Viral, Quality, and Junk Videos on YouTube: Separating Content From Noise in an Information-Rich Environment

R. Crane and **D. Sornette**
Chair of Entrepreneurial Risks
Department of Management, Technology and Economics
ETH-Zürich, CH-8032 Zürich, Switzerland

Introduction

With the rise of web 2.0 there is an ever-expanding source of interesting media because of the proliferation of user-generated content. However, mixed in with this is a large amount of noise that creates a proverbial "needle in the haystack" when searching for *relevant* content. Although there is hope that the rich network of interwoven metadata may contain enough structure to eventually help sift through this noise, currently many sites serve up only the "most popular" things.

Identifying only the most popular items can be useful, but doing so fails to take into account the famous "long tail" behavior of the web—the notion that the collective effect of small, niche interests can outweigh the market share of the few blockbuster (i.e. most-popular) items—thus providing only content that has mass appeal and masking the interests of the idiosyncratic many.

YouTube, for example, hosts over 40 million videos—enough content to keep one occupied for more than 200 years. Are there intelligent tools to search through this information-rich environment and identify interesting and relevant content? Is there a way to identify emerging trends or "hot topics" *in addition* to indexing the long tail for content that has real value?

Information about quality is contained in the dynamics

We demonstrate that this is possible based on a form of dynamic filtering. In essence, the relaxation signature following a burst of viewing activity reveals information about the quality of the content. This signature depends on the susceptibility of the social network, in addition to the type of perturbation that generated the burst.

We begin by considering two classes of perturbations: endogenous and exogenous. Their distinction—which is not required to be known *a priori*—is illustrated in figure 1. Here we show the aggregate time-series for videos appearing on the front page of YouTube along with those appearing on the 'most-viewed today' page. Videos appearing on the front-page are chosen by the editors, whereas those on the 'most-viewed today' page are 'chosen' in a collaborative sense by the collective actions of the community (by making a video the 'most-viewed'). We find that videos chosen by the editors (exogenous) have a strikingly different history than those chosen by the community (endogenous). While both classes show a power-law relaxation (inset) over one-hundred days following the peak, the videos featured by the community clearly display significant precursory growth.

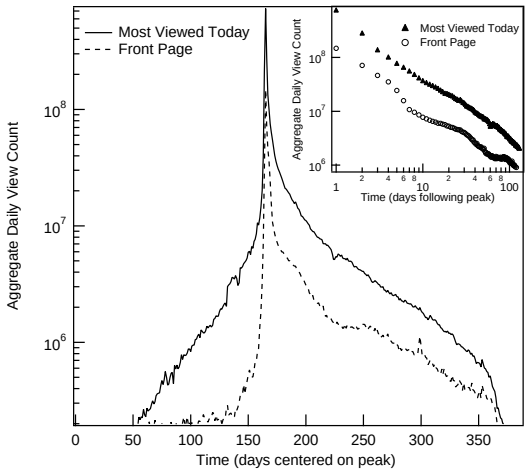

Figure 1: A non-parametric superposition of all videos appearing on the 'front page' (editorial featuring) and the 'most-viewed today' page (community featuring). One immediately sees the exogenous effect of editorial featuring, revealed by the lack of precursory growth in the view count, whereas endogenous growth is seen in the case of community featuring. Inset: power-law relaxation in the 100 days following the peak reveals long-memory effects.

Once a burst of activity has been triggered, its relaxation depends on the susceptibility of the underlying social network. If the community is "ripe" for the content, then each generation of viewers can easily pass on the video to the next generation, and one will find the view count relaxes slowly. If instead the community is "uninterested", then even a well-orchestrated marketing campaign will fail to spread through the network and one will witness a fast relaxation.

Description of Data and Model

These ideas have been formalized and tested using a massive database tracking the time-series of the daily views over 1 year for almost 5 million videos on the popular site YouTube.com.

Quantifying these effects can be achieved by studying the dynamical response of the daily view count in the context of an epidemic branching process on a social network. This model was previously applied successfully to the case of book sales (Sornette *et al.* 2004). The instantaneous view count of a video results from many factors such as featuring on YouTube, emailing (or other forms of sharing videos), embedding and linking from external websites, discussion on blogs, in newspapers, television, and from social effects in which viewers may be influenced by others in their network. The impact of these various factors may not be immediate, and this latency can be described by a response function $\phi(t - t_i)$, which on the basis of figure 1 we postulate to be a long-memory process of the form $\phi(t) \sim 1/t^{1+\theta}$, with $0 < \theta < 1$. Using this, we can describe the rate of views as a self-excited Hawkes conditional Poisson process that depends on all past events

$$\lambda(t) = V(t) + \sum_{i|t_i<t} \mu_i \phi(t - t_i) \tag{1}$$

where μ_i is the number of potential viewers influenced by a viewer at time t_i and $V(t)$ captures all spontaneous views that are not triggered by network effects.

When the network is not "ripe", corresponding to the case when $\langle \mu_i \rangle$ is less than 1, then the activity generated by an exogenous event does not cascade beyond the first few generations, and the activity is given by

$$A_{bare}(t) \sim \frac{1}{(t - t_c)^{1+\theta}} \tag{2}$$

If instead the network is "ripe" for a particular video, then the bare response is renormalized as the spreading is propagated through many generations, and the theory predicts the activity to be described as

$$A_{exo}(t) \sim \frac{1}{(t - t_c)^{1-\theta}} \tag{3}$$

If in addition to being "ripe", the burst of activity is not the result of an exogenous event, but is instead fueled by endogenous growth, the bare response is renormalized in a different way giving

$$A_{endo}(t) \sim \frac{1}{(t - t_c)^{1-2\theta}} \tag{4}$$

While these results strictly hold for an ensemble of time-series because of the stochasticity involved, we find a surprisingly large number of individual videos that seem to obey these power-law relaxations exactly. Examples of this are shown in figure 2, suggesting that we can apply this formalism to individual videos.

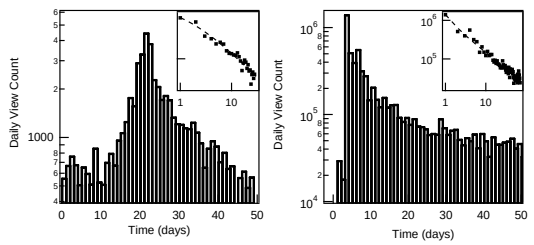

Figure 2: Examples of endogenous (left) and exogenous (right) bursts of activity for individual videos. The exponent of the power-law relaxation (inset) can be used to classify videos as *viral, quality, or junk*.

Classification of Content

As outlined above, the existence of memory in the view count dynamics implies that the relaxation signatures following a burst of activity depend on the susceptibility of the underlying social network to a particular video. We can therefore use the dynamic signature as a way of distinguishing—on the basis of the exponent of the power law governing their relaxation—between three extreme cases: *viral videos, quality videos, and junk*.

In this context, **viral videos** are those with precursory word-of-mouth growth resulting from epidemic like propagation through a social network, characterized by an exponent $(1 - 2\theta)$. **Quality videos** are similar to viral videos, but experience a sudden burst of activity rather than a bottom-up growth, and because of the "quality" of their content, subsequently trigger an epidemic cascade through the social network, relaxing with an exponent $(1-\theta)$. Lastly, **junk videos** are those that experience a burst of activity for some reason (spam, chance, etc) but do not spread through the social network. Therefore their activity is determined largely by the first-generation of viewers, and they should relax as $(1 + \theta)$.

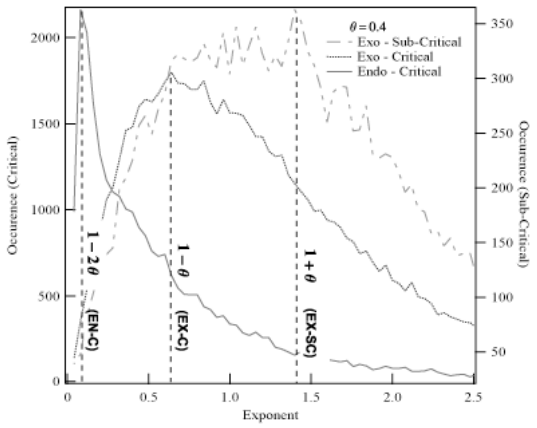

Figure 3: Exponents for videos grouped by the fraction of views contained in their most active day (peak) relative to the total. This is a natural way of separating endogenous from exogenous videos, since the former have significant precursory growth, thus lowering the fractional weight contained in the peak.

Figure 3 shows the distribution of exponents obtained by grouping videos based on the fraction of views contained in their peak relative to the total. Videos experiencing an exogenous shock should have a very high percentage because there is little precursory growth, which is opposite for the endogenous case. Immediately one sees that based on this very simple criterion, the videos naturally fall into separate exponent classes, and we can extract $\theta = 0.4$ based on this picture.

A final interesting result is that this classification does not rely on the magnitude of the largest peak, implying that identification of content can be made for large communities as well as more specialized, niche communities.

References

Sornette, D.; Deschâtres, F.; Gilbert, T.; and Ageon, Y. 2004. Endogenous versus exogenous shocks in complex networks: an empirical test using book sale ranking. *Phys. Rev. Lett.* 93(22):228701.