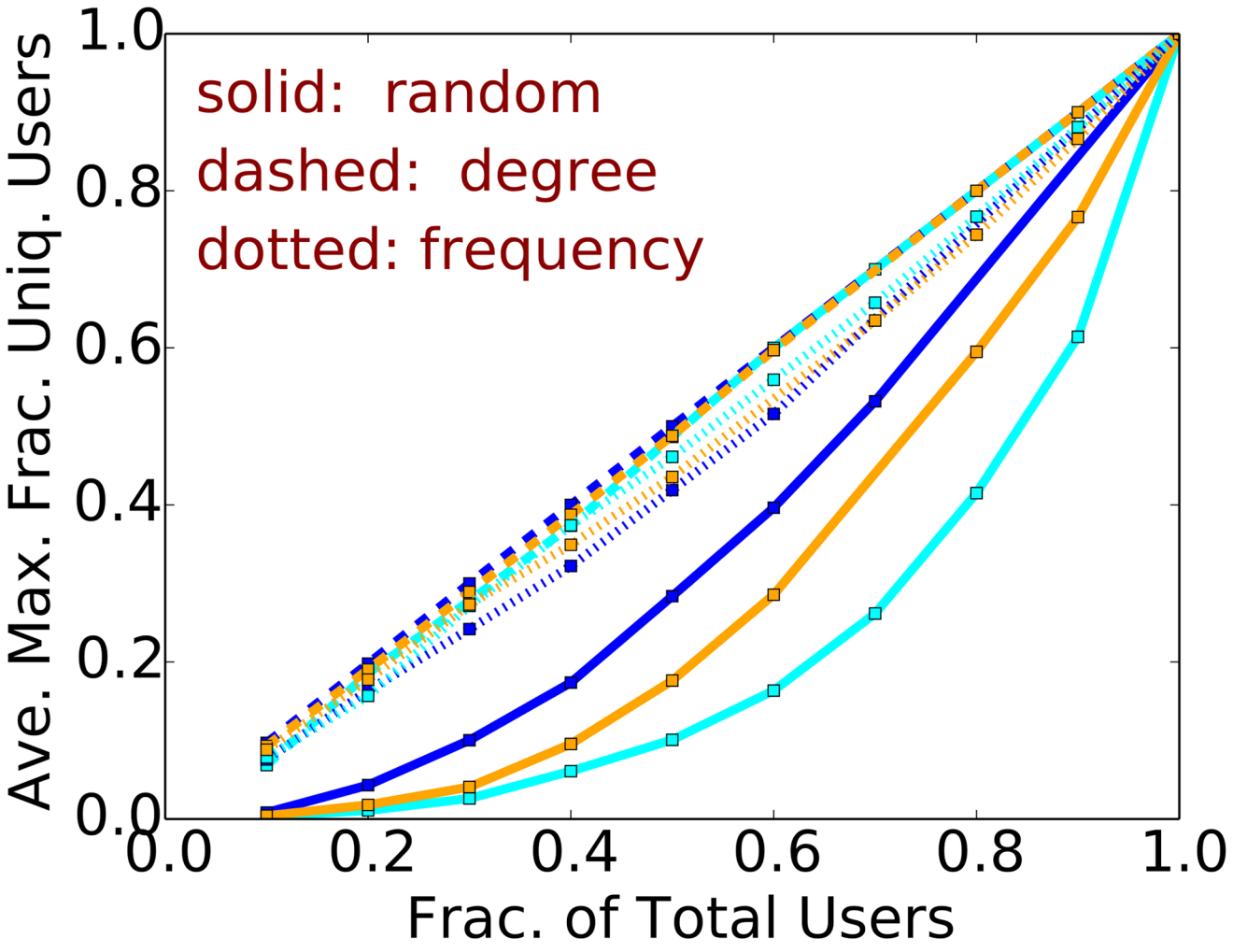

Fig 6. Greedy heuristic for CSFP. The (normalized) maximum number of unique users vs. the fraction of users selected for some of the largest cascades in different countries. Data are: blue (Mexico, Δ = 1 hour); light blue (Brazil, Δ = 4 hours); and orange (Venezuela, Δ = 4 hours).

doi:10.1371/journal.pone.0128879.g006

moment of silence in front of the gymnasium where the victims' bodies had been identified. This shows that tweets selected by our vocabulary and tracked for activity cascade formation may indeed correspond to actual protest events on the ground. The second kind is highlighted by a tweet calling for a protest on September 7, illustrating that further analysis of tweets originating from such cascades can aid in forecasting.

4.4 Cascade Feature Utility for Forecasting

Fig 7 illustrates descriptive statistics of selected features of mention and follower graph cascades in Brazil. Fig 8 shows the variables selected by the LASSO based logistic regression model. The LASSO based model finds that the probability of an event depends upon the duration and the slope of the follower and MRT graphs. These selected features are used as

Retweet-Mention Cascades	mean	stdev	median	min	max	skew	kurtosis	s.err
RT_Cascade.Duration.Median	1.02	0.13	1	1	2	6.95	48.49	0.01
RT_Cascade.Duration.Max	1.97	0.16	2	1	2	-5.74	30.97	0.01
RT_User.Slope.All.Average	5.05	1.46	4.79	1.9	16.13	2.39	10.59	0.06
RT_User.Slope.All.Max	31.53	64.97	22	4	1450	18.25	391.39	2.7
RT_User.Slope.All.Median	3.93	0.66	4	1	7	-0.73	6.67	0.03
RT_Users.In.Cascade.Average	5.8	1.55	5.53	1.9	17.53	2.03	8.64	0.06
RT_Users.In.Cascade.Max	36.36	68.81	25	4	1450	15.71	308.8	2.86
RT_Users.In.Cascade.Median	4.29	0.65	4	1	7	0.01	6.58	0.03
Total.Cascades	53.63	78.23	30	8	1077	7.44	75.49	3.26
Follower Graph Cascades								
F_Cascade.Duration.Max	2.93	2.33	2	1	18	3.44	13.38	0.1
F_Cascade.Duration.Median	1.6	1.29	1	1	10	3.39	13.64	0.05
F_User.Slope.Last.Day.Average	70.4	131.08	9.96	1.77	1361.1	4.12	25.75	5.46
F_User.Slope.Last.Day.Max	174.69	267.86	51	4	2662	3.51	19.43	11.15
F_User.Slope.Last.Day.Median	69.18	150.59	4	1	1736	4.75	34.47	6.27
F_Users.In.Cascade.Average	240.14	650.57	15.84	1.95	4798.1	4.9	26.77	27.08
F_Users.In.Cascade.Max	577.97	1289.6	78	4	9120	4.37	22.02	53.69
Total.Cascades	1106.2	2517.2	239	22	20651	5.1	30.21	104.8

Fig 7. Descriptive statistics of selected features (Brazil) for the MRT and F models. The names in the first column consist of the name of the structural feature (i.e., cascade size, duration or slope, which is the incremental increase in the size per day), and the statistical operations (i.e. median, average etc.).

doi:10.1371/journal.pone.0128879.g007

Lasso Variables and Coefficients	
(Intercept)	-0.4622
MRT_Cascade.Duration.Max	0.0143
MRT_User.Slope.All.Average	0.1253
F_User.Slope.Last.Day.Max	0.0009
F_User.Slope.Last.Day.Median	0.0008

Fig 8. LASSO Variables. Variables selected by LASSO in the cascade model for Brazil, for a training period of November 2012 through May 2013.

doi:10.1371/journal.pone.0128879.g008

		#Event	Threshold	Match	TPR	FPR	ACC.	Brier	ROC Area
Brazil	Baseline	4	0.25	15	0.25	0.18	0.71	0.27	0.54
	Model - Tweet Volume	4	0.3	10	0.75	0.59	0.48	0.2	0.44
	Model - Cascade	4	0.33	16	0.5	0.17	0.76	0.24	0.74
Mexico	Baseline	19	0.71	17	0.89	1	0.81	0.13	0.45
	Model - Tweet Volume	19	0.61	15	0.68	0	0.71	0.1	0.79
	Model - Cascade	19	0.49	20	1	0.5	0.95	0.1	0.92
Venezuela	Baseline	6	0.3	10	0.16	0.40	0.48	0.3	0.383
	Model - Tweet Volume	6	0.38	15	0.66	0.27	0.71	0.26	0.75
	Model - Cascade	6	0.37	16	0.83	0.26	0.76	0.26	0.81

Threshold: The probability above which an event is predicted to occur

Match: TP+TN

TPR,FPR: True positive rate, false positive rate

Accuracy (ACC): Match/#Days

Fig 9. Performance of the predictive models. We show the performance of the three models in terms of accuracy, brier score, and area under the ROC curve. The cascades model has the best performance accross different countries.

doi:10.1371/journal.pone.0128879.g009

explanatory variables in a generalized linear regression model [41] which confirms their significance and relevance.

4.5 Comparison Against Baseline Models

Fig 9 compares the performance of the baseline model, volume-based model and the cascade model for the three countries. For each model, we report the threshold used, true positive rate (TPR), false positive rate (FPR), accuracy (ACC), brier score, and the area under the ROC curve. The results in Fig 9 report the threshold that results in the highest accuracy in prediction.

Note that the cascade model outperforms both the baseline model and the volume-based model. Figs 10 and 11 shows the ROC for these models. Each point in the line represents a different threshold for the model.

Model Robustness Across Countries: Fig 9 shows the performance of the cascade model for Brazil, Venezuela and Mexico. For Mexico, there are 20 matches out of 21 prediction days which results in 95% accuracy. On the other hand, the cascade model results in 76% accuracy for Venezuela and Brazil. Fig 12 illustrates the ROC plots for each of the countries at various thresholds confirming Brazil and Venezuela's performance to be worse than Mexico.

4.6 Forecasting the Unexpected: The Brazilian Spring

In a recent wave of uprisings in Brazil, known as the Brazilian spring, demonstrations were organized to protest increases in bus, train, and metro ticket prices in some Brazilian cities, which quickly grew to become Brazil's largest unrest since 1992. These events involved the

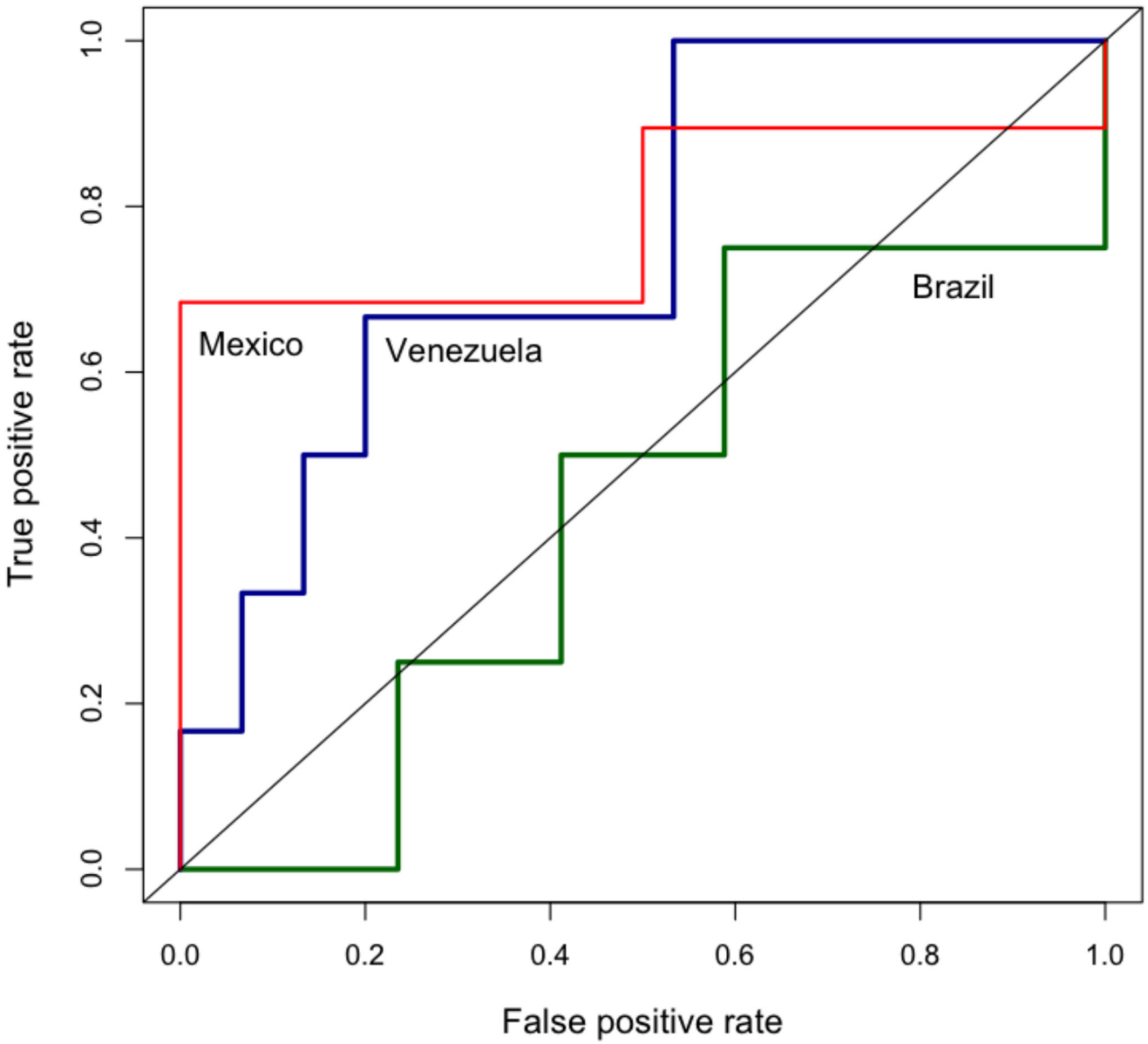

Fig 10. ROC curves for the volume-based model. We show the ROC curves for Mexico, Brazil, and Venezuela. Training period November 1, 2012 to November 9, 2013; test period November 10, 2013 to November 30, 2013.

doi:10.1371/journal.pone.0128879.g010

"General Population." We test the performance of our cascade-based prediction model by making a retrospective forecast for the events occurred in the month of June 2013 in Brazil. In the training period (November 01, 2012 to May 30, 2013), there were 131 days (out of 212) with events that involved the general population. In the test period of June 2013, there were events almost every day (29 days out of 30). The total number of events was more than 29, since there were multiple events on some days.

For this experiment, we collected 83 million tweets between November 2012 and June 2013 from Brazil. The keyword-based filtering (select if a tweet has at least 3 keywords present) resulted in 890,000 tweets which were further used to generate the graphs and the cascades.

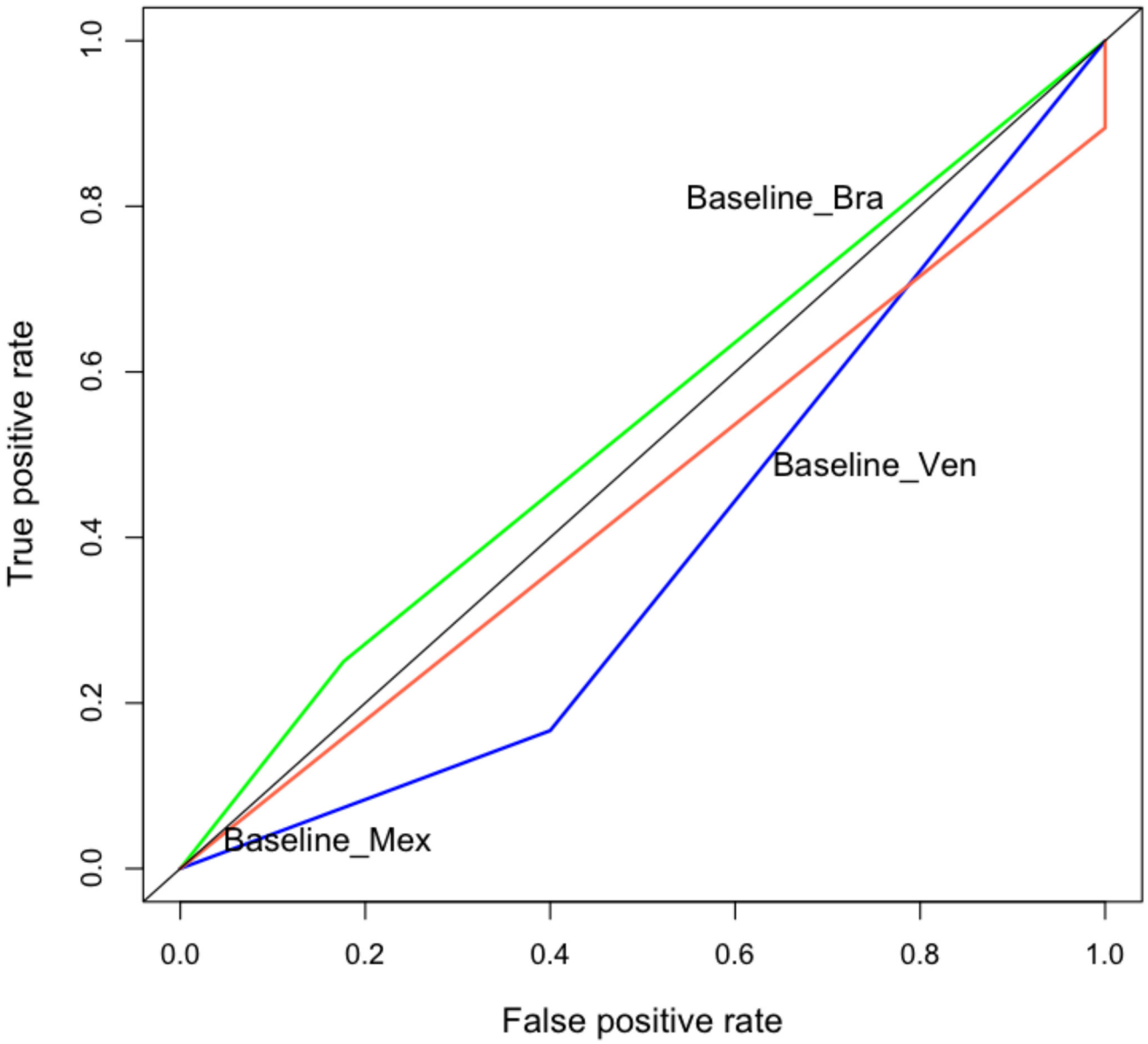

Fig 11. ROC curves for the baseline model. We show the ROC curves for Mexico, Brazil, and Venezuela. Training period November 1, 2012 to November 9, 2013; test period November 10, 2013 to November 30, 2013.

doi:10.1371/journal.pone.0128879.g011

The graph-based features were extracted for each of the cascade-based models. Fig 13 displays the performance of the cascade model for Brazil in June 2013. The model results in an area of 0.86, showing good performance. However, ROC does well when the number of events is very high. Therefore, we also plot the probabilities obtained from the regression model for the test period. Note that the peaks correspond to the days when the events become nationwide and violent. Fig 14 highlights the sudden surge in the structural features of the cascades. The cascade model results in 25 matches out of 26 alerts (when the best threshold is chosen as 0.6), a performance accuracy of 0.83 and TPR of 0.86.

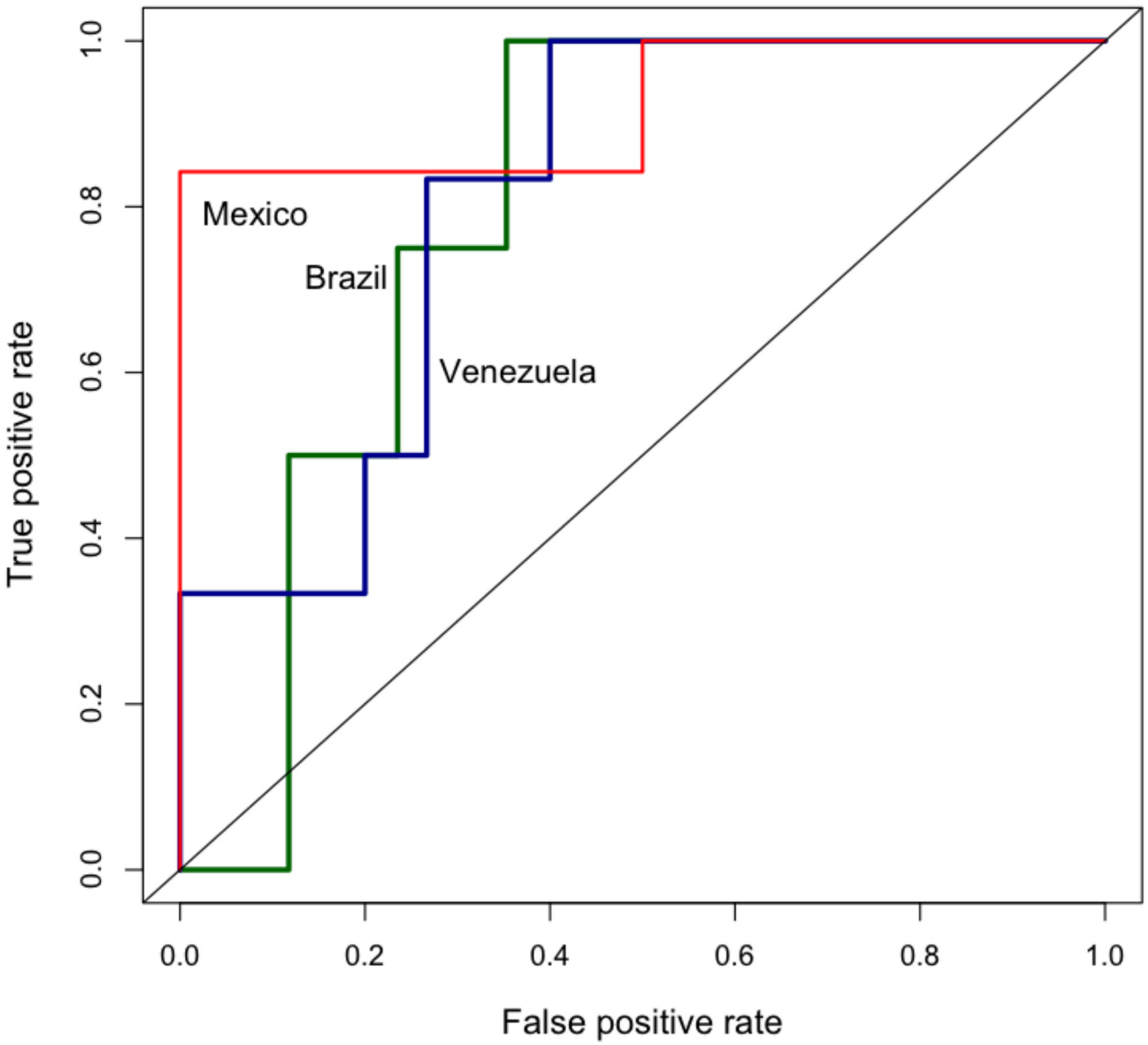

Fig 12. ROC curves for different countries. ROC curves for Mexico, Brazil and Venezuela for the cascade model. Training period November 1, 2012 to November 9, 2013; test period November 10, 2013 to November 30, 2013.

doi:10.1371/journal.pone.0128879.g012

Conclusions

Our main contributions include: (i) a detailed analysis of activity cascades arising from protest related tweets, (ii) use of cascade features for a predictive model for protest events, (iii) a rigorous formulation to explain the regimes for small and large cascades, in terms of the spectral radius and the node expansion, and (iv) characterizing critical sets for cascades, by means of the CSSP and CSFP formulations.

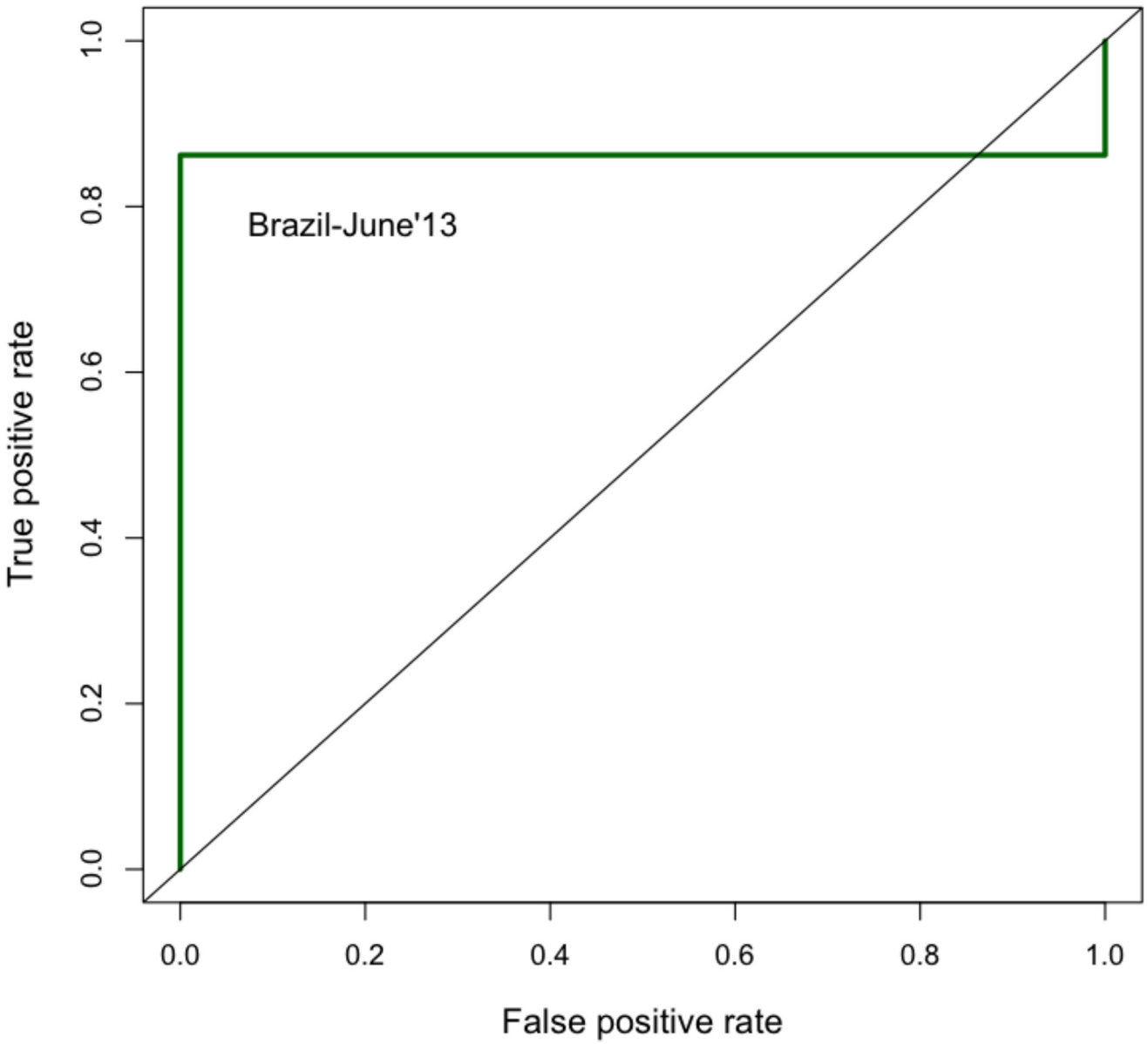

Fig 13. ROC curve for Brazil. ROC curve for different models for Brazil, for a training period of Nov 2012 through May 2013 and testing period of June 1-30, 2013.

doi:10.1371/journal.pone.0128879.g013

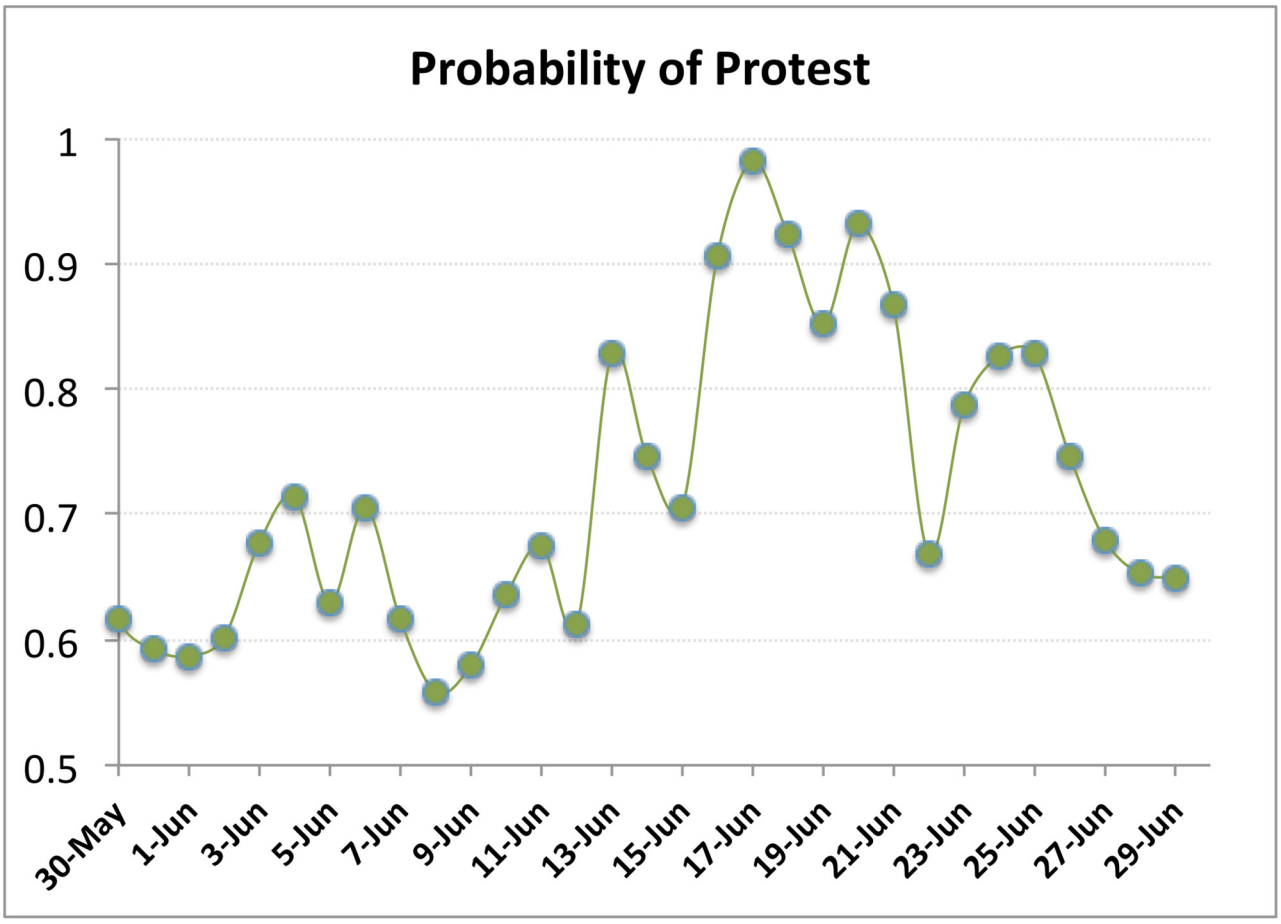

Fig 14. Cascade properties as predictors of protest. Cascade size, number of users, and number of cascades for Follower and MRT cascades in Brazil for the period November 2012—June 2013.

doi:10.1371/journal.pone.0128879.g014

Our results suggest that, despite their simplified notion, activity cascades are useful in characterizing and predicting civil unrest events. Our rigorous characterization of the conditions for having large cascades highlights the role of the overall network structure; this corroborates with other recent work on influence cascades [8].

Supporting Information

S1 Dataset. Follower cascade features.
(XLS)

S2 Dataset. MRT cascade features.
(XLS)

S3 Dataset. Keyword counts for the volume-based model.
(XLS)

Author Contributions

Conceived and designed the experiments: AM NR AV. Performed the experiments: JC GK CK. Analyzed the data: JC GK. Contributed reagents/materials/analysis tools: AM NR AV. Wrote the paper: JC GK CK AM NR AV.

References

1. Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: WWW; 2010.
2. Sankaranarayanan J, Samet H, Teitler BE, Lieberman MD, Sperling J. TwitterStand: News in Tweets. In: ACM GIS; 2009.
3. Huang C. Facebook and Twitter key to Arab Spring uprisings: report. In: The National; 2011.
4. Milne S. Egypt, Brazil, Turkey: without politics, protest is at the mercy of the elites. In: The Guardian; 2013.
5. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. The Dynamics of Protest Recruitment through an Online Network. Scientific Reports. 2011;1: . doi: 10.1038/srep00197 PMID: 22355712
6. Borge-Holthoefer J, Rivero A, Moreno Y. Locating privileged spreaders on an online social network. Physical Review E. 2012;. doi: 10.1103/PhysRevE.85.066123
7. Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W. Outtweeting the Twitterers -Predicting Information Cascades in Microblogs. In: WOSN; 2010.
8. Bakshy E, Hofman J, Mason W, Watts D. Everyones an Influencer: Quantifying Influence on Twitter. In: WSDM; 2011.
9. Kwak H, Lee C, Park H, Moon S. What is Twitter, a Social Network or a News Media? In: WWW; 2010.
10. Sun E, Rosenn I, Marlow C, Lento T. Modeling Contagion through Facebook News Feed. In: ICWSM; 2009.
11. Adar E, Adamic L. Tracking information epidemics in blogspace. In: IEEE/WIC/ACM International Conference on Web Intelligence; 2005.
12. Leskovec J, Adamic L, Huberman B. The dynamics of viral marketing. ACM Trans Web. 2007;. doi: 10.1145/1232722.1232727
13. Kempe D, Kleinberg JM, Tardos É. Influential nodes in a diffusion model for social networks. In: ICALP 2005; 2005.
14. Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. PNAS. 2008;. doi: 10.1073/pnas.0803685105 PMID: 18824681
15. Simma A, Jordan MI. Modeling Events with Cascades of Poisson Processes. In: UAI; 2010.
16. Zhou K, Song L, Zha H. Learning Social Infectivity in Sparse Low-rank Networks Using Multidimensional Hawkes Processes. In: AISTATS; 2013.
17. Ganesh A, Massoulie L, Towsley D. The effect of network topology on the spread of epidemics. Proceedings of INFOCOM. 2005;.
18. Tong H, Prakash BA, Eliassi-Rad T, Faloutsos M, Faloutsos C. Gelling, and Melting, Large Graphs by Edge Manipulation. In: CIKM; 2012.
19. Prakash BA, Chakrabarti D, Faloutsos M, Valler N, Faloutsos C. Threshold conditions for arbitrary cascade models on arbitrary networks. In: ICDM; 2011.
20. Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter. In: ICWSM; 2011.
21. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. Quantifying Wikipedia usage patterns before stock market moves. Scientific reports. 2013;3: .
22. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Scientific reports. 2013;3: .
23. Morales AJ, Losada JC, Benito RM. Users structure and behavior on an online social network during a political protest. Physica A. 2012;p. 5244–5253. doi: 10.1016/j.physa.2012.05.015
24. Tremayne M. Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street. Social Movement Studies: Journal of Social, Cultural and Political Protest. 2013;p. 110–126.
25. Yang J, Leskovec J. Patterns of Temporal Variation in Online Media. In: WSDM; 2011.
26. Hutto CJ, Yardi S, Gilbert E. A Longitudinal Study of Follow Predictors on Twitter. In: CHI; 2010.

27. Hsieh CC, Moghbel C, Fang J, Cho J. Expert vs The Crowd: Examining Popular News Prediction Performance on Twitter. In: WWW. ACM; 2013.
28. Asur S, Huberman BA. Predicting the Future With Social Media. In: WI-IAT; 2010.
29. Iyengar A, Finin T, Joshi A. Content-based prediction of temporal boundaries for events in Twitter. In: IEEE Int. Conf. on Social Computing. IEEE; 2011.
30. Wang X, Gerber MS, Brown DE. Automatic Crime Prediction using Events Extracted from Twitter Posts. In: SBP; 2012.
31. Lagi M, Bertand KZ, Bar-Yam Y. The Food Crises and Political Instability in North Africa and the Middle East; 2011. ArXiv:1108.2455v1: 15 pages. Available from: http://arxiv.org/pdf/1108.2455v1.pdf.
32. Braha D. Global Civil Unrest: Contagion, Self-Organization, and Prediction. PLoS One. 2012;. doi: 10.1371/journal.pone.0048596 PMID: 23119067
33. Radinsky K, Horvitz E. Mining the Web to Predict Future Events. In: WSDM; 2013.
34. Weber I, Garimella VRK, Batayneh A. Secular vs. Islamist Polarization in Egypt on Twitter. In: ASONAM; 2013.
35. Bell S, Cingranelli D, Murdie A, Caglayan A. Coercion, capacity, and coordination: Predictors of political violence. Conflict Management and Peace Science. 2013;. doi: 10.1177/0738894213484032
36. Sandra González-Bailón JBH, Moreno Y. Broadcasters and Hidden Influentials in Online Protest Diffusion. American Behavioral Scientist. 2013;p. 943–965.
37. Garey M, Johnson D. Computers and Intractibility; 1979.
38. Chung F, Lu L. Connected Components in Random Graphs with Given Expected Degree Sequences. Annals of Combinatorics. 2002; 6:125–145. doi: 10.1007/PL00012580
39. Saha S, Adiga A, Vullikanti AKS. Equilibria in Epidemic Containment Games. In: The 28th AAAI Conference on Artificial Intelligence (AAAI); 2014.
40. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73(3):273–282. Available from: http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x. doi: 10.1111/j.1467-9868.2011.00771.x
41. Hastie TJ, Tibshirani RJ. Generalized additive models. London: Chapman & Hall; 1990.