Modeling Content Spread in Social Media

Sayaji Hande
Adobe Research
hande@adobe.com

Vineet Gupta
Adobe Research
vinegupt@adobe.com

Sandeep Zechariah George
Adobe Research
sgeorge@adobe.com

ABSTRACT

Social networks play a major role in information spread. There have been studies in various domains, including but not limited to the diffusion of medical and technological innovations, the "word of mouth" effect in promotion of products, the spread of digital content in social media like videos, photos, etc. Information reaches us in two ways: through the source from where the information originated and through our friends (connections) in social network. The availability of large data from social networks gives us a chance to study and analyze this process in detail.

The interesting and challenging aspect of this is to study and analyze early signals to make deductive claims about future. With this note, we present a *mathematical model* which tries to capture the process of content spread in social networks from source (of the content) to individuals via their connections. We then develop an efficient technique to fit the model parameters and then apply it on data. Using initial 10-20% of data points, we predict the remaining data points and then show the comparisons with actual data.

We found out that not only our model accurately captures the process of content spread, but it can also be used to make predictions for the future spread. Another aspect of our work is that it doesn't require any priori network information, which is very difficult to create in social networks, as it makes the assessment of the same using initial (10-20%) observation data. The model is also easily implementable, scalable, interpretable and flexible.

Key Words: viral marketing, viral product design, word-of-mouth (WOM), networks, opinion makers, cascading models

1. INTRODUCTION

The theory of adoption and diffusion of new ideas or new products by a social system has been discussed in depth by Rogers [11]. Also, it has been shown that either individuals adopt an innovation independently of the decisions of others, or they are influenced by their social connections to do so. Here, the 'influence' spreads through social connections of a user to another. Be it a newly released movie or a computer-game that has been launched, they thrive on this process However, the complex structure of the social networks and heterogeneity of individuals make it far from obvious how these local correlations affect the final outcome of the diffusion process.

In the past, the focus was primarily on direct marking and efforts were invested into it. Traditional retail industry has been following this path to ensure that the visible perceptions of the product by the end users are enhanced, so as to make the brand a success.

The process in which information as well as product awareness spreads within society via connections has been studied in past. The new era industry related to digital world, such as communications, social networking among others, is fast evolving and yet to mature. The changes due to technology, new products and service offerings are born whereas a few older ones without strong footing are fast diminishing. Excess over valuation of *Facebook* is latest example of 'herd' going wild and indicating that social spread, more often than not, is a key.

Hence, consistent measurement system and models that can readily be implemented are desired. Since the last decade, lot many changes have been happening in digital, Internet and mobile communication space. Though these changes as well as the usage of new services is driven by technology, the success depends a lot on the 'perceived' benefits by the consumers.

For example, can one would have assessed the video 'Gangnam'[1] going viral apriorily? Something going viral or popular is difficult to asses prior to its launch. What we propose in this paper is, a model, with ability to asses it by usage of observation of response in first 10 to 20% time frame.

In this specific problem, modeling the measurement of end-user perceptions that drive the performance are essential. The measurements could be # of likes on particular posts, # inquiries of new product launch, # of views of video etc. There are various models proposed in past starting from Hazard model used for assessing decease spread, *Diffusion model, Cascading effects, Linear Threshold,* among others. Details of related work will be discussed in section1.2. As stated earlier, we provide a model that would help user asses 'virality' or 'popularity' early-on, without too many assumptions.

Our paper is novel in four fronts: *a) It provides a usable model that has thought process of various models, Cascading,*

[1] $http://en.wikipedia.org/wiki/Gangnam\_Style$

Diffusion, Linear threshold, in decreasing order. b) For usage of implementation on real problem, it does not seek the network structure, it learns based on initial small window, like count of folks who are influenced till that time frame. c) The Model has been parameterized and recursive equations are provided for estimation. d) This paper provides you with a methodology for estimating cumulative curve of adopters after observing initial 10-20 % of window.

The model also provides you way to customize and embed your subjective views, like any other modeling technique, such as regression line etc. In this model you would need to subjectively asses population size as well as time interval size for stages. Also, treading of estimated stage-wise p_i's need to done.

Perceptions of consumers are formed by their own experiences (or judgments) or by the opinions & experiences of their close friends and that of influencer's (opinion makers). Also, the companies in order to create or maintain their brand values, need to manage these perceptions. This paper will provide you with a measurement and estimation system, that can be used with appropriate controls in users hand for projecting the future.

Viral notifications primarily happen in a passive way (product company is not involved). For instance, when a user uses a particular product or service and passively notifies his or her friends.

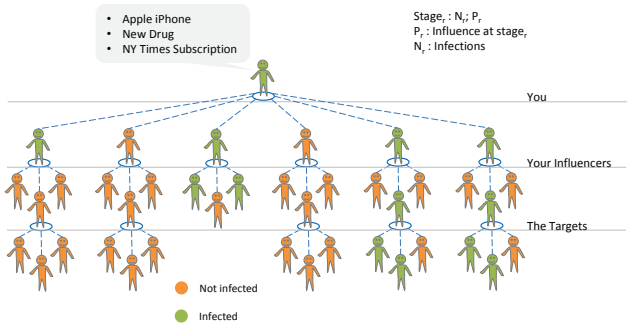


Figure 1: Infection Spread

Usage of Word-of-mouth (WOM) in the process of virality is a well known phenomena, it has been in use for centuries. For the new digital era too, companies are focusing on using it, but, there is lack of clear-cut methodology and model base which would be flexible enough to be applied in various scenarios. Be it the product marketing strategy or assessing the current strategy and its total reach.

The *aim of this paper* is to model the spreading behavior at each stage and how people get associated with it at different rate at various stages, as shown in diagram in this section. This paper provides an intuitive model with minimal assumptions, which could be easily applied in various situations. At stage 1, a few people get impacted directly through some 'seeds'. Then, Stage 2 spread happens only through folks who got influenced in Stage 1. And so on. This is a form of *Cascade models* that are referred in literature. At every stage, we then would have clear assessment of 'expected #' of influenced individuals. We will have stage-wise influence parameters. We provide a way to measure these parameters, and subsequently, using trend, provide a way to estimate cumulative # of people who would get influenced after n stages and so on.

We will then demonstrate the methodology with some real data of *YouTube* as well as *Facebook*.

Previously, some models have been proposed on similar objective, starting from Bass (1969) and subsequent papers. They will be discussed in section below.

1.1 Structure of the Paper

Next section will have relevant literature. Then, Model Formulation and Model approximations would be provided. At the end, study of empirical data is done. In the empirical data, we observe the 10-20% of observations called observation window. Estimate model parameters and project model parameters based on trend and predict the spread and compare with actuals.

It is to be noted that primary objective of this paper is link the parameters to that of model creation. After estimation, the 'trend modeling' can be done deem fit by user. In the data section called *Model Evaluation*, we have used linear trending. Also, at times, size of population that is needed, need to be judgmental and need subjective decisions as any prediction models.

1.2 Review of existing relevant Literature

Bass (1969) [1] formulated Diffusion Model, which was substantially used for projecting and assessing sales of consumer durables. Much of research work emerged after Bass (1969) in-line of Diffusion Process. New formulations in view of social networking made its way. Kempe, et al (2003) [3], Backstrom, et al (2006) [5], Kleinberg (2007) [4] among others formulated problem based on graph theory. Where as Richardson & Domingos (2001,2) [6], [6] modeled in probabilistic format. Research has also been done in extending these models, Myers et. al [7] consider diffusion model, but also take into account the out-of-network, external sources.

The book by Rogers (2003) [10] provides good source to literature related to diffusion process and its social aspects, innovator, early adopters etc. It also provides how possible influences work in society via various examples.

Watts (2002) [14] beautifully relates to why some movies become block-bluster and others do not using information-cascade in society. Pointing out that success of cascade depend on structure of network rather than innovation itself, he concludes that heterogeneous thresholds make society vulnerable whereas increasing heterogeneity in distributions make it less. He agrees that global cascade are rare and difficult to assess. In our paper, we take a different route. On basis on initial small window, we estimate the parameters and then project the future. We don't consider network structure or distributions.

Most of the models developed in last few years are based on Bass diffusion [1], Hazard modeling [12] or based on Branching Process [9]. Last decade has seen cascading models [3], [5], among others.

Bass [1], in his famous paper, developed the diffusion model, that established behavior in consumer durables and used it for forecasting sales at an aggregate level. Here, a part of population tries out a new product, and others imitate its behavior. This model base is at an aggregate level, hence, there is a little scope to see situations at various stages in time horizon.

In our model, we do agree with Bass [1] that every cus-

tomer who has adopted the product increases the probability of others adopting it in each time period after adoption. The 'adoption' in any new situation could be due to the influence from an advertisement, from an individual's friend circle, and so on. Richardson, et al (2002) [6] modeled influence probabilities in linear format and they take into account the marketing effort and then derive the spread. Whereas, Richardson, et al (2001) developed theory towards designing a campaign or searching for 'network value' customer.

Subsequently, Kempe (2003) [3] and Backstrom (2006) [5] addressed the computation issues related to the cascading model. A brief summary can be found in Kleinberg (2007) [4]. Acemoglu et al (2011) [2] discuss linear threshold model and provide upper bound for final count of adopters via some simulation evidence. They suggest that, innovations might spread further across networks with a smaller degree of clustering. The bounds were driven from those clustering.

Aral & Walker (2011; [12]) studied the effect of 'viral product design' on peer influence and social contagion using hazard modeling approach. Iyengar, et. al. (2011; [8]) study impact of opinion makers by usage of hazard model.

Valler(2010) [13] et. al. focuses on 'eigenvalue centrality' and states that "growth depends on the largest 'eigenvalue' of the matrix", likewise in this model too, as one would observe intuitively; future growth depends on the initial few P_i values.

Lans, et. al. [9] studied the reach of viral marketing. They formulated model based on branching process and initial seeds (targets). Their basic objective was to estimate the *total reach*. Also, they rightly pointed out the lack of models to predict *total reach* in literature. They focused more on the 'seeds' rather than 'generations', which we call 'stages'.

In our paper, we *do not* focus specifically on seeds and our model does not require information about network structure, which is often times dynamic.

1.3 Model Conceptualization and Business Scenarios

For formulation of a problem and the logical process flow, it would be a good idea to consider a real life problem. There are many business scenarios one can look at. For example, if you are starting a company page on *Facebook*, initially it will be exposed to many. Then subsequently some folks may subscribe to, or like that page. Based on this action, in the subsequent stage i.e. stage 2, some more friends of folks who have liked/subscribed to the page. This phenomena goes on, limiting to the total population in given 'community'.

Another examples is, on-line magazine subscription and its viral effect, if folks who subscribe discuss items from the magazine.

One can assume the '*stage probability of infection*' from an already infected source, at every stage. This can vary based on the seed population or by design. For example, as a company, one can create a launch for the 'seeds' that are the 'opinion leaders' of the community. Because of this, subsequent followers may be significantly higher that random 'seeds' chosen, indicating higher probability of spread. Please refer to Figure 1.

Above discussion is specific to *Facebook*, but model is generic and is applicable to different scenarios. It should be noted that, at every stage, we are assuming that the probability an individual getting infected depends on his infected connections, their infected connections and so on. This paper does not model an individual's weakness (or strength) of getting infected.

We also address the issue of diversity in population. For instance, the population can categorically be divided in the order of its influencing strength, and each category might have different probability of infecting its connections.

While doing empirical study on *Facebook, YouTube*, we will analyze the number of infections (i.e. 'likes' on the posts) initially and then estimate various stage-wise parameters. We will compare the behaviors of viral post versus the others, and evaluate the possible ways for prediction based on data of responses early on.

2. MODEL FORMULATION

To start with, let's take simplistic approach of spread (or virality). Here, we will call an activity of 'liking', 'sharing', 'going for a movie' etc. as getting 'infected'.

To describe the spread of infections at each stage, we provide a process-flow below. This we will further use for formulation of mathematical model as well as for creation of synthetic data analysis.

2.1 Model Process and Parameters - Conceptual Formulation

If an individual gets influenced or infected[2] at a particular stage, then, (s)he has scope to influence his connections with influence power p_r to his non-influenced friends. He would not have any chance further. This assumption is in similar lines of [3] & [5] and realistic. Of course, in base model below, people who are successfully influenced in rth stage are assumed to be 'observable' or visible in $(r+1)$th stage.

a Start with ***Population (Say N)***

b Each individual has circle of ***friends***[3], k

c ***Stage 1: Infections from original post***
Let p_1 be the probability of each person getting infected directly from the original post. This probability will depend on various parameters like: the brand value of page and the content of the post. If the population size (N) is large, and the community is fairly homogeneous, we can assume this value to be p_1 for all the members of the community. At latter stage, as discussed in **Introduction** we will further categorize people on basis of influence (ability to infect others)

d ***Stage 2: Infection from connections (friends)***
Target population are the folks that have not been infected in Stage1.
There is a chance of infection of an individual, only if one or more of his/her connections have already got infected.
Each of the infected friends in Stage1, have thus armed with p_2 probability / strength to infect the non-infected friends. This is valid for only Stage2. Hence, if one is non-infected at Stage1 and, "r" of her friends are infected, (s)he will get infected with probability $1-(1-p_2)^r$.

[2]influenced and infected and used in exchangeable manner
[3]Average # of friends are 200 in *Facebook*. This # may vary between 200 to 1000

e ***Stage r and beyond:*** This process goes on in the subsequent stages in a similar manner. Probability 'p_i's will get reduced as time passes / with further stages.

Please note that, so far, we have formulated the problem in such a way, that at various stages after stage 1, we have a probability with which an individual is able to infect another individual. Of course, one could formulate this problem in similar manner by considering the probabilities of an individual to get infected.

2.2 Mathematical Formulation

Let S_i be a random variable that indicates the # of folks getting infected in the i^{th} stage.

Key here is to study the behavior of random variables 'S_i's. To start with we will look at 1^{st} order moments[4] of these random variables. These in turn can be used from the sample to estimate the base parameters of the model.

a **Model Parameter :**

(a) p_1 : The probability of getting infected in stage-1 (i.e. from the original post)
(b) p_i: The probability by which an individual can be infected in i_{th} stage by an already infected individual. ($i > 1$)
(c) N: The population size (This may be related to number of subscribers of a particular page in *Facebook* scenario.)
(d) K_i: Number of connections of i^{th} individual
(e) K : Median value of K_i's

b **Derived Parameters:**

(a) f_i: Probability of getting infected in stage-i (only)
(b) g_i: Probability of getting infected in stage i or before
(c) $\lambda_i = Kf_i$

2.2.1 Stage-1 infections

Since $p_1 = f_1$ is the probability of an individual getting infected in Stage-1, *Expected number of infections (S_1)* would be:

$$E(S_1) = Np_1 = Nf_1 \tag{1}$$

2.2.2 Stage-2 infections

Let $I_1, I_2, .., I_N$ be the individuals in population. Let $K_1, K_2, ..., K_N$ be number of their respective friends. Let there be 'r' friends of I_l that have got infected in stage 1, where $1 \leq l \leq N$. Let R be a random variable that represents number of infected friends. Then

$$P(R = r) = \binom{K_l}{r} g_1^r (1 - g_1)^{n-r} \tag{2}$$

Let E_{1_l} be the event that I_l is not infected in stage-1, then $P(E_{1_l}) = (1 - f_1)$. Let A be the event that I_l gets infected in stage 2 and not in stage 1. As defined above, this event has a probability of f_2. Hence,

$$\begin{aligned} f_2 &= \sum_{r=1}^{K_l} P(A|R = r)P(R = r)P(E_{1_l}) \\ &= \sum_{r=1}^{K_l} P(A|R = r)\binom{K_l}{r} g_1^r (1 - g_1)^{n-r}(1 - g_1) \\ &= \sum_{r=1}^{K_l} P(A|R = r)\binom{K_l}{r} g_1^r (1 - g_1)^{n-r+1} \end{aligned}$$

Assuming that K_i's are large, and probabilities g_i's are small, we will use Poisson approximation. Let the median value of K_is be K. Then,

$$\begin{aligned} f_2 &= \sum_{r=1}^{K_l} [1 - (1 - p_2)^r]\binom{K_l}{r} g_1^r (1 - g_1)^{n-r+1} \\ &= \sum_{r=1}^{K_l} [1 - (1 - p_2)^r]\frac{e^{-\lambda_1}\lambda^r}{r!}(1 - g_1) \\ &= \sum_{r=1}^{K_l} P(B_{(r,p_2)} \geq 1)P(P_\lambda = r)(1 - g_1), \end{aligned}$$

where, $B(r, p_2)$ indicates a Binomial random variable with probability p_2 and r trials, P_{λ_1} indicates a Poisson distribution with parameter λ_1. Continuing forward,

$$\begin{aligned} \frac{f_2}{(1 - g_1)} &= \sum_{r=1}^{K_l} [1 - (1 - p_2)^r]\frac{e^{-\lambda_1}{\lambda_1}^r}{r!} \\ &= \sum_{r=1}^{K_l} \frac{e^{-\lambda_1}{\lambda_1}^r}{r!} - \frac{e^{-\lambda_1}{\lambda_1}^r(1 - p_2)^r}{r!} \\ &= \sum_{r=1}^{K_l} \frac{e^{-\lambda_1}{\lambda_1}^r}{r!} - \frac{e^{-\lambda_1}\left(\lambda_1(1 - p_2)\right)^r}{r!} \\ &= \sum_{r=1}^{K_l} \frac{e^{-\lambda_1}{\lambda_1}^r}{r!} - \frac{e^{-(\lambda_1(1-p_2))}\left(\lambda_1(1 - p_2)\right)^r}{r!} e^{-\lambda_1 p_2} \end{aligned}$$

For large K_l, $\sum_{r=0}^{K_l} \frac{e^{-\lambda_1}{\lambda_1}^r}{r!} \approx 1$ leading to,

$$\begin{aligned} \frac{f_2}{(1 - g_1)} &= \left(1 - e^{-\lambda_1}\right) - \left(1 - e^{-\lambda_1(1-p_2)}\right) e^{-\lambda_1 p_2} \\ f_2 &= \left(1 - e^{-\lambda_1 p_2}\right)(1 - g_1) \end{aligned}$$

Hence, the expected number of Stage-2 infections (S_2) is,

$$E(S_2) = Nf_2 \tag{3}$$

The above can be viewed as conditioning on only on non-infected population, and computing the expected value. Hence, it is net number of stage 2 infections alone. Now, we can move to various stages in similar manner.

2.2.3 Stage-r infections

On similar lines, based on the previous mathematical formulation, we arrive at a simplified recurrence relation determining the probability of getting infected in stage-r alone i.e. 'f_r'

$$f_r = \left(1 - e^{-\lambda_{r-1}p_r}\right)(1 - g_{r-1}) \tag{4}$$

[4]equating sample moments to actuals from sample and solving is called moment estimates

Hence, the expected number of Stage-r infections (S_r) is,

$$E(S_2) = Nf_r \tag{5}$$

Where, the derived parameters are:

$$\begin{aligned} f_1 &= g_1 = p_1 \text{ and } \lambda_1 = Kp_1 \\ \text{For all } r > 1, & \\ f_r &= \left(1 - e^{-\lambda_{r-1} p_r}\right)(1 - g_{r-1}) \\ g_r &= f_r + g_{r-1} \\ \lambda_r &= Kf_r \end{aligned}$$

The above recurrence equations provide us way to estimate parameters stage-wise. In *'Model Evaluation'* section we provide examples with its usage.

3. CASE OF HOMOGENEOUS INFLUENCE

Let i represent stages for $i = 1, 2, ...$ Probability of an individual infected in $i-1$th stage infecting non-infected individual is p_i.
Then, if N is size of population (community), and, S_r is rth stage infections and K be median size of circle of individual connections, then - following are infection equation, which has recurrence relation and would be used for estimating parameters via moment estimates in section5, of the paper.

$$E(S_r) = Nf_r \tag{6}$$

Where, the derived parameters are:

$$f_1 = g_1 = p_1 \text{ and } \lambda_1 = Kp_1 \tag{7}$$

For all $r > 1$,

$$\begin{aligned} f_r &= \left(1 - e^{-\lambda_{r-1} p_r}\right)(1 - g_{r-1}) \\ g_r &= f_r + g_{r-1} \\ \lambda_r &= Kf_r \end{aligned}$$

4. CASE OF HETEROGENEITY OF INFLUENCE'S

There may be various clusters in a population with different rate of spread etc. We demonstrate how one can generalize above model to such a scenario. The process is simple.

Let there be m various categories of influences, p_{ij} representing similar probability for ith stage from jth group. Now, we will have K_js, N_j's, where $\sum K_j = K$ and $\sum N_j = N$. Then

$$\begin{aligned} f_r &= [\prod_{i=1}^{m}(1 - e^{-\lambda_{(r-1)i}}) \\ &- \prod_{i=1}^{m}(1 - e^{-\lambda_{(r-1)i}(1-p_{ri})})e^{-\lambda_{(r-1)i} p_{ri}}](1 - g_{r-1}) \end{aligned}$$

If S_r is the number of infections in stage-r ,

$$E(S_r) = Nf_r$$

Where, the derived parameters are:

$$\begin{aligned} f_1 &= g_1 = p_1 \text{ and } \lambda_{1j} = K_{.j} f_1 \\ \text{For all } r > 1, & \\ f_r &= [\prod_{i=1}^{m}(1 - e^{-\lambda_{(r-1)i}}) \\ &- \prod_{i=1}^{m}(1 - e^{-\lambda_{(r-1)i}(1-p_{ri})})e^{-\lambda_{(r-1)i} p_{ri}}] \\ &\times(1 - g_{r-1}) \\ g_r &= f_r + g_{r-1} \\ \lambda_{rj} &= K_{.j} f_r \end{aligned}$$

Based on observations we can estimate parameters p_i's, which could be used for "similar" posts in future.

Please note in above discussions and derivations, in Stage-i, we have taken the probabilities of individuals getting infected from their connections to be same i.e. p_i. We will elaborate this model further to include the diversity of getting infected in population. For instance, individuals who are opinion makers in society are likely to have substantial influence i.e. they infect their connections with a higher probability as compared to others. This will be addressed in next section.

It is to be noted that the parameter p_is can vary substantially from individual to individual. Later, this problem can address by creating two sets within population where one is follower another in influential etc.

4.1 Heterogeneity of Influencer's

In the previous section, we assumed that every individual has an equal ability to influence his/her friends. In the real world, this is not true. While designing viral campaign one would need to consider the individual probabilities related to influence. Here, in this paper, we are looking at various stages within the population and how spread is likely to grow. At times, the whole set of population is a combination of various clusters, which are homogeneous within. For example, some clusters may be very closely knit where as some may not be at the same level. Influence in one cluster may spread faster than the other. Hence, we attempt to provide extension of previous section, on similar lines for this kind of issue. The techniques are similar.

Let there be m type of people in the population. As stated earlier, this is an extension of methodologies discussed in previous section. Thus,

$$N = N_1 + N_2 + ... + N_m$$

(N_i is the size of population of i^{th} kind.)

4.1.1 'm' type of people in the population

For an individual I_l, we assume K_{l_i} of his friends belong to population of type-i

a **Model Parameter:**

(a) p_1 : The probability of getting infected in stage-1 (i.e. from the original post)

(b) p_{ij}: The probability by which an individual can be infected in i_{th} stage by an individual of population type j. ($i > 1$ and $1 \leq j \leq m$)

(c) N: The total population size

(d) K_{ij}: Number of connections of i^{th} individual of population type j

(e) $K_{.j}$: Median value of K_{ij}'s, where i varies.

b **Derived Parameters:**

(a) f_i: Probability of getting infected in stage-i (only)

(b) g_i: Probability of getting infected in stage1 or stage 2 or ... stage i-1 or stage i

(c) $\lambda_{ij} = K_{.j} f_i$

4.1.2 Stage-1 infections

As mentioned before, Stage-1 infections spread due to the post itself and there is no role of 'friends'. Thus, the results would be similar to the case when only single type of people are in the population.
The probability of getting infected in stage-1 is f_1, and S_1 is the number of infections in stage-1, then the expected value of S_1 is:

$$E(S_1) = Nf_1 \tag{8}$$

4.1.3 Stage-2 infections

Now we evaluate the probability f_2 of an individual getting infected in stage-2. This would depend on connections which are already infected and their respective population categories. Similar methodology and approximations as described in the previous section have been used below:

$$\begin{aligned}
f_2 &= \sum_{r_1=1,r_2=1\ldots r_m=1}^{K_{l1},K_{l2},\ldots,K_{lm}} [1 - \prod_{i=1}^{m}(1-p_{2i})^{r_i}] \\
&\quad \times (\prod_{j=1}^{m} \binom{K_{lj}}{r_j} f_1^{r_j}(1-f_1)^{K_{l_j}-r_j})(1-g_1) \\
\frac{f_2}{(1-g_1)} &= \sum_{r_1=1,r_2=1\ldots r_m=1}^{K_{l1},K_{l2},\ldots,K_{lm}} [1 - \prod_{i=1}^{m}(1-p_{2i})^{r_i}] \prod_{j=1}^{m} \frac{e^{-\lambda_{1j}}\lambda_{1j}^{r_j}}{r_j!} \\
&= \sum_{r_1=1,r_2=1\ldots r_m=1}^{K_{l1},K_{l2},\ldots,K_{lm}} (\prod_{j=1}^{m} \frac{e^{-\lambda_{1j}}\lambda_{1j}^{r_j}}{r_j!} \\
&\quad - \prod_{j=1}^{m} \frac{e^{-\lambda_{1j}(1-p_{2j})}(\lambda_{1j}(1-p_{2j}))^{r_j}}{r_j!} e^{-\lambda_{1j}p_{2j}})
\end{aligned}$$

Assuming that K_{l_i}s are large,

$$\begin{aligned}
\frac{f_2}{(1-g_1)} &= \prod_{i=1}^{m}(1-e^{-\lambda_{1i}}) - \prod_{i=1}^{m}(1-e^{-\lambda_{1i}(1-p_{2i})})e^{-\lambda_{1i}p_{2i}} \\
f_2 &= [\prod_{i=1}^{m}(1-e^{-\lambda_{1i}}) \\
&\quad - \prod_{i=1}^{m}(1-e^{-\lambda_{1i}(1-p_{2i})})e^{-\lambda_{1i}p_{2i}}](1-g_1)
\end{aligned}$$

If S_2 is the number of infections in stage-2,

$$E(S_2) = Nf_2 \tag{9}$$

4.1.4 Stage-r infections $r > 1$

$$\begin{aligned}
f_r &= [\prod_{i=1}^{m}(1-e^{-\lambda_{(r-1)i}}) \\
&\quad - \prod_{i=1}^{m}(1-e^{-\lambda_{(r-1)i}(1-p_{ri})})e^{-\lambda_{(r-1)i}p_{ri}}](1-g_{r-1})
\end{aligned}$$

If S_r is the number of infections in stage-r ,

$$E(S_r) = Nf_r \tag{10}$$

Where, the derived parameters are:

$$f_1 = g_1 = p_1 \text{ and } \lambda_{1j} = K_{.j} f_1$$

For all $r > 1$,

$$\begin{aligned}
f_r &= [\prod_{i=1}^{m}(1-e^{-\lambda_{(r-1)i}}) \\
&\quad - \prod_{i=1}^{m}(1-e^{-\lambda_{(r-1)i}(1-p_{ri})})e^{-\lambda_{(r-1)i}p_{ri}}] \\
&\quad \times (1-g_{r-1}) \\
g_r &= f_r + g_{r-1} \\
\lambda_{rj} &= K_{.j} f_r
\end{aligned}$$

Based on observations we can estimate parameters p_{ij}s, which could be used for "similar" posts.
One can derive a simplified version for Bi-heterogeneous case.

5. MODEL EVALUATION

The mathematical model described in this paper, can be applied to variety of social-media data and it can be used to predict future trends. The objective of this section is to demonstrate possible usage and applications. Once the parameters of the model are estimated and model fitments are done, one can observe the trends and assess their future values using various methodologies available in literature, including but not limiting to, regression and time series.

These *subjective* decisions are needed to be assessed by the user as they may depend on the data being investigated and domain expertise.

- Population Size N: For example, for *NY Times* post on *Facebook* we can easily take N as # of page likes and then asses virality of particular post. Where as for *YouTube* we took a multiple of total range. The multiple being 2 to 8. One would have to use subjective assessment to judge the Target population of interest (N).
- Median # of connections K: For *Facebook* we took this as 250, where as for *YouTube* we took it as 400.
- Estimation of p_i's from observation window: We have used simple moment estimators.
- Predicting trend of p_i's: As this is methodology paper, in examples below, after estimating p_i's in initial observed window of 10%, we have taken easier approach. One can use *time-series* or other known methods on p_i's and then use predicted p_i's (and then other parameters) to project the cumulative infection curve.

There are various multitudes of social networks available, which provides different types of interactions i.e. One to One, One to Many, Many to Many. Here is a list of applicable categories (but not limited to) where the model could be applied with the right data :

a Blogs and Forums

b Social media sharing services like Video (YouTube, Vimeo), Photos (Flickr, Picasa, Instagram), Audio (Pandora, Lastfm), Bookmarks (Stumble Upon, Delicious).

c Social Networks like Full Network (*Facebook*, Google+), Microblogging (Twitter , Plurk), Professional (LinkedIn, Xing).

d Social News (Digg, Reddit)

e Location Based Networks (FourSquare, Latitude)

As stated earlier, our main objective in this section is to demonstrate practical application of the described models. In the following sub sections we are going to attempt to evaluate the theory by applying it for data gathered from two unique sources.

a In section 5.2, we are going to discuss about four New York Times on its *Facebook* page.

b In section 5.3, a YouTube Viral phenomenon - "Gangnam Style" Music Video.

c In section 5.4. we explore more examples, taken from *Facebook* and YouTube.

5.1 Implementation Methodology

It is clear that the equation (7) would be used along with $E(S_r) = Nf_r$. This forms as 'moment'[5] estimates. Of course, one could make use of some efficient methodologies such as maximum likelihood, etc among others. Once we have set observation window, which is typically 10% to 20% of overall data, we solve for the related parameters using the recurrence equations. Then, based on the trend of some stable function of p_i's that are observed, we predict the future. The details are provided in subsequent examples below.

5.2 NY Times - *Facebook*

In this section we evaluate mathematical model discussed in this paper using "*Facebook*" posts, posted by "The New York Times"[6]. This evaluation introspects 'likes' l for four unique posts over the first 12 hours of posting, sampled at interval δt which is about 10 minutes.

Assumptions

a Its assumed that every individual on the network has equal infection capability; hence we are using **homogeneous** equation; see Equation: **4**.

b An individual, 'liking' a post is in effect considered **'infected'**.

c The model requires us to define **stages**; for simplicity of evaluation we have defined each stage as the data sampling interval δt, which averages about 10 minutes.

d Population of a network is considered to be the subscribers of that page being evaluated, in this case, # of Likes on New York Times page.

e Average number of friends is considered as **200**[7].

f Initially, we assume $p_j = 0.0002$,where $j \geq 2$ and we solve for the model parameters using the equations described in the Section: **3**[8].

Model Parameters

a $N = 2,504,817$; see Assumption: 4.

b $k = 200$; see Assumption: 5.

c $\lambda_i = kf_i$; λ_i defines the number of individuals infected at stage i.

d $g_i = f_i + g_{i-1}$; g_i is the probability of getting infected in any of the previous stages, 1 to $i-1$.

e $f_i = (1 - e^{-p_i \lambda_{i-1}})(1 - g_{i-1})$; f_i is the probability of getting infected in stage i.

f $f_1 = g_1 = p_1$ and $\lambda_1 = kp_1$; Refer Equation: 7.

g $p_1 = N/\delta l_1$, From Equation **6**; l_i is the number of people infected in Stage 1.

Data Parameters

a $\delta l = l_i - l_{i-1}$; δl is the number of 'likes' recorded for the post in δt time, in other words, δl defines the number of infections that occurred at that **stage** in time δt : see Assumption: 2 & 3.

b $e_i = \delta l - f_i N$; e_i is the error parameter, which is used to re-estimate p_i values, this is based on Equation: **5**.

δt	Stage	Time Stamp	l	δl	p_i	λi	G_i	F_i	Err e_i
	1	09-08-2012 01:59	1577	1577	**6E-04**	0.126	6E-04	6E-04	**0**
0.007639	2	09-08-2012 02:10	3047	1470	**0.005**	0.117	0.001	6E-04	**7E-06**
0.006944	3	09-08-2012 02:20	4082	1035	**0.004**	0.083	0.002	4E-04	**1E-07**
0.00625	4	09-08-2012 02:29	4865	783	**0.004**	0.063	0.002	3E-04	**2E-07**
0.007639	5	09-08-2012 02:40	5548	683	**0.004**	0.055	0.002	3E-04	**5E-07**

Figure 2: Data Snapshot Table - Post A from New York Times

Evaluation

a For each post at each stage λ_j, g_j,f_j ande_j, where $j \geq 2$, are evaluated using recurrence relation described in Equation: **4** based on an estimated $p_j = 0.0002$; See Assumption: **6**.

b This results in $|e_j| > 0$.

c In order to estimate appropriate values of p_j, we adjust p_i and re-evaluate λ_j, g_j,f_j ande_j, where $j \geq 2$, using the simplified recurrence relation such that $e_j \to 0$, See Assumption: **6**.

[5]Equating sample expected values to equations

[6]New York Times on *Facebook*: www.facebook.com/nytimes

[7]http://arxiv.org/abs/1111.4503 states median number of friends on *Facebook* is around 200.

[8]We solve for the equations by re-estimate all values of p_i, such that $e_i \to 0$; see Data Parameter: **2**

Now we have a list of p_i values, and its observed that there is some jitter in values of δt . Note that this jitter is caused due to the sampling interval in Facebook not being uniform.

5.2.1 *Predicting 'Like' Trends*

In this section we are going to attempt to predict 'Like' trends by observing part of the values evaluated in the previous section. Here we observe a part of the data and the evaluated from p_1 to p_r for time t_r and then try to predict values of l for all δt's.

Here Stage 1 to Stage r is considered as the **observation window** for predicting future cumulative likes l.

One could use some model of *Time Series analysis* for evaluating the trends of p_i's but for simplicity here we are going to use *simple linear extrapolation* of cumulative p_i's (See Equation: **11**) to evaluate δl for each stage.

$$H(t) = \sum_{0}^{r} p_i \tag{11}$$

Algorithm

a Get all values of p_i, where $0 \leq i \leq r$ and plot cumulative values of p_i (See Equation: **11**) against t.

b Curve fit using a Linear Equation on the plot obtained.

c Extrapolate estimated cumulative p_i based on Linear Equation.

d Evaluate λ'_i, g'_i, f'_i using the recurrence relations explained in Section **3**.

e Calculate $\delta l' = N, f'_i$, iterate steps **4** & **5** for each evaluated p_i.

f Plot $\delta l, \delta l'$ against t.

Flowchart

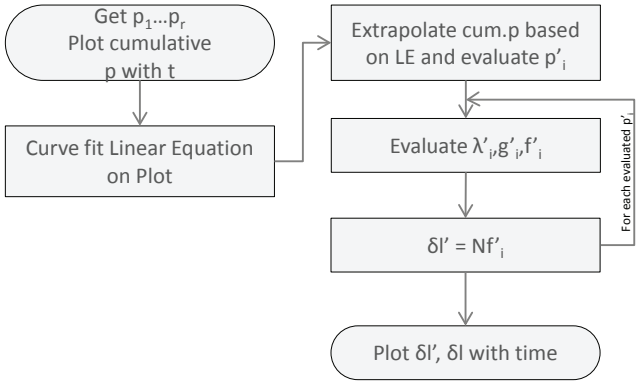


Figure 3: Flowchart to predict future $\delta l'$ and plot them against actual δl

Parameters

a $p'_i = \sum_0^n p_i - \sum_0^{n-1} p_i$; Here p'_i are the values estimated from the curve fit Linear Equation, based on Stage 2 from Algorithm in Flowchart 3.

b $\lambda'_i = kf'_i$

c $g'_i = f'_i + g'_{i-1}$; Expresses the estimated probability of getting infected in any of the previous stages 1 to $i-1$.

d $f'_i = (1 - e^{-p'_i \lambda'_{i-1}})(1 - g'_{i-1})$; Defines the estimated probability of getting infected in stage i.

e $\delta l' = Nf'_i$; Defines the number of estimated individuals infected at stage i, this is based on Equation: **6**.

f $N = 2,504,817$ and $k = 200$ See Assumption **4** and **5**.

Here, we use a fraction of data points of our collected data to draw a cumulative of p_i from p_1 to p_r curve. On this curve, simple curve fitting technique is used to express the curve using a linear equation.

From the linear equation,we extrapolate expected values of cumulative p_i's for the observation window. And evaluate $\delta l'$ as per the flowchart described in Figure: **3**

We plot cumulative l(*green*) with l'(*red*). Figure: **4** is an example of this plot, in section 5.4, we have several examples of these plots.

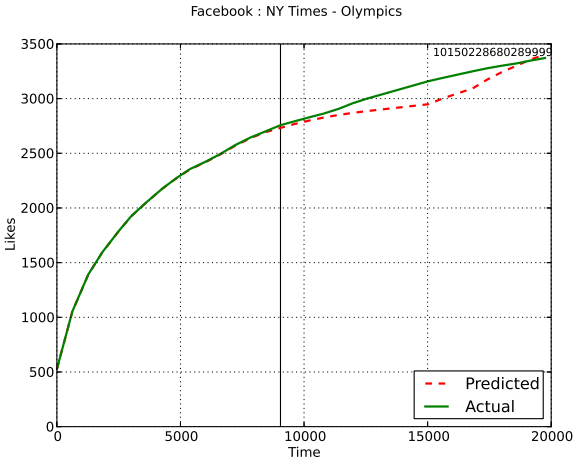


Figure 4: Predicted likes(*red*) with Actual likes(*green*) for post A, with observation window marked by vertical line.

This example provides evidence of applicability of the model described in this paper. From this experiment, we can conclude, with the simplistic approach of extrapolation of cumulative p_i can be used to predict future trends of 'likes' with sufficient accuracy. Better techniques along with longer observation period can result in higher accuracy for predicted 'likes'.

5.3 'Gangnam' Style

"Gangnam Style" is a single by South Korean pop artist "PSY".The song and it's music video went viral in August 2012 currently grossing over 1 billion views, we have extracted data and evaluated using our technique.

5.3.1 *Data Extraction*

YouTube shares certain statistics regarding the video in chart form, that is dynamically generated using the Google Charts API, from this URL, we are able to gather the following information.

- **Begin Date** :18 July 2012; The date from which the monitoring is available.
- **End Date** : 07 January 2013; The date to which the monitoring is available.
- **Data Range** : $[0-1,385,470,291]$ Values in set $[0,Range]$, in Number of Views.
- **Data Points** : A set of 100 data points d_i in percentage of Data Range.

From this, we evaluate 100 (length of data points) number of equal divisions of time between the 'Begin' and 'End' date, to this date we associate $Range*d_i$ number of views.

δt	Stage	Time Stamp	I (views)	δI (views)	p_i	λ_i	g_i	f_i	Err e_i
	1	17-07-2012 11:02	883222	883222	**1E-03**	0.4	1E-03	1E-03	**0**
1.23	2	18-07-2012 16:33	1766444	883222	**0.003**	0.4	0.002	1E-03	**1E-05**
2.46	3	21-07-2012 03:36	2649666	883222	**0.003**	0.4	0.003	1E-03	**4E-08**
2.46	4	23-07-2012 14:38	3532888	883222	**0.003**	0.4	0.004	1E-03	**2E-08**
1.23	5	24-07-2012 20:09	4416110	883222	**0.003**	0.4	0.005	1E-03	**-0**

Figure 5: Data Snapshot Table - Gangnam Style

Key Assumptions

a N, the population size $= 3.5 * Range$, i.e. $4,849,146,018$.

b $k = 400$ average number of friends in the network[9].

c Viewership may not be completely unique, this data may include multiple views by the same individual.

With these data and assumptions, we plot Actual Views vs Predicted Views in Figure: 6 based on the Algorithm described in flowchart Figure: 3

[9]Here we assume average number of friends to be $2x$ *Facebook*, given that you don't need to be logged in to interact with the video, and sharing of YouTube videos can happen via other social networks.

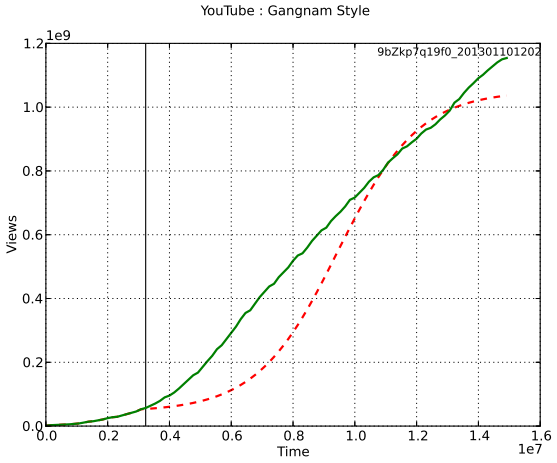


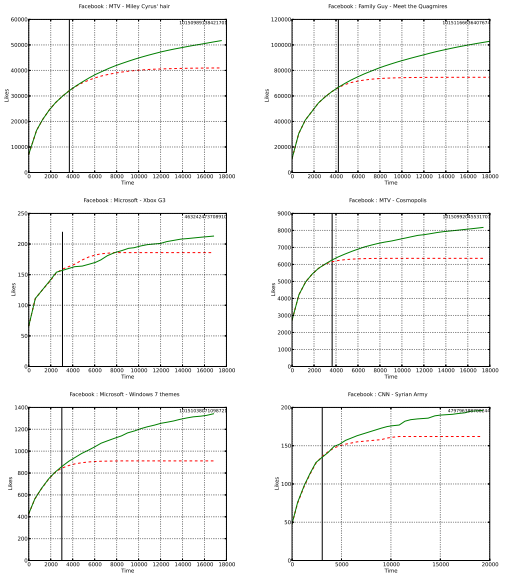
Figure 6: Predicted Views(*red*) **with Actual Views**(*green*) **for "Gangnam Style" with observation window marked**

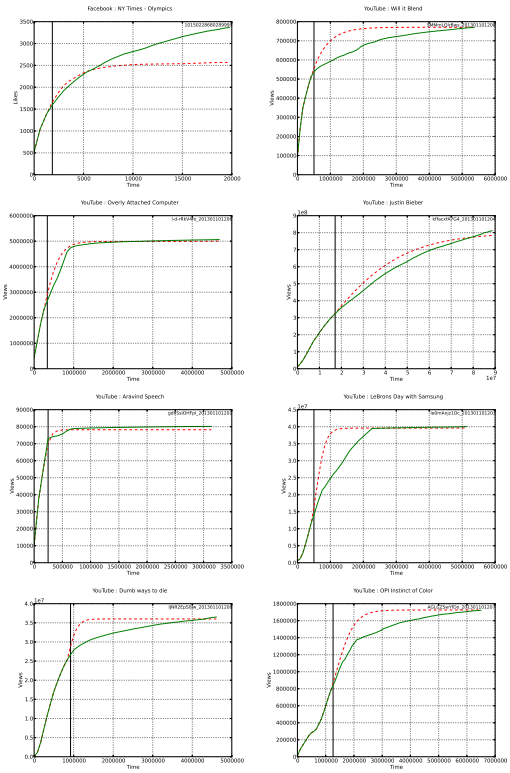
Section 5.4 presents our evaluations on a few 'Facebook' and 'YouTube' posts.

5.4 Additional Emperical Evaluations

Accuracy of prediction of *Facebook* posts are lower, due to sampling interval issues, discussed in section 5.2. Also note that the community size of each page of N, is at best a reasonable estimate.

Here,
- Actual line is represented as *green*, continous lines.
- Prediction line is represented as *red*, dashed lines.
- X-Axis is the elapsed time in seconds(s).
- Y-Axis is the 'Views' in *YouTube* & 'Likes' for *Facebook*.





Even with loose assumptions of N, which we assumed to be the 100% of the range obtained from the data, and data point approximations, we were able to accurately model predicted viewership within an acceptable error range. We may be able to improve the accuracy, if we gain access to more accurate values of N, and data points only include unique viewership.

6. ACKNOWLEDGMENT

We are thankful to Dr Ritwik Sinha for providing much needed initial references and pointers to earlier work. We also would like to thank Ramesh Srinivasaraghavan for helping us out with the logical flow of the text.

7. REFERENCES

[1] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, January 1969.

[2] Asuman Ozdaglar Daron Acemoglu and Ercan Yildiz. Diffusion of innovations on social networks. *IEEE Conference on Decision and ControI, Orlando, FL*, December 2011.

[3] EÂt' va Tardos David Kempe, Jon Kleinberg. Maximizing the spread of infuence through a social network. *SIGKDD*, 2003.

[4] Jon Kleinberg. Cascading behavior in networks: Algoritmic and economic issues. *Algorithmic Game Theory, edited by Boam Nisan et al; Cambridge University Press*, 2007.

[5] Jon Kleinberg Lars Backstrom, Dan Huttenlocher. Group formation in large social networks: Membership, growth, and evolution. *KDD*, 2006.

[6] Pedro Domingos Matthew Richardoson. Mining the network value customer. *Seventh International Conferance on Knowladge Discovery and Data Mining*, 2001.

[7] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[8] Thomas W. Valente Raghuram Iyengar, Christophe Van den Bulte. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, March 2011.

[9] Jehoshua Eliashberg Berend Wierenga Ralf Van der Lans, Gerrit van Bruggen. A viral branching model for predicting the spread of electronic word of mouth. *Marketing Science*, 29(2):348–365, March 2010.

[10] Everett Rogers. Diffusion of innovations, 5th edition. *ISBN-10: 0743222091*, 2003.

[11] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 1995.

[12] D. Walker S. Aral. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1629, September 2011.

[13] Nicholas Valler, B. Aditya Prakash, Hanghang Tong, Michalis Faloutsos, and Christos Faloutsos. Epidemic spread in mobile ad hoc networks: Determining the tipping point. In Jordi Domingo-Pascual, Pietro Manzoni, Sergio Palazzo, Ana Pont, and Caterina M. Scoglio, editors, *NETWORKING 2011 - 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011, Proceedings, Part I*, volume 6640 of *Lecture Notes in Computer Science*, pages 266–280. Springer, 2011.

[14] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 9, pp. 5766-5771*, Apr. 30, 2002.