Information Relaxation is Ultradiffusive

ABSTRACT

We investigate how the overall response to a piece of information evolves and relaxes as a function of time in social networks like Reddit, Digg and Youtube. This response or popularity is measured in terms of the number of votes/comments that the story (or article) accrued over time. We find that the temporal evolution of popularity can be described by a universal function whose parameters depend upon the system under consideration. Unlike most previous studies, which empirically investigated the dynamics of voting behavior, we also give a theoretical interpretation of the observed behavior using ultradiffusion.

Whether it is the inter-arrival time between two consecutive votes on a story on Reddit or the comments on a video on Youtube, there is always a hierarchy of time scales in information propagation. One vote/comment might occur almost simultaneously with the previous, whereas another vote/comment might occur hours after the preceding one. This hierarchy of time scales leads us to believe that the dynamical response of users to information is ultradiffusive in nature. We show that a ultradiffusion based stochastic process can be used to rationalize the observed temporal evolution. We demonstrate that the results predicted by simulating the ultradiffusion process are in close correspondence to the actual observations.

I INTRODUCTION

Recent years have seen an exponential increase in the popularity and growth of online social networks. Although these networks have not yet replaced the traditional sources of information like the news websites, they are fast becoming alternate conduits of information. Be it the Arab Spring [1], the floods in Thailand in 2011 [2], the Hurricane Sandy in 2012 [3] or the Boston marathon bombing [4] or the recent bus -train crash in Ottawa in 2013 [5], social media has started playing an increasingly important role in the dissemination of information and crisis management. The propensity of viral spread of information through these social networks, have led to their growing and successful use in marketing campaigns (like Nike's "finding greatness" campaign [6]) and even electoral campaigns. Their effectiveness in spreading information have made them powerful mouthpieces for stars and have created their own league of celebrities like Justin Bieber and Lady Gaga. Thus, understanding the dynamics of information response in these networks is crucial to characterize, predict and improve their role as useful information channels.

There has been a lot of work to understand information flow on networks. Work in this direction includes the study of information spread in the blogosphere [12], the virality of information content [17] and recommendations [19], role of social networks in information spread [4] and so on. The role of individuals in the spread of information and their ability to influence others has also been actively studied [8,28]. Often, the underlying network through which information spreads is not known and there have been attempts to infer near-optimal networks that might lead to the observed contagion. Unlike [10], the objective of this paper is *not* to infer the structure of the social network responsible for the observed temporal dynamics.

The focus is the study of the temporal evolution of user response to information and factors that can rationalize the observed behavior without any knowledge or assumptions about the underlying social network. The response to information (like a story or video) is often in terms of rebroadcasts (like votes or comments). The number of rebroadcasts accrued over time gives a measure of popularity growth of that piece of information with time. Like [18], we empirically investigate this temporal evolution for different social media. However, unlike most of the prior research in this direction, we also provide a theoretical interpretation for these observations using ultradiffusion. Employing ultradiffusion to understand this temporal phenomena gives us invaluable insights into the relaxation of information response. It can serve as a powerful analytic tool can complement work that attempts to explain information spread based on the network properties.

One rebroadcast can occur immediately after the previous while another rebroadcast might occur a few minutes after the previous. Further, another rebroadcast might occur hours after the one preceding it. Hence there exists a hierarchy of time scales for the relaxation of information response. Some examples of hierarchical systems are taxonomy of objects, concepts, events and relationships; structure of organisms having atomic cells as basic building blocks; ontology in information science so on and so forth. A common feature of hierarchical systems is the ex-

[1] http://smallwarsjournal.com/jrnl/art/what-the-arab-spring-tells-us-about-the-future-of-social-media-in-revolutionary-movements

[2] http://www.techinasia.com/thailand-flood-social-media-innovation/

[3] http://www.theguardian.com/world/us-news-blog/2013/feb/20/mta-conedison-hurricane-sandy-social-media-week

[4] http://www.voanews.com/content/multi-social-media-play-huge-role-in-solving-boston-bombing/1649774.html

[5] http://www.huffingtonpost.ca/2013/09/18/oc-transpo-via-rail-crash-ottawa_n_3948062.html

[6] http://www.copywriterjournalist.com/2012/08/05/nike-find-your-greatness-campaign-videos-london-2012-olympics/

©ASE 2012

istence of many timescales leading to anomalous dynamics of relaxation. Physicists have described such dynamics using *ultradiffusion* [1, 14]. Due to the hierarchical time scales of information response, we postulate that ultradiffusion can also be used to describe this phenomena.

We study the temporal evolution of information response in social media like Twitter, Reddit and Youtube. We quantify this dynamics using different measures of rebroadcast such as votes or comments. We observe that the information relaxation or the popularity growth follows a universal function irrespective of the social media. The parameters of the function can be learned from the data. In contrast to the Poisson process, ultradiffusion can successfully explain the exponential nature of this function.

Ultradiffusion shows that the state space of the associated stochastic process has a hierarchical structure and the observations are a consequence of transitions within this structure. Based on the observed temporal response evolution, we can make powerful deductions about the the underlying hierarchical structure and the ultradiffusion process. For example, if the observed temporal trend is exponential, we can deduce that the hierarchical structure is finite and learn its parameters from the observations.

Ultradiffusion is closely associated with the notion of ultrametric distance. Along with common distance metrics like the Euclidean and Manhattan distance, it is another important though rarely used distance metric. One of our key contributions is to show the ultrametricity of information response. We expose the ultrametric distance between any two rebroadcasts (votes or comments). Ultradiffusion can then be simulated using transitions in the hierarchical state space of the stochastic process where the probability of transition is a function of the distance between states.

We demonstrate that that the probability of rebroadcasting predicted by the ultradiffusive stochastic process is very closely correlated with the observed probability of rebroadcasting irrespective of the social media considered. Our model seems to describe dynamics of voting on stories in Digg and Reddit and commenting on videos on Youtube with high accuracy. We also show that using ultradiffusion to explain temporal response relaxation phenomena is not just limited to social media. The temporal evolution of downloads of a research paper [15] can also be explained using ultradiffusion.

In the next Section we discuss related work. In Section III, we introduce ultrametricty and ultradiffusion. We describe the datasets in Section IV followed by measurements and observations (Section V). We show how ultrametricity can be used to model ultradiffusion. Ultradiffusion is then used to explain the behavior observed. In Section VI, we describe the experiments, followed by the results and evaluation (Section VII). We conclude with discussion and future work in Section VIII.

II RELATED WORK

Reference [26] define a class of searchable networks. Their definition of searchablity is based on the assumption of hierarchy of the social space like a community. The distance between two people is assumed to be ultrametric. We on the other hand claim that observed dynamics or response to information could be explained using ultradiffusive stochastic process. An underlying ultrametric structure of the social network could be one of the many potential causes for the ultrametricity in dynamics.

A lot of network dependent research has been done on information flow. It includes work on how information spreads in a social network and the effect of the structure and clustering of the network on this spread [2, 7, 12, 17–19, 27] Spreading processes have been extensively studied in epidemiology which also make implicit or explicit assumptions about network dynamics. Models like SIS (susceptible-infectious-susceptible) and SIR (susceptible-infectious-recovered) [3,13] implicitly assume a fully connected underlying network. To take the heterogeneity into account, [16] represented interactions between nodes as a directed graph to study the spread of viruses. To investigate the propagation of diseases, [22] assumed the structure of the network to be scale-free. [6, 23] emphasize on the importance of the largest eigenvalue of the adjacency matrix of the network in a virus propagation model. [21] showed the effect of having a network with power-law degree distribution with an exponential cutoff where information spreads. [11, 28] argued that information flow might not be analogous to viral epidemics because even if the underlying network is the same, how people spread information is different from how viruses spread. [9] propose Orthogonal Sparse LOgistic Regression method to predict epidemic outbreaks. [10] attempt to a near-optimal structure of the social network responsible from the observed contagions. The role of individuals within a network [8, 24, 28] and the role of social networks in information spread [4] has also been extensively studied.

All of these studies, investigate or try to understand some empirical property or characteristics of information flow as a function of the underlying social network. On the other hand, like [25] we attempt to understand the response to information as a function of time independent of the underlying network. [25] show that initial evolution of votes on stories in Digg and Youtube could be used to forecast their future popularity. We on the other

©ASE 2012

hand study the temporal evolution pattern of rebroadcasts (votes or comments) on Reddit, Digg and Youtube and show that it can be modeled using ultradiffusion.

While there has been work [5,20] on the inter-arrival times between emails (events), we focus on the evolution of a event occurrences or response to information with time.

III ULTRAMETRICITY AND ULTRADIFFUSION

One of the important characterization of information is the relaxation of its response or its popularity trace over time. Here, we provide an in depth empirical, experimental and theoretical analysis of this trace. If a social news aggregator is considered and the information is a news story, its popularity with time is measured by the number of votes it collects over time. Similarly, if the social media is a video sharing website and the information propagating is a video shared, the observed phenomenon indicating change in popularity might be the number of comments that the video receives with time. We take some real-life examples of temporal evolution of response to information. We consider the growth of popularity of stories on Reddit, news on Digg and videos on Youtube. We also study the number of downloads of a research paper after it was widely publicized [15].The quantifiable measure of response to information in the four cases considered are: the increase in the number of votes on Digg and Reddit stories over time, the temporal increase in the number of comments on Youtube videos and the number of downloads from a university server with time. We define the observable response (voting/ commenting/downloading) as the *rebroadcast phenomena*. Each vote/comment/download is an *event*. The temporal evolution of the rebroadcast phenomena is called the *counting process*.

Ultrametricity: The observed counting process could perhaps be rationalized as ultrafiffusive if it has signatures associated with an ultradiffusive stochastic process. Ultradiffusion is a stochastic process and like any stochastic process comprises of a state space S which is a collection of random variables $(X_t, t \in T)$ where t represents time, X_t is the random variable associated with a rebroadcast/ event at time t. The collection of timestamps at which event occurrences are considered is T. X_t is 1 when an event occurs and is 0 otherwise.

Unlike Poisson process which assumes that the event occurrences are independent of each other, ultradiffusion elicits that a later event might be caused by or correlated to an earlier event or a combination of earlier events. The earlier event in turn might be independent or it might be correlated to a combination of even earlier events. This leads to a hierarchical causal/correlational model of prior event occurrences which can be used to predict the occurrence of a new event. In other words, ultradiffusion proposes that the observed pattern of events is a consequence of an underlying hierarchy of states.

In this hierarchical model, an event temporally nearer to the occurring event has a greater probability of affecting it. In other words the correlation between two events is determined by a notion of "closeness" or distance between them. Ultradiffusion hypothesizes that this distance between two event occurrences is ultrametric. A set of states $X_t \in S$ form an *ultrametric space* and function $d(X_{t_i}, X_{t_j}) \rightarrow R$ is an *ultrametric distance metric* if:

1. $d(X_{t_i}, X_{t_j}) \geq 0$ (non-negative)
2. $d(X_{t_i}, X_{t_j}) = 0$ if $i = j$
3. $d(X_{t_i}, X_{t_j}) = d(X_{t_j}, X_{t_i})$ (symmetry)
4. $d(X_{t_i}, X_{t_j}) \leq max(d(X_{t_i}, X_{t_k}), d(X_{t_k}, X_{t_j}))$ (*ultrametric property i.e.* the distance between any arbitrary events X_{t_i} and X_{t_j} is less than or equal to the maximum of the distance between state X_{t_i} and X_{t_k} (any other state) and that between X_{t_k} and X_{t_j}).

Ultradiffusion: The autocorrelation function $P_{X_{t_i}}(t)$ is the probability of finding the system at the initial state X_{t_i} after time t. Ultradiffusion is best studied by resorting to the master equation describing the evolution of this autocorrelation function with time and is given by [1]:

$$\frac{dP_{X_{t_i}}}{dt} = \sum_{j=1}^{N} \epsilon_{X_{t_i} X_{t_j}} P_{X_{t_j}} \tag{1}$$

Here ϵ is transition matrix, whose element corresponding to states X_{t_i} and X_{t_j}, $\epsilon_{X_{t_i} X_{t_j}}$ satisfies the ultrametric property:

$$\epsilon_{X_{t_i} X_{t_j}} \geq min(\epsilon_{X_{t_i} X_{t_k}}, \epsilon_{X_{t_k} X_{t_j}}) \tag{2}$$

for distinct i,j and k. Here N is the number of states. The conservation of probability requirement entails that for $i = j$ i.e. the diagonal elements of the transition matrix, $\epsilon_{X_{t_i} X_{t_i}} = -\sum_{i \neq j} \epsilon_{X_{t_i} X_{t_j}}$.

The transition matrix ϵ can be any monotonically decreasing function of the ultrametric distance function $d(X_{t_i}, X_{t_j})$ for distinct states X_{t_i} and X_{t_j}. Without loss of generality we take $\epsilon_{X_{t_i}, X_{t_j}} = e^{(-\mu d(X_{t_i}, X_{t_j}))}$ for $i \neq j$ where μ is the scaling function.

©ASE 2012

Bachas and Huberman [1] have shown that the exact solution for the autocorrelation function associated with Equation 1 (for an ultrametric space defined by a hierarchical tree) is of the form :

$$P_{X_i}(t) = \alpha + \sum \beta_n e^{-\gamma_n t} \quad (3)$$

The constants α, β_n and γ_n can be derived analytically as shown in the Appendix. When the number of states is finite, the resultant autocorrelation function is exponential in nature. Otherwise, it follows a power law as shown in the section on ultradiffusion in the Appendix. Next, we describe the dataset studied.

IV DATASETS

We study data from social networking sites, Digg, Youtube and Reddit. We also investigate the evolution of downloads from a web server [15].

Reddit: Reddit [7] is a social news and entertainment website. It has a homepage and other pages like "rising", "controversial", "new" where stories are ranked based on different criteria. Registered users can post a links to hyperlinked stories/news elsewhere or some can write their own ("self") stories/news in text format. Other users can vote in favor of or opposing these submissions. These votes determine the ranking of the story in the different ordered lists/pages of stories hosted by this website. The data [8] we used has the votes (and the timestamps of these votes) on more than 16000 links in Reddit. There were 31,927 unique users and more than 7 million votes. Of these 9700 stories/links having more than 50 votes which comprise our dataset. As expected the distribution of votes over stories follows a power law distribution, with many stories receiving few votes and few stories receiving many votes. We study, the temporal evolution of votes in each of these stories and on average.

Digg: Digg[9] is a social news aggregator, which like Reddit, allows users to posts links to news stories. When a story is submitted by a user, other users can vote on it. The number of diggs or votes received by a story shows how popular a story is and plays an important role in its ranking with respect to other stories on the front page of Digg. The temporal voting behavior on a story is indicative of the story's popularity on Digg with time. The dataset we used [18] comprises of voting on stories promoted to Digg's front page over a period of a month in 2009 (July). It contains 355K votes on 2060 popular stories made by 100K distinct users.

Youtube: Youtube[10] is a popular video sharing platform. An individual can upload a video on Youtube others can view it. Often some videos receive word-of-mouth appreciation and spread through friendship networks and become viral. For example, one of the viral videos in 2012 "Gangnam Style" reached over a billion views [11]. People also have the option of commenting on these videos. The commenting behavior is the quantifiable measure of the change in information popularity over time. We extracted the comments and the timestamps at which these comments were posted for more than 20 videos of advertisements in 2012 and 2013 on Youtube having more than 7500 comments. We study the evolution of the number of comments accrued by these videos and model it using ultradiffusion.

Server Downloads: Johansen et al. [15] studied the response of people to new information. They investigated and modeled the number of downloads of a research paper from an authors webpage in Niels Bohr Institute. This dataset was collected after the paper was publicized by a Danish newspaper. We show here that ultradiffusion gives an alternate (possibly better) explanation of the observed variation of downloading activity with time.

V MEASUREMENTS AND OBSERVATIONS

Like [18], we observe that most stories/videos in most social media accrue a lot of new votes/comments initially after being posted. Gradually the number of new votes/comments slow down with time and finally it reaches a saturation point, beyond which hardly any new votes/comments are received. Let time t be $t = 0$ when the story s is submitted. If for this story s, $n_s(t)$ is the number of votes at time t and M is the total number of votes at saturation, then the probability distribution of votes with time, $p(t, s)$, is given by $n_s(t) = M \cdot p(t, s)$. Using non-linear least curve data fitting, we fit the observed rebroadcasting behavior $p(t, s)$ with an exponential curve given by:

$$\bar{p}(t, s) = h_1 \cdot (1 - e^{-h_2 t}) \quad (4)$$

For a story s, the expected number of votes at time t ($E[N(t)]$), in terms of probability of no votes at time t ($\bar{p}_0(t, s)$) is defined by $E[N(t)] = M * (1 - \bar{p}_0(t, s)) = M * (\bar{p}(t, s))$.

We see that in most cases, for each story/video/download considered, this curve gives a very good fit with the ob-

[7] www.reddit.com

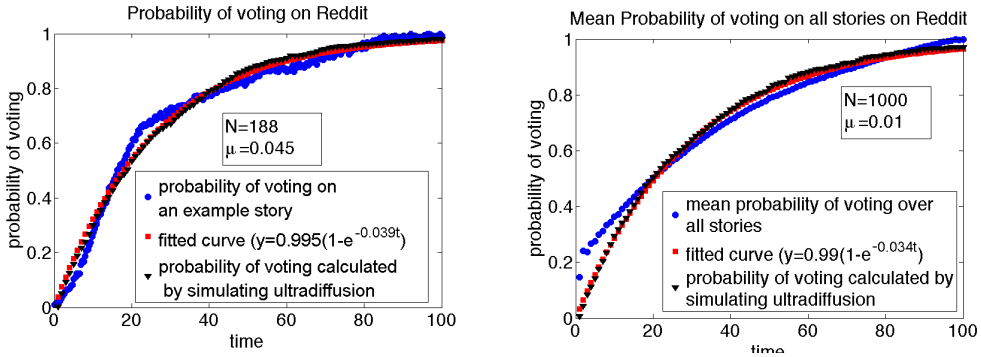
[8] http://www.reddit.com/r/redditdev/comments/bubhl/csv_dump_of_reddit_voting_data/

[9] www.Digg.com

[10] www.youtube.com

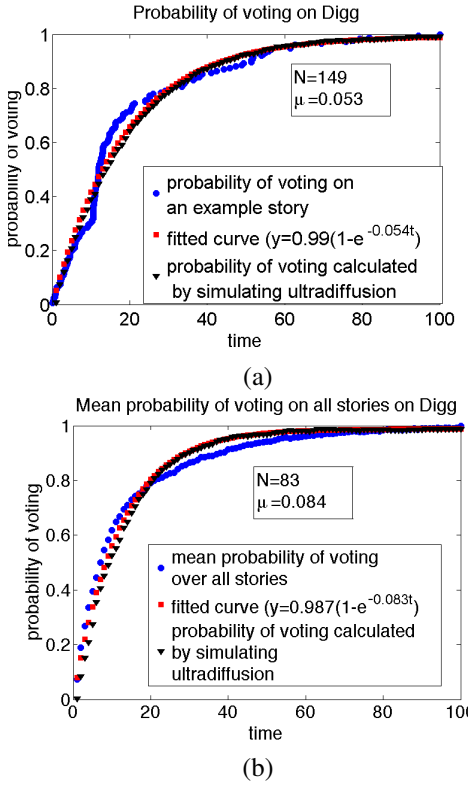
[11] http://articles.latimes.com/2012/dec/17/business/la-fi-ct-viral-video-top-ten-20121218

©ASE 2012



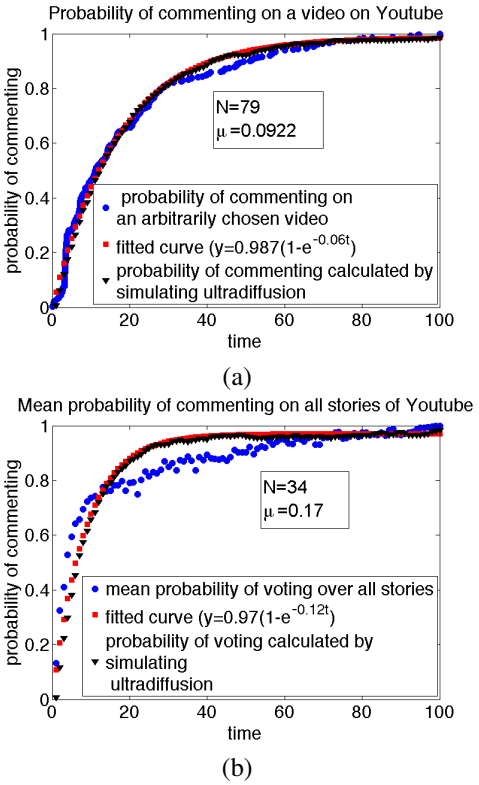
(a) (b)

Figure 1: (a)The observed probability of voting on Reddit with time for a sample story (in blue). The exponentially fitted curve is in red ($R^2 = 0.98$) and the calculated probability of voting from a simulated ultradiffusion process is shown in black ($R^2 = 0.98$). (b)The blue curve gives the observed mean probability of voting on Reddit. The exponential curve fitting is given in red ($R^2 = 0.96$). The mean probability of voting generated by an ultradiffusion simulation is given in black ($R^2 = 0.93$).



(a) (b)

Figure 2: (a)The probability of voting on Digg with time for a sample story. The observed probability of voting is in blue. The fitted exponential curve is in red ($R^2 = 0.976$) and the probability of voting obtained by simulating ultradiffusion is in black ($R^2 = 0.976$) (b)The average probability of voting over all stories is shown in blue and the fitted curve is shown in red ($R^2 = 0.97$). The average probability of voting simulated using ultradiffusion is in black ($R^2 = 0.95$).



(a) (b)

Figure 3: (a) The probability of commenting on a sample story Youtube. The curve in blue represents the observed probability of commenting. The red curve shows the fitted curve ($R^2 = 0.986$) and the probability of commenting obtained by simulating ultradiffusion is shown in black ($R^2 = 0.99$). (b) The mean probability of commenting is shown in blue. The fitted curve is in red ($R^2 = 0.80$). The probability of commenting derived from ultradiffusion simulation is in black ($R^2 = 0.80$)

©ASE 2012

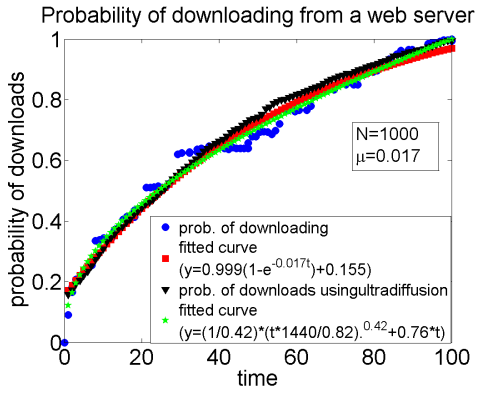


Figure 4: The probability of downloads of a paper from Niels Bohr Insititute server with time is shown in blue [15]. The curve in green ($y = (1/0.42) * (t * 1440/0.82).^{0.42} + 0.76 * t$)is one postulated to describe the phenomena in [15] ($R^2 = 0.988$). The curve shown in red ($R^2 = 0.98$) is the fitting of the universal function $y = h_1(1 - e^{-h_2 t}) + h_3$ ($h_1 = 0.999, h_2 = 0.017, h_3 = 0.155$). The black triangles show the evolution of downloads simulated by ultradiffusion($R^2 = 0.97$).

served probability density. To illustrate this, consider an example story sampled arbitrarily for Reddit, Digg and Youtube, shown in Figures 1(a), 2(a) and 3(a) respectively. The blue curves in Figures 1(a) 2(a) and3 (a) shows the probability of votes or comments with time on an example story or video in Reddit, Digg and Youtube respectively. The curves in red are the exponential fits (Equation 4). We see that these fits very well explain the observed phenomena as demonstrated by the high values of R^2 ($R^2 > 0.9$).

Similarly, in Figures 1 (b) 2(b),3 (b), the plots in blue, represent the mean probability density of voting/commenting with time averaged over all stories considered for each of the social media datasets. Again the exponential plots in red successfully explain the blue curves.

In the case of server downloads, the authors postulated that a functional curve of the form $N(t) = \frac{a}{1-b} t^{1-b} + c * t$ (in green) best describes ($R^2 = 0.988$) the evolution of the number of downloads or hits on the server with time (in blue). We note that this functional curve has a component increasing linearly with time. However, we have observed that most social phenomena and the counting process associated with them saturate with time. The downloads on a research paper wears down with time and probably does no go on forever linearly. Thus these observations only capture the transient state of the evolution of downloads. Instead, our universal curve not only describes the transient stage of the evolution very well but also can model saturation. We digitized the graph in Figure 1 of that paper. We show that the universal function $h_1(1 - e^{-h_2 t}) + h_3$ in red well describes the evolution of the number of votes with time ($R^2 = 0.98$).

Based on these observations, is it then possible to make interesting deductions about a probable stochastic process that could be used to explain them? Poisson process is a stochastic process often used to count and model the number of events occurring with time. In our case the occurring events are the votes, comments or downloads. However, if we assume the observed process in Poisson, then the probability of votes/comments would change linearly with time as we show in the Appendix. Contrary to that, we observe an exponential change in the probability of voting with time. One plausible reason for the Poisson process not being successful in describing the phenomena is because it assumes independence of events. However, the probability of a person commenting on a video might depend on a previous occurrence of commenting (maybe by his friend who initiated him to the video). Thus, the observable events like each vote/ comment/download might not be completely independent events. Rather they might be caused/correlated with each other leading to a hierarchy of correlation or causation between occurrences. This is a signature of an ultradiffusive process described above. In the next section we show how the expected number of votes and observed universal curve can be explained using ultradiffusion.

Ultrametricity of Response of Information Relaxation: If the process is indeed ultradiffusive, then there must exist an ultrametric space on which distances between occurrences is defined. We discover such an ultrametric space.

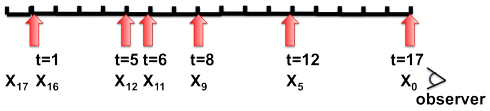
Let a time-series $X_{t_0}, X_{t_1} \cdots X_{t_n}$ ($t_i < t_j$, if $i < j$) be associated with event e. Let t_n be the bounding time on observations or the current time. The distance between two events X_{t_i} and X_{t_j} be given by

$$d(X_{t_i}, X_{t_j}) = \begin{cases} |max(t_n - t_i, t_n - t_j)|, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

A function so defined satisfies the *ultrametric distance metric properties* defined in the previous section and the associated space is ultrametric.

Let us illustrate this distance further with a toy example. Consider a story s whose voting trace is shown in Figure 5. In other words votes(events) occur at times $t = 1, 5, 6, 8, 12$ and 17. The time of observation is $t = 17$. With each occurrence of a vote at time t_i, we associate a random variable $X_{t_n - t_i}$ where t_n is time of observation or current time. The random variable $X_{t_n - t_i}$ describes the occurrence (or nonoccurence) of an event at time $t_n - t_i$

©ASE 2012



(a)

Figure 5: A timeline of events (rebroadcasts). The red arrows in the timeline show when the events occur at times $t = 1, 5, 6, 8, 12, 17$. Matrix in Equation 6 shows the distance between event in the i^{th} row and the j^{th} column.

prior to observation.

The distance between the different voting occurrences is shown by the matrix in Equation 6. For instance, as we see in Equation 6, the distance between the event occurring at time $t = 6$, X_{11} and that at time $t = 12$, X_5 is given in the row corresponding to X_{11} and the column corresponding to X_5. This distance can be seen to be 12 from this look up matrix. The distance between X_{17} (state when no votes occur) and X_{11} is 6.

$$\begin{array}{c} \\ X_0 \\ X_5 \\ X_9 \\ X_{11} \\ X_{12} \\ X_{16} \\ X_{17} \end{array} \begin{array}{c} \begin{array}{ccccccc} X_0 & X_5 & X_9 & X_{11} & X_{12} & X_{16} & X_{17} \end{array} \\ \begin{pmatrix} 0 & 17 & 17 & 17 & 17 & 17 & 17 \\ 17 & 0 & 12 & 12 & 12 & 12 & 12 \\ 17 & 12 & 0 & 8 & 8 & 8 & 8 \\ 17 & 12 & 8 & 0 & 6 & 6 & 6 \\ 17 & 12 & 8 & 6 & 0 & 5 & 5 \\ 17 & 12 & 8 & 6 & 5 & 0 & 1 \\ 17 & 12 & 8 & 6 & 5 & 1 & 0 \end{pmatrix} \end{array} \quad (6)$$

The existence of many time-scales in rebroadcasting behavior and our ability to construct an ultrametric space to represent the rebroadcasting behavior, leads us to speculate that response to information is an ultradiffusive process.

To simulate ultradiffusion, we first need to have a transition matrix. We exposed the ultrametric space for information response and the associated ultrametric distance above which we use to create the transition matrix ϵ using transition rates $\epsilon_{X_{t_i},X_{t_j}} = e^{(-\mu d(X_{t_i},X_{t_j}))}$. The eigen spectrum of this transition matrix is provided in the Appendix.

Information Relaxation is Ultradiffusive: Using the transition matrix defined above and Equation 1, we can calculate the probability of being in any state at time t. Let the starting vector be $P_{X_{t_i}}(0) = \chi_{X_{t_i}}$ where $\chi_{X_{t_i}}$ is a characteristic function (i.e. $P_{X_{t_i}}(0)$ is a vector whose element corresponding to X_{t_i} is 1 and it is zero elsewhere). The autocorrelation function is $P_{X_{t_i}}(t)$ which corresponds to state X_{t_i}.

$P_{X_{t_N}}(t)$ defines the probability of being in state X_{t_N} at time t. If X_{t_N} is analogous to the non-rebroadcasting (voting, commenting) state (like X_{17} in Figure 5) and $P_{X_{t_N}}(t)$ is analogous to the probability of not rebroadcasting till time t, then taking M as the maximum number of rebroadcasts possible, $M*(1-P_{X_{t_N}}(t))$ is the expected number of rebroadcasts at time t. Since there are no votes initially, therefore the probability of not voting $i.e.$, $P_{X_{t_N}}(0) = 1$ and $P(0) = \chi_N$.

If the number of states in the transition matrix are finite, we know that the decay of the autocorrelation function is exponential [1]. Specifically for transition matrix defined using ultrametric distance in Equation 5 (and whose eigen spectrum is discussed in details in the Appendix), we have $P_{X_{t_i}}(t) = \sum_{j=1}^{t_N} e^{(\lambda(j)t)} V_i^2(j)$. Therefore, $P_{X_{t_N}}(t) = \frac{t_N-1}{t_N} * e^{-t(t_N \dot{e}^{(-\mu(t_N-1))}} + \frac{1}{t_N}$ and the expected number of rebroadcasts after time t is $E(R(t)) = M*(1-P_{X_{t_N}}(t)) = \frac{1}{t_N}*(1-e^{-t(t_N e^{(-\mu(t_N-1))}})$.

VI EXPERIMENTS

We hypothesize the ultrametricity of information response and that relaxation of information response is ultradiffusive. When the information is a story shared on Reddit or Digg or a video shared on Youtube, we claim its growth in popularity can be rationalized using a ultradiffusion based stochastic process.

To verify this claim, for each story/news/video in the social media we simulate its spread using an ultradiffusion process. The transition matrix describing the ultradiffusion process is determined by the ultrametric distance $d(X_{t_i}, X_{t_j})$ between events X_{t_i} and X_{t_j} are defined as in Equation 5. The transition rates between X_{t_i} and X_{t_j} are given by $e^{-\mu d(X_{t_i},X_{t_j})}$. However, prior to simulation, we need to determine the values of t_N (the bounding time on observations) and μ (the scaling factor) for the ultradiffusion process to be simulated. We learn these variables from the observed data on rebroadcasting behavior. We use h_1 and h_2 in Equation 4 to learn the values of t_N and μ. The length of the hierarchical tree t_N is given by $t_N = \frac{1}{1-h_2}$. The decay parameter μ is deduced as $\mu = log(\frac{t_N}{h_1})/(t_N - 1)$.

Once we have deduced t_N and μ, we can then use these values to calculate the transition matrix ϵ encoding the ultradiffusion process. We assume that the ultradiffusion process starts at state X_{t_N} $i.e$ $P_{X_{t_N}}(0) = \chi_{X_{t_N}}$ in Equation 1. As explained above, $P_{X_{t_N}}(t)$, gives the probability of not voting which is 1 at $t = 0$. We then simulate the ultradiffusion process using the learned parameters t_N and μ. The rebroadcasting behavior obtained by simulating ul-

©ASE 2012

tradiffusion is then given by

$$\hat{p}(t,s) = 1 - P_{X_{t_N}}(t) = \frac{1}{t_N}(1 - e^{-(tt_N e^{(-\mu(t_N-1))})}) \quad (7)$$

We plot the observed probability curve $p(t,s)$ in blue, the fitted probability curve $\bar{p}(t,s)$ (in red) and the probability curve obtained from simulated ultradiffusion $\hat{p}(t,s)$ (in black) with time t for every story/news/video s investigated in the different social media websites.

To get the overall trend of information spread in the different social media, we average the probability of voting across different stories/news/videos for each social media website. Then we simulate an ultradiffusion process generating this overall trend. In other words, just like for individual stories, for each social media, we fit the mean probability of voting with an exponential curve. We use this fit to determine the t_N and μ of an ultradiffusion process. We then calculate the mean probability of voting with time using this ultradiffusion process and compare it with the observed mean probability of voting with time. We evaluate the goodness of fit using R^2 which measures how successful the fit is in explaining the variation of the data. High value of R^2 indicates a good fit.

Irrespective of the social media studied, we observe that for most stories/news/videos s, the generated rebroadcasting behavior by simulating ultradiffusion represented by $\hat{p}(t,s)$ closely agrees with the observed rebroadcasting behavior given by $p(t,s)$. We also observe that the ultradiffusion process seems to explain the overall rebroadcasting behavior in the different social media satisfactorily. We demonstrate these results in greater details in the next section.

VII RESULTS AND EVALUATION

In this section we present our observations of how the cumulative probability density of votes in Reddit and Digg and comments in Youtube and (downloads [15]) changes with time. For most of the stories the rebroadcasting behavior calculated using the ultradiffusion model fits very well with the observed probability of voting with time. We calculate the observed probability of rebroadcast for each individual story in Reddit, Digg and Youtube as shown in Figures 1(a),2(a) and 3(a) in blue, the best exponential fit to this curve in red and the predicted probability of voting using ultradiffusion in black. Multiplying probability of rebroadcasts (votes) with the maximum number of rebroadcasts(votes) gives the expected number of votes at time any time t for that story. Similarly we also calculate and plot the observed mean probability of rebroadcasts over a collection of stories in each of these social media as as shown in Figures 1(b),2(b) and 3(b) (in red), the exponential fit (in red) and process simulated by ultradiffusion (in black). We see a high correspondence between the actual and the simulated rebroadcasting behavior irrespective of the social media considered.

Reddit: We investigate the voting behavior for individual stories in the Reddit dataset (an example is shown in Fig. 1 (a)) and the average voting behavior over all stories studied (Fig. 1 (b)). We evaluate how successful ultradiffusion is in explaining the probability of voting (both for each individual stories and for the average over all stories on Reddit) using R^2. In most cases we get a very high value of $R^2(> 0.93)$ as demonstrated in Fig. 1. This implies that ultradiffusion is successful in describing rebroadcasting behavior in Reddit.

Digg: In Fig. 2 (a), we give an example of the popularity growth of an individual news article on Digg in terms of diggs. In Fig. 2 (b) we plot the average popularity growth. Again high value of R^2 both for individual stories and for the average of of all stories on Digg ($R^2 = 0.95$) as shown in Fig. 2 demonstrates that ultradiffusion explains with high accuracy the rebroadcasting behavior on Digg.

Youtube: Just like in Reddit and Digg, we also measure the probability of commenting on individual stories with time(Fig. 3 (a)) and mean probability of commenting over all Youtube videos collected over time (Fig. 3 (b)). Again on evaluating the goodness of fit, we get a high value of $R^2(> 0.8)$, suggesting that ultradiffusion very well explains the data.

Server Downloads: We observe in Figure 4, that ultradiffusion can also be successfully describe the evolution of the number of downloads from a server with time. The observed probability of downloads with time is shown in blue and the probability predicted using ultradiffusion is shown in black . Again we see a very high correspondence between the actual observations of the counting process and that predicted by ultradiffusion($R^2 = 0.97$)

We successfully demonstrate that in each of the four datasets the predicted probability of votes by simulating ultradiffusion process (both for individual stories and the overall social) is very close to the observed probability of voting. Therefore, we claim that the rebroadcasting behavior generated by the ultradiffusion process is very similar to the actual rebroadcasting behavior observed.

VIII CONCLUSION

We study the relaxation of human response to information on online social networks. For news, stories or videos, this

©ASE 2012

response is in terms of votes or comments that the story, news or video receives over time. We notice that the inter-arrival time between rebroadcasts is not homogeneous and there exists a hierarchy of time scales. Some consecutive rebroadcasts occur almost simultaneously whereas others might occur with along interval within them. This leads us to postulate that these occurrences of these events might not necessarily be completely independent with some events being correlated with each other, which is the signature of an ultradiffusive process. We discover a universal function describing the observed counting process regardless of the social media considered; be it the probability of voting on a story on Reddit or Digg or the probability of commenting on a video on Youtube with time. The parameters of this function can be learned from the dataset under consideration. The universal function observed can indeed be well described using ultradiffusion. On the other hand, it cannot be derived using the Poisson process.

We know that for transition matrices corresponding to finite trees, the decay of the autocorrelation is *exponential* with time. However, if the hierarchical tree is infinite, there is a *power law* decay of the autocorrelation function with time (*cf.* Appendix). Our ability to rationalize information relaxation using ultrasiffusion gives us the power to predict the future evolution of the rebroadcasts over time. Future work includes predicting the observable behavior in other social networks and disease spread.

References

[1] C. P. Bachas and B. A. Huberman. Complexity and ultradiffusion. *J. of Physics A: Mathematical and General*, 20(14):4995, 1987.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. KDD '06, pages 44–54, New York, NY, USA, 2006. ACM.

[3] N. T. J. Bailey. *Mathematical Theory of Infectious Diseases*. Oxford University Press, 1975.

[4] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. WWW '12, New York, USA, 2012. ACM.

[5] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[6] C. Castellano and R. Pastor-Satorras. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.*, 105:218701, Nov 2010.

[7] D. Centola. The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996):1194–1197, Sept. 2010.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[9] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, and S. Yang. Cascading outbreak prediction in networks: a data-driven approach. KDD '13, pages 901–909, New York, NY, USA, 2013. ACM.

[10] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. WSDM '13, pages 23–32, New York, NY, USA, 2013. ACM.

[11] Greg Ver Steeg, R. Ghosh, and K. Lerman. What stops social epidemics? In *ICWSM*, 2011.

[12] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[13] H. W. Hethcote. An immunization model for a heterogeneous population. *Theor. Pop. Biol.*, 14:338–349, 1978.

[14] B. A. Huberman and M. Kerzberg. Ultradiffusion-the relaxation of hierarchical systems. *Journal of Physics A: Mathematical and General*, 18(6):4470, 1985.

[15] A. Johansen and D. Sornette. Download relaxation dynamics on the {WWW} following newspaper publication of {URL}. *Physica A: Stat. Mechanics and its Applications*, 276:338 – 345, 2000.

[16] J. Kephart and S. White. Directed-graph epidemiological models of computer viruses. In *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pages 343–359, 1991.

[17] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? WWW '10, pages 591–600, New York, USA, 2010. ACM.

[18] K. Lerman and R. Ghosh. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

©ASE 2012

[20] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, 2008.

[21] M. E. J. Newman. *Physical Review E*, (1):016128, July.

[22] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr 2001.

[23] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos. Got the flu (or mumps)? check the eigenvalue! 2010.

[24] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 113–114, New York, NY, USA, 2011. ACM.

[25] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.

[26] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296(5571):1302–1305, 2002.

[27] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality Prediction and Community Structure in Social Networks. *Scientific Reports*, 3, Aug. 2013.

[28] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, June 2004.

IX APPENDIX

Poisson Process Let ρ be the rate of arrival of a Poisson process. The probability of the number of votes at time t being k is given by
$P(N(t) = k) = (e^{-\rho t}\rho t^k)/k!$

Since $\sum_{k=0}^{\infty}(e^{-\rho t}\rho t^k)/k! = 1)$ we get:
$E[N(t)] = \sum_{k=0}^{\infty} k(e^{-\rho t}\rho t^k)/k! = \rho t$
If T_0 is the time at saturation, then the probability of an event occurring is given by $\frac{t}{T_0}$. Hence the probability of an event occurring varies linearly with time.

Ultradiffusion Reference [1] gives the exact solution for the autocorrelation at state L which is taken as a leaf of a hierarchical tree:

$P_L(t) = \frac{1}{N} + \sum_{n=1}^{root}(\frac{1}{N_{L_{n-1}}} - \frac{1}{N_{L_n}})e^{(-\frac{t}{\tau_{L_n}})}$
Here N is the number of leaves of the tree, L_n is the node in the n^{th} level of the shortest path from leaf L to the root. The level of the leaf L is minimum ($n = 1$) and that of the root is maximum. N_{L_n} are the number of final descendants or leaves of node L_n. $\frac{1}{\tau_{L_n}}$ is given by:
$\frac{1}{\tau_{L_n}} = N_{L_n}e^{(-h_{L_n})} + \sum_{i=2}^{root}(N_{L_i} - N_{L_{i-1}})e^{(-h_{L_i})}$
Here h_{L_i} is the height of node L_i. The height of the leaves is the minimum and that of the root is the maximum. Thus $\alpha = \frac{1}{N}$, $\beta_n = \frac{1}{N_{L_{n-1}}} - \frac{1}{N_{L_n}}$ and $\gamma_n = \frac{1}{\tau_{L_n}}$ in Eq. 3.

Power Laws If we consider a hierarchical tree, in which every node produces b offsprings and Δh is the height between two levels, then the silhouette of the tree is given by $s = \frac{1}{\Delta h}logb$.
If $s < 1$ the autocorrlation function averaged over all initial conditions is given by

$\bar{P}(t) = \sum_{m=1}^{\infty}(b-1)b^{-m}e^{(-t(be^{-\Delta h})^m \frac{e^{\Delta h}-1}{e^{\Delta h}-b})} = Dt^{-v}$
Here $v = \frac{s}{1-s}$ and $D = \Gamma v(\frac{e^{\Delta h}-1}{e^{\Delta h}-b})^{-v}\frac{b-1}{logb}v$. [1] also show that uniformly random trees have the same dynamic exponent(silhouette) as completely ordered uniform trees.

Eigen Spectrum of the Transition Matrix We construct a transition matrix for a time series of events $X_{t_1}, X_{t_2} \cdots X_{t_N}$ ($t_i < t_j$, if $i < j$) whose ultrametric distance metric is defined by $d(X_{t_i}, X_{t_j}) = |max(t_N - t_i, t_N - t_j)|$ if $i \neq j$ and 0 otherwise. For this matrix we construct a hierarchical tree in the state space in following [1] and calculate the normalized eigenvalues using the characteristic function where $X_{t_i}(j) = 1$ if if X_{t_i} is a descendant of X_{t_j} and is 0 otherwise. N_y are the number of descendants or leaves generated by y. We take y as a son of y_1. We note that $v_{t_i}(y)$ is a basis for the subspace the eigenvectors span, one for each son of y. $v_{t_i}(y) = \frac{1}{N_y}X_{t_i}(y) - \frac{1}{N_{y_1}}X_{t_i}(y_1)$.

Taking the degeneracy and linear independence under consideration, we derive analytically that if there are t_N leaves, then the normalized eigenvectors are given by

$$V_i(j) = \begin{cases} 1/\sqrt{t_N} & \text{if } j = 1 \,\forall\, i \\ 1/\sqrt{(j-1).(j)} & \text{if } 1 < j \leq t_N \text{ and } i < j \\ -\sqrt{j-1/j} & \text{if } 1 < j \leq t_N \text{ and } j = i \\ 0 & \text{if } 1 < j \leq t_N \text{ and } i > j \end{cases}$$

The corresponding eigenvalues are given by $\lambda(j) = -((j-1)e^{(-\mu(j-1))} + \sum_{i=j}^{t_N} e^{(-\mu(i-1)))}$ if $1 < j \leq t_N$ and 0 if $j = 1$.

©ASE 2012