Insights into Internet Memes

Christian Bauckhage
Fraunhofer IAIS
Bonn, Germany

Abstract

Internet memes are phenomena that rapidly gain popularity or notoriety on the Internet. Often, modifications or spoofs add to the profile of the original idea thus turning it into a phenomenon that transgresses social and cultural boundaries. It is commonly assumed that Internet memes spread virally but scientific evidence as to this assumption is scarce. In this paper, we address this issue and investigate the epidemic dynamics of 150 famous Internet memes. Our analysis is based on time series data that were collected from Google Insights, Delicious, Digg, and StumbleUpon. We find that differential equation models from mathematical epidemiology as well as simple log-normal distributions give a good account of the growth and decline of memes. We discuss the role of log-normal distributions in modeling Internet phenomena and touch on practical implications of our findings.

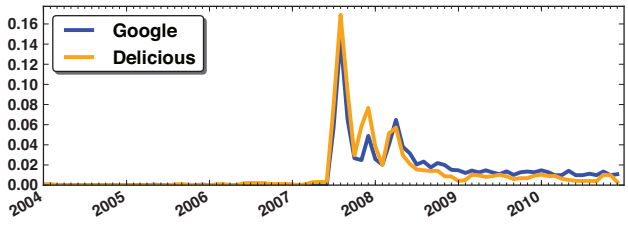
Introduction

The term *Internet meme* refers to the phenomenon of content or concepts that spread rapidly among Internet users. It alludes to a theory by Dawkin (1976) who postulates *memes* as a cultural analogon of genes in order to explain how rumors, catch-phrases, melodies, or fashion trends replicate through a population. Whether or not memes do really exist is heatedly debated and we do not intend to join that discourse. Instead, our discussion in this paper focuses on observable characteristics of Internet memes that resemble those of viral spread and epidemic outbreaks.

In their basic form, Internet memes propagate among people by means of email, instant messaging, forums, blogs, or social networking sites. Content-wise, they usually consist of offbeat news, websites, catch phrases, images, or video clips (see Figs. 1 and 2). Put in simple terms, Internet memes are inside jokes or pieces of hip underground knowledge, that many people are in on.

Internet memes typically evolve through commentary, imitations, or parodies, or even through related news in other media. Most Internet memes spread rapidly; some were observed to go in and out of popularity in just a matter of days. Memes are spread in a voluntary, peer to peer fashion, rather than in a compulsory manner. Their proliferation through social communities does not follow predetermined paths and usually defies efforts to control it.

(a) instances of the "chocolate rain" meme



(b) two time series (retrieved from Google Insights and Delicious) reflecting the rise and decline in popularity of this Internet meme

Figure 1: Example of an Internet meme. On April 22, 2007, singer Tay Zonday (upper left) posted a home-made music video on YouTube. The catchy tune and somewhat awkward performance apparently appealed to a large audience: as of this writing, the "chocolate rain" video has been viewed more than 57,000,000 times and was frequently spoofed and re-contextualized.

As of late, the phenomenon of Internet memes has itself attracted growing public interest. Popular web sites such as *knowyourmeme.com*, *memedump.com*, or *memebase.com* view them as a form of art and provide accounts of the origin and evolution of famous memes.

Professionals in public relations and advertising, too, have embraced Internet memes. In viral marketing, there are examples of memes that were purposely designed to create publicity for products or services. Finally, political campaigning increasingly attempts to create Internet memes to shape opinion. They are supposed to create an image of trendiness but often interest in the content is for purposes of trivia or frivolity rather than for information.

Given the public interest in Internet memes, it is sobering to see that many aspects of the phenomenon are still poorly

Figure 2: Instances of the "o rly?" meme. It is disputed wether it originates from somethingawful.com or 4chan.org.

understood. Knowledge as to the dynamics of meme spread is still more qualitative than quantitative and conclusions appear to be drawn from episodic rather than from analytic evidence. As a consequence, models that would allow for assessing the success of a viral campaign in its early stages or for predicting the longevity or peak circulation of a rising meme remain elusive to this date.

At the same time, scientific interest in the topic is noticeably increasing as more and more researchers in web data mining and social network analysis are beginning to study Internet memes. With the work reported here, we want to contribute to these efforts. In particular, we are interested in the temporal dynamics of Internet memes and study models for predicting the evolution of their popularity. Our analysis is based on time series that were collected from *Google Insights* as well as from three social bookmarking services, namely *delicious.com*, *digg.com*, and *stumbleupon.com*.

We report on characteristic similarities and differences among the data from the different sources. Our analysis reveals that the user communities of the considered services appear to have different interests and show behaviors that reflect different aspects of Internet memes.

Moreover, we study the use of models from mathematical epidemiology and log-normal distributions in modeling the temporal dynamics of Internet memes. We observe that both provide accurate accounts for our data and we discuss our findings with respect to the link structure of social graphs centered around Internet memes. Finally, we apply our models in an attempt to predict the future evolution of various Internet phenomena.

Our presentation proceeds as follows: next, we review related work and discuss it with respect to the approaches followed in this paper. Then, we introduce the time series data that forms the empirical basis for our study. We analyze similarities and differences among the data from different sources and then introduce mathematical models of outbreak data and apply them to characterize Internet memes and their evolution. We conclude by summarizing our results.

Related Work

Work related to Internet memes and their dynamics is found in the areas of web intelligence and social network analysis. Several authors attempt to identify influential members in a community so as to contain the spread of misinformation or rumors (Budak, Agrawal, and Abbadi 2010; Shah and Zaman 2009). Others propose models of how events disseminate through online communities and use these to track memes through specific social media (Adar and Adamic 2005; Lin et al. 2010) or to investigate the interplay between social and traditional media (Leskovec, Backstrom, and Kleinberg 2009). Although these contributions touch on outbreak analysis and peak intensity modeling, they are not particularly concerned with time series analysis and do not develop tools for forecasting the future development of a rampant meme.

Outbreak analysis for trend prediction, however, is an active area of research in epidemic modeling (Britton 2010). Moreover, similarities in the spread of diseases and rumors have been noted for long (Dietz 1967) and are thought to be an emergent property of the scale-free nature of social- or communication networks (Keeling and Eames 2005; Lloyd and May 2001; Pastor-Satorras and Vespignani 2001). This has led to several applications of traditional epidemic modeling in the context of web technologies. Examples include mechanisms to curtail the activity of computer viruses (Bloem, Alpcan, and Basar 2009) or attempts to infer social relations from observations of information propagation among individuals (Myers and Leskovec 2010).

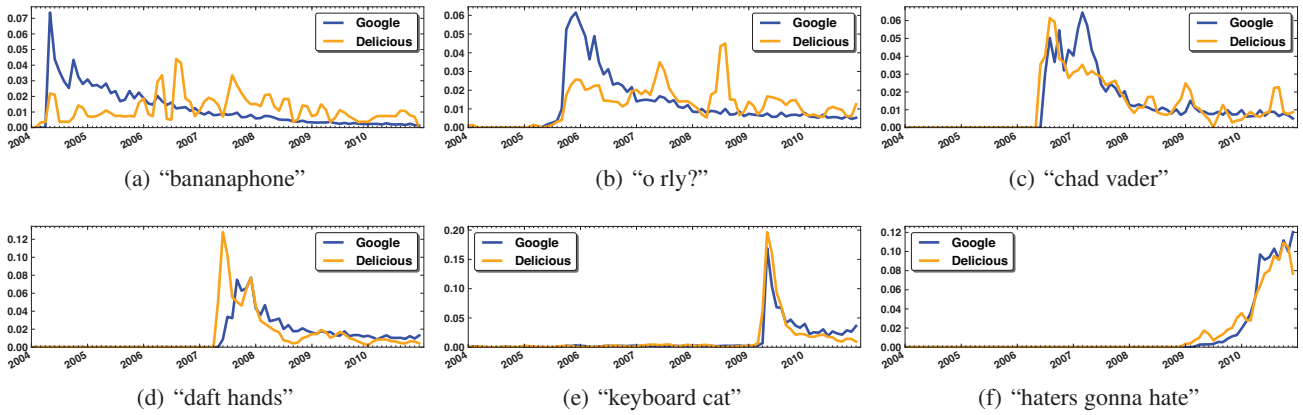
Work more closely related to what is reported here is due to Yang and Lescovec (2011) and Kubo et al. (2007). The former cluster time series obtained from a micro blogging service in order to predict future interest in a topic. The latter investigate the temporal evolution of content in bulletin boards and report that a simple stochastic compartment model gives a good account of the process. Concerned with Internet memes, we could not corroborate these findings. While the time series analyzed by Kubo et al. quickly tail off, temporal distributions that characterize meme popularity are, in their vast majority, heavily skewed and long-tailed. Our results reported below indicate that more elaborate compartment models and log-normal distributions capture this behavior more accurately. Log-normal distributions are known to accurately model a wide range of long-tail phenomena (Limpert, Stahel, and Abbt 2001) including Internet measurements such as communication times or the growth of the web graph (Downey 2005; Mitzenmacher 2004). They were also found to characterize frequency distributions of bookmarks or recommendations in bookmarking or recommender services (Wu and Huberman 2007; Lescovec, Adamic, and Huberman 2007) as well as to represent the response dynamics of social systems to sudden exogenous events (Crane and Sornette 2008).

Stochastic compartment models and log-normal distributions will be discussed again in more detail in a later section.

Data Collection and Preprocessing

In this paper, we analyze the characteristics of a collection of 150 Internet memes. Table 1 lists a subset of 120 of these memes; the remaining 30 examples are part of our analysis but we avoid mentioning them because they are memes that

- are of repugnant, offensive or highly controversial nature (this includes so called gross out memes which often center around bizarre sexual practices; we also ignore memes centered around acts of violence or torture (of animals) as well as so called screamer memes that are intended to invoke a state of horror or nervous shock in their audience)



(a) "bananaphone" (b) "o rly?" (c) "chad vader"

(d) "daft hands" (e) "keyboard cat" (f) "haters gonna hate"

Figure 3: Examples of normalized time series gathered from Google Insights and Delicious. The data indicates how interest in different Internet memes developed over time. From these examples, it seems that the later a meme occurred on the Internet, the higher the degree of correlation of the corresponding time series.

- are of political nature (e.g. activist memes that promote political ideas or malign political opponents)
- are related to personal or commercial web sites.

For each meme, we gathered data from Google Insights that characterize how its popularity or notoriety developed over time.

Google Insights is a service by Google that provides statistics on queries terms users have entered into the Google search engine. It provides weekly summaries of how frequently a query has been used in the time since January 1st 2004 and allows for narrowing down to regions and categories. For our study, we retrieved overall worldwide statistics. Note that Google Insights does not reveal absolute search counts. Rather, the data is normalized such that the peak search activity for a query is scaled to a value of 100. Data obtained from Google Insights therefore indicates relative search frequencies and does not allow for estimating absolute public interest in a topic.

When available, we also collected time series from Delicious, Digg, and StumbleUpon.

Delicious is a social bookmarking service for storing web bookmarks. It has a search facility that summarizes when and how many bookmarks were tagged with a query term. The data is returned in form of summaries covering up to three months but can be easily converted into average daily activity counts. Unlike Google Insights, Delicious thus allows for estimating absolute user activities related to a topic.

Digg is a social news service where users can vote on web content submitted by others. It provides a search API that returns topic related activities of the community. Information is available on a per day basis but, compared to Google Insights or Delicious, there is considerably less usage data.

StumbleUpon is a discovery engine that recommends web content that has been entered by its users. We used the available API to determine at which points in time users commented on content related to our 150 memes. Again, the data is available on a per day basis but is much sparser than in the case of Google Insights or Delicious.

The collected data were converted into a format representing average monthly activities for the period from January 2004 to December 2010. This resulted in discrete time series $\boldsymbol{z} = [z_1, z_2, \ldots, z_T]$ covering a period $T = 84$ months where z_1 represents activities related to a meme in the month of January 2004 and z_T represents the corresponding activities for December 2010.

In order to compare meme related activities across different sources, the data were turned into discrete probability vectors $\boldsymbol{x}$ where $x_t = z_t / \sum_i z_i$. Examples of the resulting normalized time series are shown in Fig. 3

Onset times were determined using the discrete Teager-Kaiser operator

$$TK(x_t) = x_t^2 - x_{t-1}x_{t+1} \tag{1}$$

which is a signal processing technique to detect abrupt variations in a data stream. For each of our time series, the earliest such variation was said to define the onset time t_o. Model fitting in later stages of our analysis was done using truncated time series $\boldsymbol{x} = [x_{t_o}, \ldots, x_T]$.

Immediate Observations and Implications

Looking at the time series in Fig. 3, it seems that over the years there is a growing correlation between the frequencies of meme related queries to Google and activities of the Delicious community. While Internet memes that appeared more than five years ago show different temporal patterns for the two sources, the corresponding time series of memes with onset times later than 2006 seem more closely correlated.

In an attempt to quantify this observation, we examined *weighted average annual correlations* between series from Google Insights and their counterparts from the other services. For each year $y \in \{2004, \ldots, 2010\}$, we considered

$$\text{avgcorr}(y) = \frac{w_y}{N_y} \sum_{\{\boldsymbol{x} | t_o \in y\}} \text{corr}(\boldsymbol{x}, \boldsymbol{x}') \tag{2}$$

where $N_y = |\{(\boldsymbol{x}, \boldsymbol{x}') \mid t_o(\boldsymbol{x}) \in y\}|$ and $\boldsymbol{x}$ is a Google time series whose onset time t_o falls into year y and $\boldsymbol{x}'$ denotes the corresponding data from either of the other services.

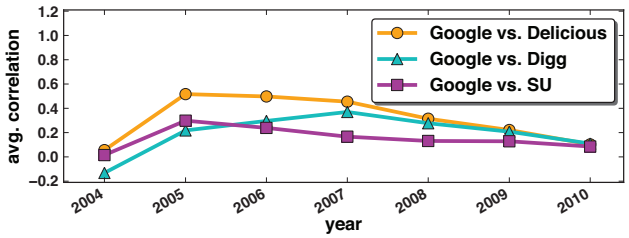


Figure 4: Weighted average annual correlations between meme related time series retrieved from Google Insights, Delicious, Digg, and StumbleUpon. Weighted correlations between Google searches and the activities in social bookmarking services reach peak values for memes with onset times in the years between 2005 and 2007.

The weights $w_{2004} = 7/7, \ldots, w_{2010} = 1/7$ are chosen to penalize larger correlations due to shorter sequence lengths. Figure 4 indicates that, on average, the largest correlations are between the Google and the Delicious time series. For all three services, we observe peak correlations in the years from 2005 to 2007; the rather small values for 2010 suggest that the seemingly increasing correlations in Fig. 3 are indeed an artifact of shorter observation times.

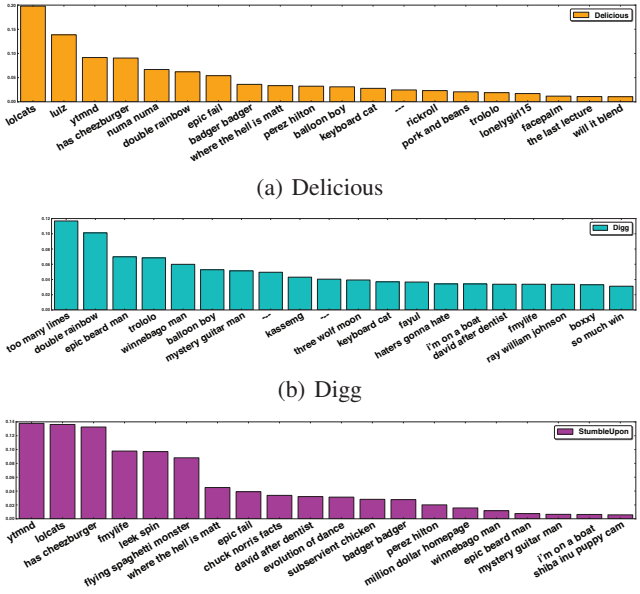
Using the data that allow for the assessment of absolute meme related activities, we compared interests and behaviors of different communities and determined the *average daily activities since onset time* ranked in descending order.

Figure 5 shows the twenty highest ranking memes according to their per day popularity in the Delicious, Digg, and StumbleUpon communities. Given the onset times in Tab. 1, we note that, in the case of Digg, all of the top ranking Internet memes emerged during the last two years. This reflects Digg's role as a social news service: if content or stories that just showed up on the Internet are posted at Digg, users react quickly to the news. Therefore, the shorter the time since onset, the more meme related daily activity there is. On the other hand, memes that have been around for a while hardly provoke further reactions from the Digg community.

We also observe that about a quarter of the memes that are most popular among the StumbleUpon community have to do with rather artistic content ("ytmnd", "fmylife", "flying spaghetti monster", "where the hell is matt", "mystery guitar man"). This is in contrast to the most popular memes determined from Delicious which coincide with memes that are known for their considerable popularity and wide circulation on the Internet. We therefore conjecture that users of recommendation engines are more after sophisticated content than after mundane jokes or fads.

Modeling Meme Dynamics

In part, the work reported in this paper was motivated by a striking observation made while tracking Internet memes using Google Insights: the query frequencies for almost every meme known to have originated later than January 2004 were displayed as a positively skewed curve with a considerably long tail (see again Fig. 3). Characteristics like these



(a) Delicious

(b) Digg

(c) StumbleUpon

Figure 5: Top 20 Internet memes according to average daily activity observed in data retrieved from Delicious, Digg and StumbleUpon. Memes labeled '–' have been garbled for their controversial nature. Memes that are popular among Delicious users are very popular in general; memes that rank high at Digg are very recent; for StumbleUpopn, a larger percentage of popular memes centers around artistic content.

are known from data on daily infectious rates of epidemics and are often studied using stochastic models. In this section, we investigate the use of two classes of models and argue that and why log-normal distributions are well suited to represent the temporal dynamics of Internet memes.

Compartment Models

Compartment models are an established approach to describe the progress of an epidemic in a large population. Typically, the population is thought of as being divided into disjoint fractions of those who are susceptible (S) to the disease, those who are infectious (I), and those who have recovered (R). Some models consider further compartments but they all assume that an individual belongs to one group only. Transitions between groups are constrained by the structure of the model; the SIRS model, for instance, is concerned with transitions of the form $S \to I \to R \to S$ which are governed by the following differential equations

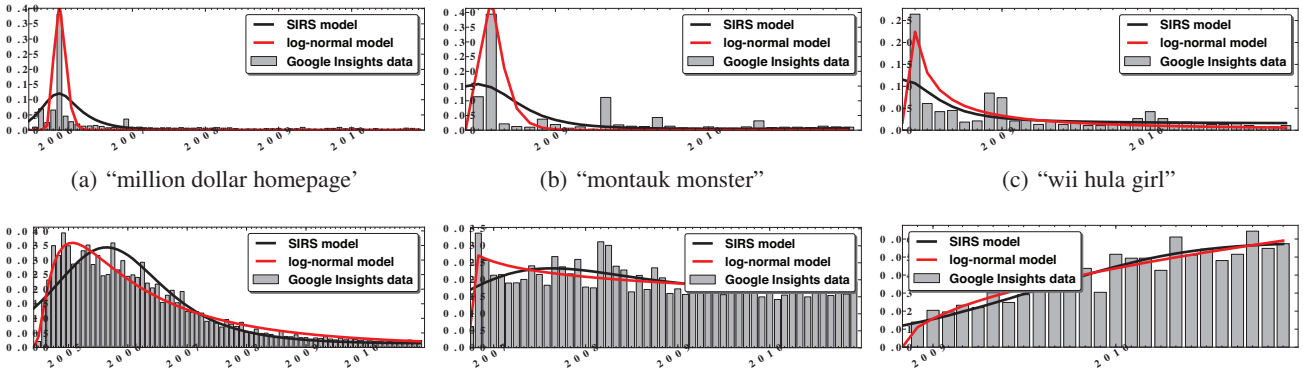
$$\dot{S}(t) = -\beta I(t) S(t) + \phi R(t) \quad (3)$$

$$\dot{I}(t) = \beta I(t) S(t) - \nu I(t) \quad (4)$$

$$\dot{R}(t) = \nu I(t) - \phi R(t) \quad (5)$$

where $S(0) = 1-\epsilon$, $I(0) = \epsilon$, and $R(0) = 0$. The parameter β is the rate of infection, ν is the rate of recovery, and ϕ denotes the average loss of immunity.

Slightly simpler models (of type SI, SIS, SIR) have been used to study information dissemination within web-based



(a) "million dollar homepage' (b) "montauk monster" (c) "wii hula girl"

(d) "salad fingers" (e) "laughing baby" (f) "so much win"

Figure 6: Examples of SIRS and log-normal fits to Google Insights time series that characterize the evolution of interest in different Internet memes. The examples in the top row show pathological cases that are not well accounted for by either model. This occurs if a meme is characterized by a single burst of popularity or by a sequence of such bursts. The bottom row shows more accurate fits for memes of slowly declining, or almost constant, or even constantly growing popularity.

communities (Kubo et al. 2007; Myers and Leskovec 2010) and were reported to give a good account of the interaction dynamics in social networks. We therefore examined the use of stochastic compartment approaches (SIR, SEIR, and SIRS) in modeling the temporal dynamics of meme spread.

The general assumption is that meme related time series $\boldsymbol{x} = [x_{t_o}, \ldots, x_T]$ available from Google Insights correspond to the infectious rates $I(t)$ of epidemic processes.

Note, however, that systems of differential equations as in (3) – (5) are nonlinear so that model fitting is non-trivial. In order to estimate the parameters that would fit a compartment model to a time series of meme related search frequencies, we therefore resorted to Markov Chain Monte Carlo methods. Given observational data for a meme, we generated 1000 proposal distributions using random parameterizations of a compartment model. Suitable parameters that would match the infectious rates of the model to the given time series were then determined in an iterative weighted resampling process. Among the tested compartment models, we found SIRS type models to provide the best explanations of meme activity data.

Figure 6 shows examples of corresponding best matching curves. We note that SIRS models reproduces the general behavior of memes, but, in particular for memes that are characterized by bursty activities, tend to underestimate the early early contagious stages of the meme. This indicates that stochastic compartment models with constant parameters lack the flexibility required to accurately describe the temporal dynamics of Internet memes. While variants with time-dependent parameters might add further flexibility, they would disproportionately increase the difficulty of parameter estimation.

Log-Normal Models

Log-normal distributions have bee successfully used to model frequency distributions of bookmarks or recommendations as well as to characterize response dynamics of social systems (Wu and Huberman 2007; Lescovec, Adamic, and Huberman 2007; Crane and Sornette 2008). They implicitly provide means for the modeling of time-dependent growth and decline rates and therefore appear as an auspicious alternative in studying the temporal dynamics of Internet memes.

A random variable x is log-normally distributed, if $\log(x)$ has a normal distribution. Accordingly, the probability density function of such a random variable is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\big(\log(x) - \mu\big)^2\right). \qquad (6)$$

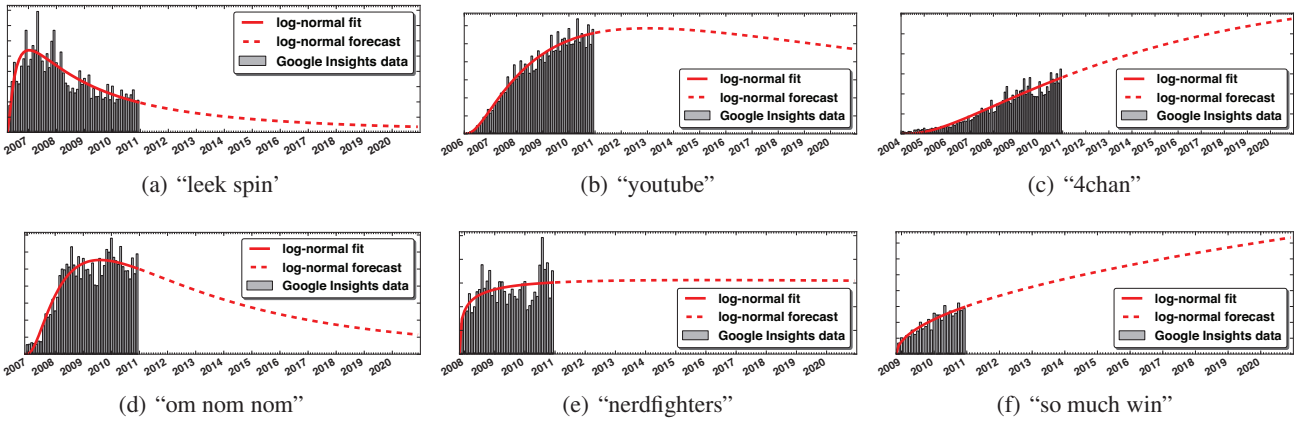
The distribution is only defined for positive values, skewed to the left, and often long-tailed. The mean μ and standard deviation σ of $\log(x)$ define the exact form of the curve.

It can be shown that log-normal distributions are generated by multiplicative processes. Such processes are commonly applied to describe growth and decline in biological or economic systems. Suppose a process starts with a quantity of size x_0 which then, at each time t, may grow or shrink in terms of a percentage of its current size. In other words, the process is governed by a time-dependent random variable γ_t such that

$$x_t = \gamma_t \, x_{t-1}. \qquad (7)$$

Although multiplicative processes and their corresponding log-normal distributions are known to provide accurate models for a variety of Internet related phenomena (Mitzenmacher 2004; Downey 2005), we are not aware of any previous work where they would have been used to study the characteristics of Internet memes.

For each of the 150 time series $\boldsymbol{x} = [x_{t_o}, \ldots, x_T]$ that were obtained from Google Insights, we determined the best fitting log-normal distribution using least squares optimization. Table 1 lists the resulting parameters μ and σ for a subset of 120 memes and Fig. 6 illustrates the behavior of six of the models we obtained this way. Overall, we found log-normal distributions to provide a highly accurate account of the temporal dynamics of the memes under consideration.



(a) "leek spin' (b) "youtube" (c) "4chan"

(d) "om nom nom" (e) "nerdfighters" (f) "so much win"

Figure 7: Forecasts of the future evolution of six popular memes and Internet phenomena according to the log-normal model.

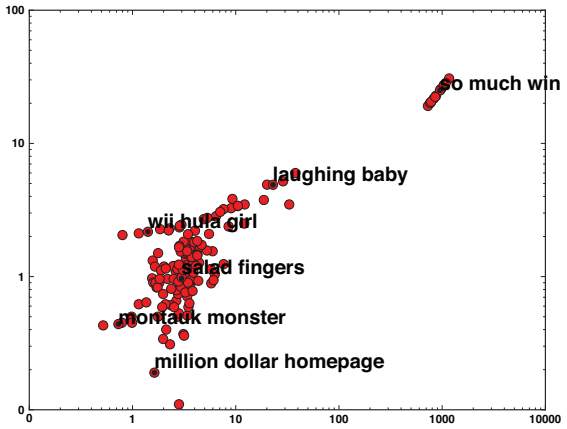


Figure 8: Two-dimensional embedding of 150 Internet memes in (μ, σ) plane where μ and σ are the shape parameters of the log-normal distribution. The majority of memes is found in a cluster represented by the "salad fingers" meme. See Fig. 6 for the appearance of the time series of the six memes whose names are shown here.

In order to quantify this impression, we performed χ^2 goodness of fit tests. With respect to all 150 memes considered here, we found the p-values of SIRS and log-normal models to exceed a confidence threshold of 0.9 in about 70% of the cases. Yet, in 83% of the cases, the p-values obtained for log-normal fits exceeded those of the corresponding SIRS fits. We also determined the Kullback-Leibler divergence

$$D_{KL}(\boldsymbol{x}|\boldsymbol{f}) = \sum_t x_t \log \frac{x_t}{f_t} \qquad (8)$$

between each time series $\boldsymbol{x}$ and its best fitting model $\boldsymbol{f}$. Table 1 lists the resulting D_{KL} measures (closer to 0.0 is better) for SIRS and log-normal fits. In 55% of the cases, we found the log-normal fits to yield better D_{KL} measures than the best fitting SIRS model.

The upper row in Fig. 6 indicates that even in pathological cases where χ^2 tests and D_{KL} measures signal a low quality fit, the log-normal model still provides an acceptable description of the general behavior of the meme. Cases for which both models yield a rather poor account typically correspond to memes that are characterized by either a single burst of popularity or by sequences of such bursts usually due to rekindled interest after news reports in other media. The majority of Internet memes, however, are characterized by time series that are positively skewed and long-tailed. In these cases, as well as for memes that appear not to have reached peak popularity yet, log-normal distributions provide accurate descriptions (see the lower row in Fig. 6).

Implications and Application to Prediction

At this point is important to note that, in contrast to stochastic compartment models such as the SIRS model, log-normal approximations do not model processes and mechanisms of meme spread but summarize corresponding time series.

Nevertheless, the good quality of log-normal fits to meme related time series provides interesting insights. Work by Dover, Goldberg, and Shapira (2010) has established connections between temporal observations of rates of infection (by rumors or marketing messages) and network topologies or links structures of social groups. In particular, it was shown that temporally log-normal diffusion rates indicate networks of log-normal link distributions. In the context of meme spread on the Internet, this is interesting, because it has been observed that although the Internet globally constitutes a scale free graph, it locally consists of homogeneous sub-graphs of log-normal connectivity (Pennock et al. 2002). Therefore, at least for the majority of Internet memes whose temporal penetration data is well represented by log-normal distributions, we conjecture that they spread through rather homogenous communities of similar interests and preferences instead of through the Internet at large.

An immediate application of log-normal models of meme related time series is to apply the resulting descriptions in order to produce a compressed representation of memes in the space spanned by the shape parameters μ and σ. Figure 8 shows the corresponding two-dimensional embedding of the 150 memes considered in this paper. We find the ma-

Table 1: 120 Internet memes and their statistics.

		SIRS	log-normal		
meme	onset	D_{KL}	μ	σ	D_{KL}
all your base	<01/04	0.01	3.81	1.68	0.05
badger badger	<01/04	0.01	20.12	4.90	0.01
bubb rubb	<01/04	0.02	3.53	1.86	0.09
schfifty five	<01/04	0.03	3.29	1.37	0.07
weebl and bob	<01/04	0.02	3.32	1.49	0.09
bert is evil	<01/04	0.01	3.57	1.62	0.06
gunther ding dong	<01/04	0.17	3.37	1.22	0.04
subservient chicken	03/04	1.42	1.15	2.11	0.25
bananaphone	04/04	0.41	4.03	2.21	0.04
salad fingers	06/04	0.16	2.99	0.97	0.02
i love bees	06/04	1.76	1.55	0.97	0.36
pure pwnage	08/04	0.17	3.39	0.71	0.06
zoomquilt	09/04	0.19	3.23	1.49	0.11
llama song	09/04	0.13	3.00	0.88	0.03
hopkin green frog	10/04	0.86	2.29	2.30	0.14
crazy frog	10/04	0.15	2.66	0.66	0.14
numa numa	12/04	0.05	3.44	1.48	0.05
full of win	01/05	0.45	727.91	19.11	0.38
boom goes the dynamite	03/05	0.46	4.90	2.70	0.17
leeroy jenkins	04/05	0.41	2.93	1.54	0.02
o rly	04/05	0.10	2.76	0.74	0.13
ytmnd	04/05	0.03	3.11	0.91	0.02
flying spaghetti monster	05/05	0.19	3.62	1.46	0.18
ya rly	06/05	0.09	2.83	1.06	0.07
pedobear	06/05	0.03	5.79	0.89	0.02
million dollar homepage	08/05	0.91	1.63	0.19	0.79
asian backstreet boys	09/05	0.61	1.71	0.83	0.08
chuck norris facts	09/05	0.21	2.62	1.10	0.23
laughing interview	09/05	1.03	2.54	1.20	1.18
no wai	09/05	0.08	5.51	2.09	0.02
peanut butter jelly time	10/05	0.08	3.45	2.08	0.04
crazy robot dance	10/05	2.21	1.59	0.90	2.47
charlie the unicorn	10/05	0.06	3.54	0.81	0.08
diet coke mentos	10/05	0.47	2.31	0.31	0.48
one red paperclip	10/05	0.41	2.35	0.61	0.42
ask a ninja	12/05	0.09	2.66	0.79	0.03
funtwo	12/05	0.08	2.94	0.84	0.08
do a barrel roll	01/06	0.09	1167.77	30.73	0.02
evolution of dance	03/06	1.27	1.82	0.90	0.45
loituma	03/06	0.27	2.19	0.96	0.04
la caida de edgar	04/06	1.10	1.57	1.32	0.03
leek spin	04/06	0.12	3.85	1.27	0.02
giant enemy crab	04/06	0.36	6.44	2.84	0.09
chad vader	06/06	0.22	2.71	1.15	0.04
lonelygirl15	07/06	1.36	0.98	0.50	0.43
music is my hot sex	07/06	0.40	2.82	0.11	0.66
shoop da whoop	08/06	0.04	5.98	1.55	0.01
lulz	08/06	0.02	4.02	0.94	0.03
noah takes a photo ...	08/06	0.32	9.02	3.26	0.15
mudkips	08/06	0.02	3.62	0.94	0.03
monorail cat	09/06	0.11	3.77	1.21	0.02
will it blend	09/06	0.25	3.75	1.40	0.13
caramelldansen	09/06	0.11	3.40	0.51	0.09
laughing baby	10/06	0.02	22.95	4.89	0.01
epic fail	11/06	0.04	6.27	1.04	0.02
om nom nom	12/06	0.02	4.34	0.93	0.02
it's over 9000	12/06	0.09	976.64	25.66	0.02
lol wut	02/07	0.03	4.11	1.20	0.01
facepalm	02/07	0.03	33.00	3.48	0.01
crank that	03/07	0.37	2.12	0.40	0.20

Table 1: 120 Internet memes and their statistics.

		SIRS	log-normal		
meme	onset	D_{KL}	μ	σ	D_{KL}
has cheezburger	03/07	0.01	3.45	0.90	0.02
allison stokke	03/07	0.93	1.67	0.90	0.66
rickroll	03/07	0.12	3.25	0.76	0.12
lolcats	04/07	0.02	3.89	1.07	0.02
fukken saved	04/07	0.19	4.52	1.66	0.03
daft hands	05/07	0.10	2.51	0.96	0.05
dramatic chipmunk	05/07	0.91	2.85	2.34	0.14
i like turtles	05/07	0.20	37.83	5.99	0.11
powerthirst	05/07	0.05	3.20	1.05	0.06
chocolate rain	06/07	0.60	2.85	1.70	0.08
my new haircut	06/07	0.15	2.37	0.81	0.08
raymond crowe	07/07	1.16	1.65	1.19	0.15
benny lava	08/07	0.20	3.21	1.17	0.03
leave britney alone	08/07	1.37	0.80	2.05	0.33
techno viking	08/07	0.20	12.21	3.48	0.06
daft bodies	10/07	0.32	3.96	1.81	0.02
tinaecmusic	10/07	0.80	2.02	1.19	0.33
charlie bit me	10/07	0.09	4.26	1.46	0.07
magibon	10/07	0.15	2.78	0.92	0.11
nerdfighters	11/07	0.09	18.72	3.76	0.02
the last lecture	12/07	0.29	2.09	0.62	0.26
tron guy	02/08	0.20	7.62	3.21	0.20
vernon koekemoer	02/08	1.44	1.15	0.62	0.14
interior crocodile alligator	02/08	0.03	5.26	1.57	0.03
push button receive bacon	03/08	0.46	3.11	0.37	0.37
pork and beans	03/08	0.18	1.75	0.83	0.14
ninja cat	04/08	0.07	2.99	0.99	0.13
where the hell is matt	05/08	0.12	2.82	1.66	0.03
wii hula girl	05/08	1.39	1.41	2.17	0.33
ran ran ru	05/08	0.17	28.76	5.19	0.04
bert ernie rap	06/08	0.41	10.51	3.40	0.12
montauk monster	06/08	2.20	0.73	0.44	0.83
totally looks like	06/08	0.15	961.69	25.30	0.03
i dunno lol	07/08	0.16	1043.92	27.38	0.02
scarlet takes tumble	09/08	0.72	1.89	1.11	0.08
shiba inu puppy cam	09/08	1.80	0.99	0.45	0.67
yo dawg	10/08	0.03	4.67	1.73	0.03
shut down everything	10/08	0.11	4.24	1.44	0.07
so much win	11/08	0.11	952.35	25.00	0.01
boxxy	12/08	0.51	7.07	3.06	0.11
fmylife	12/08	0.38	1.98	0.74	0.07
courage wolf	12/08	0.13	1033.81	27.28	0.02
david after dentist	01/09	1.09	2.98	2.45	0.25
i'm on a boat	01/09	0.52	1.83	0.96	0.02
kia hamster	02/09	0.50	855.73	22.34	0.48
haters gonna hate	02/09	0.02	3.16	0.36	0.03
this is photobomb	03/09	0.14	3.09	1.12	0.02
keyboard cat	04/09	0.18	2.87	2.41	0.02
three wolf moon	04/09	0.63	5.31	2.76	0.12
socially awkward penguin	04/09	0.08	1082.39	28.40	0.03
crasher squirrel	07/09	2.59	0.52	0.43	0.24
balloon boy	09/09	3.88	-2.60	1.18	0.19
french the llama	11/09	0.31	8.53	2.38	0.12
hipster kitty	11/09	0.28	12.10	2.50	0.04
winnebago man	01/10	0.07	1.98	0.34	0.07
epic beard man	01/10	2.49	-0.82	1.63	0.11
trololo	01/10	0.80	1.36	0.64	0.17
double rainbow	06/10	0.38	1.77	1.50	0.01
too many limes	06/10	0.99	2.10	1.15	0.01
fayul	10/10	1.87	4.24	1.85	0.00

jority of memes clustered around the "salad finger" meme which corroborates the observation that time series of meme related activities are typically skewed and long-tailed. We also note a distinct cluster of memes on the top right. These are memes or Internet phenomena for which the mean μ was estimated to be large therefore indicating a pattern of still increasing popularity.

The existence of such memes led us to attempt a forecast of their future evolution according to the corresponding log-normal model. Figure 7 depicts 10-year forecasts for a collection of six memes or meme related web sites. Certainly, these forecasts will have to be taken with a grain of salt for they do not envision possibly disruptive events. Nevertheless, our predictions look plausible. In contrast to related recent work (Yang and Lescovec 2011), they were obtained without having to learn predictive models from large amounts of data.

Conclusion

The term *Internet meme* is used to describe evolving content that rapidly gains popularity or notoriety on the Internet. As of late, Internet memes have attracted increased public interest and a growing number of web sites and communities are dedicated to this topic. Moreover, professionals in marketing and campaigning have embraced Internet memes as a way to build rapport with trendy communities.

Given the growing interest in Internet memes, there is sur-

prisingly little scientific work on the phenomenon so far. In particular, data-driven models that would allow for characterizing the dynamics of a meme or even for forecasting its longevity or peak circulation are scarce.

In this paper, we investigated the temporal dynamics and infectious properties of 150 famous Internet memes. Our analysis was based on time series that were collected from Google Insights, Delicious, Digg, and StumbleUpon. From this data, we identified distinct interests in the corresponding communities. Among other results, we saw that users of the Digg social news service predominantly react to recent memes and users of the StumbleUpon recommendation engine appear to be interested mostly in sophisticated memes.

We also examined the use of different mathematical models of epidemic spread in the context of Internet memes. We found that elaborate traditional compartment models with constant parameters give a good account of the growth and decline patterns of memes yet lack the flexibility to characterize short-lived bursts of meme related activity. Log-normal distributions, on the other hand, implicitly account for time-dependent growth and decline rates. We found log-normal distributions to yield accurate summaries of the temporal dynamics of Internet memes; in statistical significance tests, we found that for 70% of the 150 memes considered in this paper the probability of a log-normal model underlying the observed data distribution exceeded 90%. Taking into account the fact that log-normal diffusion processes indicate networks of log-normal link distributions (Dover, Goldberg, and Shapira 2010) and the observation that the globally scale free Internet graph appears to contain many log-normal subgraphs (Pennock et al. 2002), we conjecture that the majority of currently famous Internet memes spreads through homogenous communities and social networks rather than through the Internet at large.

Acknowledgements

We want to thank our anonymous reviewers for helpful remarks and pointers to the recent literature. Also, we gladly acknowledge the help of Fabian Beckmann, Lukas Havemann, Josua Sassen, and Fabian Thorand who participated in the Fraunhofer Talent School 2010 and helped with initial experiments. Finally, we thank Tansu Alpcan for insightful discussions on stochastic models of network dynamics.

References

Adar, E., and Adamic, A. 2005. Tracking Information Epidemics in Blogspace. In *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence*.

Bloem, M.; Alpcan, T.; and Basar, T. 2009. Optimal and Robust Epidemic Response for Multiple Networks. *Control Engineering Practice* 17(5):525–533.

Britton, T. 2010. Stochastic Epidemic Models: A Survey. *Mathematical Biosciences* 225(1):24–35.

Budak, C.; Agrawal, D.; and Abbadi, A. E. 2010. Limiting the Spread of Misinformation in Social Networks. In *Proc. ACM Int. Conf. on WWW*.

Crane, R., and Sornette, D. 2008. Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. *PNAS* 105(41):15649–15653.

Dawkin, R. 1976. *The Selfish Gene*. Oxford University Press.

Dietz, K. 1967. Epidemics and Rumors: A Survey. *J. of the Royal Statistical Society A* 130(4):505–528.

Dover, Y.; Goldberg, J.; and Shapira, D. 2010. Uncovering Social Network Structures through Penetration Data. unpublished working paper.

Downey, A. 2005. Lognormal and Pareto Distributions in the Internet. *Computer Communications* 28(7):790–801.

Keeling, M., and Eames, K. 2005. Networks and Epidemic Models. *J. Royal Society Interface* 2(4):295–307.

Kubo, M.; Naruse, K.; Sato, H.; and Matubara, T. 2007. The Possibility of an Epidemic Meme Analogy for Web Community Population Analysis . In *Proc. int. Conf. on Intelligent Data Engineering and Automated Learning*.

Lescovec, J.; Adamic, L.; and Huberman, B. 2007. The Dynamics of Viral Marketing. *ACM Tans. on the Web* 1(1):5.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the Dynamics of the News Cycle. In *Proc. ACM Inf. Conf. on Knowledge Discovery and Data Mining*.

Limpert, E.; Stahel, W.; and Abbt, M. 2001. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51(5):341–352.

Lin, C.; Zhao, B.; Mei, Q.; and Han, J. 2010. PET: A Statistical Model for Popular Events Tracking in Social Communities. In *Proc. ACM Inf. Conf. on Knowledge Discovery and Data Mining*.

Lloyd, A., and May, R. 2001. How Viruses Spread Among Computers and People. *Science* 292(5520):1316–1317.

Mitzenmacher, M. 2004. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics* 1(2):226–251.

Myers, S., and Leskovec, J. 2010. On the Convexity of Latent Social Network Inference. In *Proc. Conf. on Neural Information Processing Systems*.

Pastor-Satorras, R., and Vespignani, A. 2001. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters* 86(14):3200–3203.

Pennock, D.; Flake, G.; Lawrence, S.; Glover, E.; and Gilles, C. 2002. Winners Don't Take All: Characterizing the Competition for Links on the Web. *PNAS* 99(8):5207–5211.

Shah, D., and Zaman, T. 2009. Rumors in a Network: Who's the Culprit? In *Proc. NIPS Worshop on Analyzing Networks and Learning with Graphs*.

Wu, F., and Huberman, B. 2007. Novelty and Collective Attention. *PNAS* 104(45):17599–17601.

Yang, J., and Lescovec, J. 2011. Patterns of Temporal Variation in Online Media. In *Proc. ACM Int. Conf. on Web Search and Data Mining*.