# Video Genetics: A Case Study from YouTube

John R. Kender
Columbia University
1214 Amsterdam, MC 0401
New York, NY 10027
jrk@cs.columbia.edu

Matthew L. Hill
IBM Research
19 Skyline Dr., P.O. Box 704
Hawthorne, NY 10532
mh@us.ibm.com

Apostol (Paul) Natsev
IBM Research
19 Skyline Dr., P.O. Box 704
Hawthorne, NY 10532
natsev@us.ibm.com

John R. Smith
IBM Research
19 Skyline Dr., P.O. Box 704
Hawthorne, NY 10532
jsmith@us.ibm.com

Lexing Xie
IBM Research
19 Skyline Dr., P.O. Box 704
Hawthorne, NY 10532
xlx@us.ibm.com

## ABSTRACT

We explore in a single but large case study how videos within YouTube, competing for view counts, are like organisms within an ecology, competing for survival. We develop this analogy, whose core idea shows that short video clips, best detected across videos as near-duplicate keyframes, behave similarly to genes. We report work in progress, on a dataset of 5.4K videos with 210K keyframes on a single topic, which traces *sequences*, not bags, of "near-dups" over time, both within videos and across them. We demonstrate their utility to: cleanse responses to queries contaminated by over-eager YouTube query expansion; separate videos temporally according to their responses to external events; track the evolution and lifespan of continuing video "stories"; automatically locate video summaries already present within a video ecology; quickly verify video copying via a direct application of the Smith-Waterman algorithm used in genetics—which also provides useful feedback for tuning the near-dup detection and clustering process; and quickly classify videos via a kind of Lempel-Ziv encoding into the categories of news, monologue, dialogue, and slideshow. We demonstrate a number of novel visualizations of this large dataset, including a direct use of the Matlab black-body "hot" false-color map, together with the GraphViz package, to display the gene-like inheritance of viral properties of keyframes. We further speculate that, as with genes, there are "functional roles" for semantic categories of clips, and, as with species, there are differings rates of "genetic drift" for each video genre.

## Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

## General Terms

Algorithms, Design, Management

## Keywords

Near-duplicate keyframes; video ecology visualization; video evolution; video genetics; video mashups; video memes; video Smith-Waterman matching; video species.

## 1. INTRODUCTION

### 1.1 The Genetics of Videos

Video data differs from purely textual data in several important ways. It is often easier to generate, but it is harder to manipulate, compare, and query over. Consequently, social repositories of video data such as YouTube depend on verbal title, tags, and descriptions for searching, even though each of these textual records can bear little relation to the actual video content. Sometimes this inaccuracy is deliberate for spamming purposes, but often it is due to the honest polysemy of words. One can and does expect to find that the precision of an answer to a query is inexact. Further, retrieval may even be complicated by the server's heuristic attempts at query expansion, such as including videos whose supposed relevance to the query comes mainly from their popularity, or from their collocation within ill-formed playlists whose other videos are in fact legitimate responses.

Examination of parts of the YouTube ecologies suggest that the users tend not to repeat others' entire videos whole and unaltered—although they may repeatedly repost copies of their own. But neither do they often create highly edited original stories, except for the common degenerate form of slideshows made of random still shots meant to accompany music or monologue. Instead, the unit of video propagation appears to be the clip, a short video segment consisting of a single frame, a short shot, or a few contiguous shots. These

*MM'10,* October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

are most often modified only in relatively minor ways, such as with the superimposition of text or watermark “bugs”.

We have investigated how videos are promulgated in depositories, by tracking the sources, reuse, and view counts of these clips. We analyze them through finding and matching their simplest proxies: single key frames that are nearly-duplicated across videos. In deriving algorithms to detect, cluster, track, compare, and display these viral “near-dups”, we noted that their behavior tended to resemble that of genes; indeed, some of our tools turned out to be rediscoveries of some tools of genetic research. We find that the virality of videos is best thought of as the survival of these “video genes”.

## 1.2 The Full Analogy

We present a fuller analogy in Table 1. We report that this mapping between an ecosystem and a social video depository has already suggested to us useful representations and powerful tools that we had not originally considered.

| Concept | Biology | vs. | Videos |
|---|---|---|---|
| Environment | An ecology | > | YouTube |
| Discrete form | Organism | > | Video |
| Observables | Phenotype | > | Genre |
| Composition | Genotype | = | ”Grammar” |
| Encoding | Amino acid, ACGT | < | Near-duplicate |
| Entire genome | DNA | > | Clip sequence |
| Functional | Amino acid family | ? | Virality rank |
| Nonfunctional | Intron | > | Editorialization |
| End protect | Telomeres | > | End credits |
| Categorization | Phylogenetic tree | < | Timeline trace |
| Seq alignment | Smith-Waterman | = | Dyn Time Warp |
| Multi align | Clustal | > | Clustal-like? |
| Misalignment | Genetic gap | = | Segment bug |
| Repeat detect | Electrophoresis | < | Lempel-Ziv |
| Display color | By am.acid family | < | By “hotness” |
| Natural selection | Food, reproduction | = | Views |
| Mutualism | Symbiosis | < | Playlist |
| Reproduction | Sexual | < | Mashup |
| Reproduction | Asexual | < | Copying |
| Interbreeding | Within specie | > | Within topic |
| Evolution | Genetic drift | ? | Story drift |

**Table 1: A social video depository acts like a biological world.**

Some of the parallels are straightforward. Videos follow grammar rules of good formation (genotypes), and have immediately observable differences by genre (phenotypes). The sequences within and across short clips tend to be preserved, except for editorial inclusions (introns) and superfluous beginnings and endings (telomeres). Their relationships to each other can be traced by their reuse of near-duplicate shots (phylogenetics), discoverable through time-flexible matching algorithms (Smith-Waterman and Clustal). Near-dups are sometimes dropped or interrupted (genetic gaps), but their basic patterns of repetition are indicative of structure, which can be displayed graphically (electrophoresis). Competition for views can be enhanced by coexistence on playlists with related videos (symbiosis), or by copying (asexual reproduction) or by mashups (sexual reproduction), leading to evolution of content (genetic drift).

Nevertheless, we note that there are some differences in the orders of magnitude that distinguish the two domains. As noted in the table, compared to the genes that make up the DNA of an organism, near-duplicate keyframes of a video are made from a much larger alphabet, and each video encodes a much shorter sequence of them. Although this simplifies video analysis, it impairs the use of many genetic visualization techniques, such as the simple false-coloring of base pairs or amino acids in displays of lineage or similarity. (In particular, neither the SAX nor Intelligent Icons visualizations [7] apply either.) On the other hand, videos come with timestamp information, so their inheritance and evolution are much easier to know with certainty. One of the most important parallels, however, is that genetics encodes information within grammatical sequences; unlike bag-of-X approaches, sequential temporal information, both intra- and inter-video, encodes and yields critical information about genre and ancestry.

## 1.3 Related Work

There are a number of related works on studying the lifetimes and interaction of communication artifacts; we cannot do justice to them all here. We note that political phrases have been tracked and studied over time [6], as has the dissemination of newsletters [5], as well as the distribution through “fanning” of Facebook news feeds [10].

Among their conclusions are that usage patterns have long tails, and are not easily modeled analytically; they tend to follow roughly a power or log-normal distribution. There does not seem to be any features of a user or environment that can predict virality at birth, but early growth patterns relative to peers might [2]. The work of [8] on song ratings suggests that the best predictor of virality is virality itself: the rich get richer, even if early riches are accidental.

The closest work to our investigations may be that of [3], who found that YouTube videos tended to fall into just three categories: viral, quality, and junk, and that each approximately follows a power law, but each with a different exponent.

Unlike most of the efforts on social network analysis, the focus of this research is on the relationships of the artifacts, rather than of the artists themselves. Generating attractive new video content is difficult work, and it often only appears in response to external world events. More common are borrowings and incremental modifications, which do not appear to require any deep commitment to a particular subcommunity. Compared to studies of groups of scientific researchers or political activists, for example, here the ecological forces are simple and the propagation tools are cheap. Evolution rather than affinity appears to be the dominant phenomenon.

# 2. THE CASE STUDY

## 2.1 Data Collection

During three months of 2009, we collected 200K videos and their metadata from YouTube, on 22 different topics. We report a case study here on one of these, the “Iran” dataset, for which our spider repeatedly made about a dozen queries through the YouTube API, each of the form, “Iran election”, “Iran president”, “Iran controversy”, “Iran killing”, etc. These queries retrieved about 5.4K unique videos, eventually yielding about 210K keyframes. A second topic, “Housing”, provided ground truth for tuning our algorithms: about 15K near-dup pairs and 25K non-near-dup pairs from this dataset were hand-annotated and verified.

To compare keyframes, we first normalized them spatially and through histogram equalization, then masked away the four corners of each image to reduce interference by added text and graphics, then extracted features from the remaining cross-shaped region based on HSV color correlograms. Each keyframe's resulting 332-component descriptor was then indexed for k-nearest neighbor lookup with respect to the $L^2$ distance metric. We used the Fast Library for Approximate Nearest Neighbor (FLANN) package; for performance, we set a heuristic limit on number of neighbors searched, equal to $\sqrt{N}$ where $N$ is the size of the keyframe set. We kept only those neighbors within an adaptive threshold that was sensitive to descriptor complexity. Equivalence classes of near-duplicates resulted from an efficient union-find variant for the transitive closure of these neighbor relations. On this dataset, this method created about 2.5K near-dup classes; over all datasets, precision was 98%, recall was 80%, and F1 was 88%, at a total cost of $O(N\sqrt{N})$ per dataset.

## 2.2 Global Analyses

By examining these near-dup classes, it was not hard to discover that YouTube appears to answer queries using a (proprietary) query expansion algorithm to enhance recall, but often at the cost of lowered precision. It appears to select very popular but off-topic videos if they happen to share a playlist with other videos that directly satisfy the query, and it appears to weight this selection proportionately by viewcount. Since many playlists are off-topic and ill-specified ("My Favorite Videos"), contamination of the response set results; in the case study dataset, the 4 most popular videos were so off-topic that they shared no near-dup keyframes with any other videos at all; likewise, for 7 of the top 10. For videos inthe top 10% of popularity (which gathered almost all the total views), the likelihood of having a near-dup was below the average of the dataset as a whole, which was about .58. However, noting their lack of near-dups, these off-topic "invasive species" are easily detected, and they in fact do not "interbreed".

In general, this particular YouTube topic domain, like many others, is characterized by a very unequal distribution of views, as measured by the Gini coefficient [4] used in biodiversity studies. Gini computes a state of pure equity as a value of 0: applied here, it would mean each video has exactly the same number of views. A Gini coefficient of 1 indicates that all resources are possessed by a single entity: here, one video would have all the views. The Gini coefficient of the case study dataset is .94, whether one looks only at those videos with near-duplicates, or includes those without. This value far exceeds the Gini coefficient of inequality for the distribution of monetary wealth in any country, which has its maximum at about .7 for Namibia.

Additionally, we have noted that the member keyframes of a near-duplicate class tend to be temporally collocated preferentially with members of only very few other near-duplicate classes. Further, the temporal order of these pairings tends to be preserved across videos. This can be seen by examining bigram statistics, which record the frequency of adjacent pairs of near-dups within a video. If $B$ is a matrix that counts at $B(i, j)$, the number of times in the video ecology that a near-dup from class $i$ is followed by a near-dup from class $j$, then $B$ is shown to be heavily asymmetric. A standard definition of asymmetry represents $B$ as the sum of a purely symmetric matrix $S = (B + B^T)/2$ and a purely anti-symmetric matrix $K = (B - B^T)/2$, and defines the amount of asymmetry as $0 \leq a = \|K\|/\|B\| \leq 1$, where the norm can be chosen as appropriate; the extremes values of $a$ occur exactly where expected. For some norms, our case study dataset has $a = .6$, and if the diagonal (self-succession of near-dups) is omitted, $a = .7$. Observation confirmed the severe non-Markovian property of near-dups, with many identical long chains of near-dup classes repeating within and across videos. One conclusion is that the unit of video virality is more properly that of the keyframe *sequence*, that is, a clip; here we use individual near-dups as an efficient proxy.

# 3. OVERVIEW VISUALIZATION

In this work in progress, the genetic inheritance of videos has been defined via the simplest possible approximation: as an unweighted directed link between a pair of videos that share at single near-dup, but restricted to only those pairs which are immediately temporally adjacent. This is essentially a case of transitive reduction (that is, the inverse to transitive closure), where each near-dup induces its own single time-based simple path through the video ecology. (There is no metadata available from YouTube that could help to further determine true inheritance.)

We illustrate this with a machine-generated graph, from Matlab through GraphViz, which presents in greatly reduced and symbolic form the inheritances of near-dups within the case study dataset on Iran. We note at a glance that, although there are many isolated components, videos tend to generate and interbreed in synchrony with important events. The two most prominent tangles correspond to the Iran election protests of June 13 and the death by sniper fire of a young woman on June 20. We also note that initial videos about an event tend to acquire more views than later ones, and they also tend to be subject to more evolution (although we have not yet defined and completed measurements for these hypotheses.)

Because keyframes are clustered according to near-dup similarity, and these classes are then sorted and numbered according to their cardinality, we can use either this near-dup class number, or the cardinality itself, as a rough indicator of near-dup virality. We visualize this, using the Matlab "hot" false-color scheme, which is based on blackbody radiation (black, red, orange, yellow, white in that order); see Figure 1. A virally "hot" keyframe shows up as white, and an infrequently used one as black. The visualization uses a fixed size box for each video, and shows within the box, as vertical stripes, all of the near-dup keyframes detected in temporal order within those videos with at least 15K views; each box has a kind of cytogenic banding pattern. Even though this encoding is many-to-one (2.5K near-dup classes into 64 false colors), a video with many viral near-dups should show up as a box that is close to all white: see the close-up in Figure 2.

This visualization suggests three observations. First, any news summary videos already present in the ecology show up immediately as white boxes (which are actually visible in Figure 1 under magnification). Secondly, mashups show up as a confluence of inheritance arrows onto a box with a pronounced striped pattern, reflecting the video's alternation of many "hot" near-dups with "cold" unpopular interjected title slides and editorializing frames. Third, many extended near-dup sequences have been transmitted essentially in their entirety.

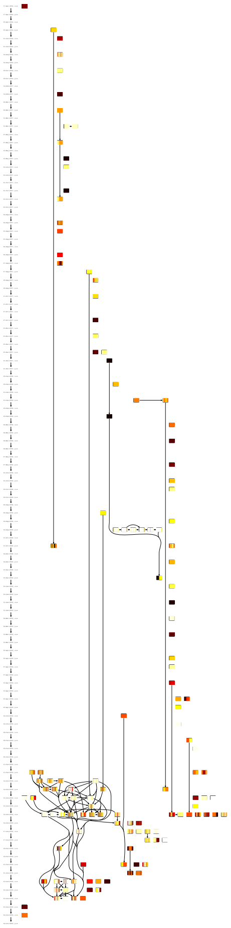


**Figure 1: A visualization of the inheritances of all test set videos with more than 15000 views, displayed as boxes against a timeline indicating their first posting time; time goes down. Arcs trace the recurrence of one or more near-duplicate frames. Near-dups of a video are color-encoded as vertical stripes within the boxes according to viral "hotness". Isolated videos tend to be off-topic. Note two posting clusters, corresponding to the events of a demonstration and of a sniper shooting, with much copying between them.**

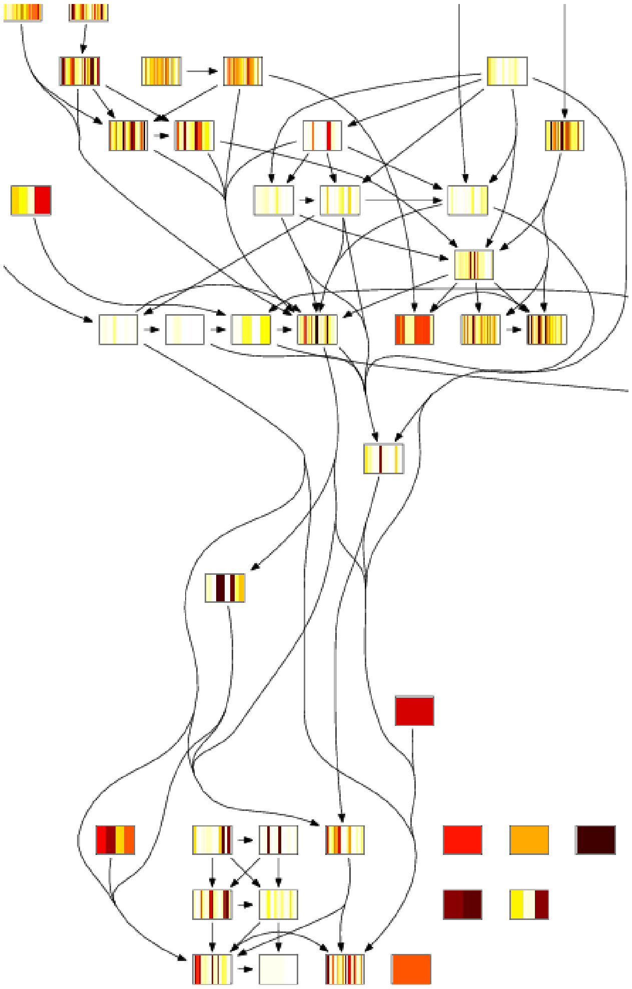


**Figure 2: A closeup of the two major Iran events. A newscast summary video appears as the white box in the lowest row. A mashup appears as the striped box with large in-degree, in the middle of the row of seven boxes about one-third down.**

## 4. VIDEO SPECIE DETAILS

### 4.1 Full Video Similarity via Smith-Waterman

Since a full video is now represented by a sequence of near-dup class numbers, two videos can be be compared using an extension of the Smith-Waterman algorithm of genetics [9], which is a simplified and discretized form of dynamic time warping that assigns penalties to component mismatches and to sequence gaps. Because the algorithm works on a derived sequence of integers (and not on features, images, or clips), it is fast. An example of its use in aligning two videos is given in Figure 3, which shows before and after alignments, where the near-dup classes have been assigned false colors arbitrarily for illustration purposes; gaps and mismatches are perceived instantly. Since many long videos are reposted with only a few added editorializations at their beginnings or trimmed credits from their endings, they are especially informative. Any gaps within matched pairs of them tend to indicate a failure of the keyframe detector, and any mismatches tend to indicate an overgranularity of keyframe clustering. (A similar observation can be made about the internal self-matching of talkshow and slideshow videos.) These inaccuracies can provide very useful non-manual ground truth feedback for tuning the sensitivities of the earlier algorithms that detect and cluster the near-duplicates.

As a bonus, the Smith-Waterman match induces a true distance metric on the videos. The total number of gaps necessary to form a minimal cost match between a pair of sequences is in fact positive-definite, symmetric, and subadditive. We can therefore create an induced distance matrix of a video ecology in order to study the sparsity of its interconnectivity. Both visualization and direct computation indicate that in our case study dataset, interrelationships are on the order of only 1%. This is illustrated in Figure 4, which uses the heuristic Reverse Cuthill-McKee algorithm to

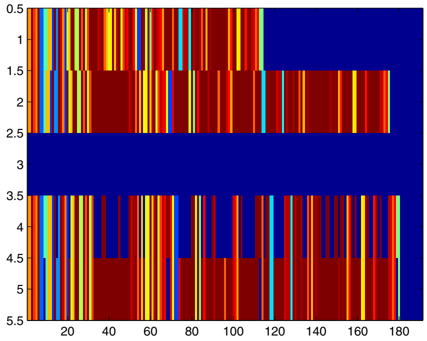


**Figure 3: Rows 1 and 2: Two similar videos, with near-dups classes assigned random false-colors to demonstrate video similarity. Rows 4 and 5: Same two videos aligned using Smith-Waterman.**

permute the interconnection matrix to dramatize its sparsity. This means that in general, despite much inheritance, any two specific videos retrieved using a query are unlikely to be related to each other, simply given the large number of videos in the total collection. The YouTube ecology may not distribute views equitably, but it certainly is diverse.

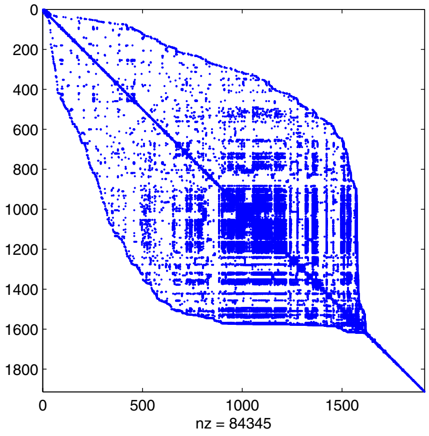


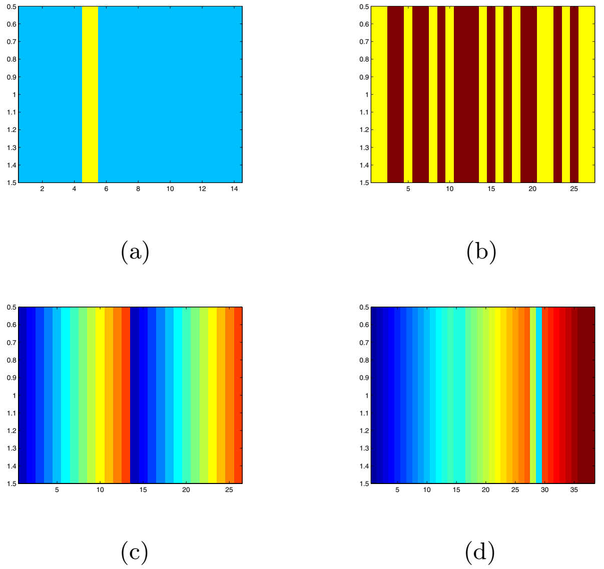
**Figure 4: The graph of video genetic relationships is very sparse; their interconnection matrix shows a density of 1% and a narrow "bandwidth".**

## 4.2 Genre via Motifs

We have noted that by observing the patterns of near-dup repetitions *within* a video, it is often easy to tell the video genre, regardless of near-dup content, by using what geneticists call motifs. We illustrate this with an example that is very roughly the equivalent of gel electrophoresis in genetics.

We represent the video as a sequence of near-dup class numbers, which are integers. By applying a degenerate form of Lempel-Ziv encoding [11] to this sequence, we can capture the periodicity of the content. For example, if the video is of the form $ABAABABBA$, where $A$ and $B$ are two (large, positive) near-dup class numbers, this sequence can be compressed losslessly into 002132213, where each (small, non-negative) code integer captures the LZ "distance" of each original integer to its most recent previous occurrence. Ignoring any of the 0s, which indicate new content, this LZ code can now be histogrammed to find peaks, which will corresponding to length of the periods of repetition of content. In the example, this histogram would be 232, corresponding to 2 instances of the code of 1, 3 of the code of 2, and 2 of the code of 3.

Except for some expected blurring, we have found that videos consisting of a single static frame or of a monologue—which is almost the same thing–have LZ code histograms with modes at 1; dialogues such as host-plus-guest talk shows have histogram modes at 2 (as in the example above); slideshows accompanying music or spoken text have modes at $P$, where $P$ is the number of independent slides that are cycled through; and news programs and mashups have no discernible period at all, since their codes are mostly 0s. These patterns also easily show up, of course, in false-color displays, as shown in Figure 5. One preliminary observation is that viral near-dups tend *not* to come from videos with periodicity; they tend to originate from news.



(a)

(b)

(c)

(d)

**Figure 5: False color representations of near-dup classes for videos of different genres: (a) Monologues basically repeat with period 1, (b) Dialogues have a period of 2; (c) Slide shows have a long period; (d) News and mashups have very little repetition and have no period.**

# 5. SPECULATION AND FUTURE WORK

This is work in progress and many areas are under investigation; the following is a list.

Color correlogram data are only an approximation to content, and are known to oversegment during camera movements; some object recognition would be helpful. Little is known about the distribution of near-dups across an ecology; in our data sets, they do seem to follow a Zipf-like power law, but with an exponent of about −2/3, that is, with a longer and more diverse tail than Zipf expects. Just as amino acids are grouped into families with common functional roles, near-dups probably occur in families that have common semantic meanings (e.g., commentator, crowd, outdoors, establishing shot, etc.). Inheritance can and should be defined more realistically as occurring due to multiple inheritances within a temporal window, rather tahn through single near-dup transmission. Although for some topics a video summary may already exist in the ecology, as a topic drifts, it may become necessary to construct one instead, probably through some disciplined automatic mashup algorithm. Multiple sequence alignment is provably NP-hard, with $V$ videos of length $L$ taking $O((2L)^V)$ time; so, various heuristic approaches, like the Clustal family of algorithms in genetics will be necessary to find and verify all near-duplicate *videos*. Basing a taxonomy of genres solely on period is primitive; there are probably better grammar-based means [1] derived from constraints on how one must tell a story, which will require a kind of compiler to parse near-dups into their functional roles, and sequences of them into composite meaning. The amount of "genetic drift" (topic drift) probably varies from genre to genre, but it may be near-constant within one. We have in fact already noted that every news video inheritance stream appears to evolve faster than any music video stream does. Of course, much of this research is probably dependent of other variables of the metadata (topic, genre, source, etc.), so therefore any universal principles about analysis or display can only be inferred after many case studies. We have in fact noticed some virality differences between "acute" topics (like the troubles in Iran) and "chronic" ones (like the swine flu epidemic): the latter appear to evolve more slowly. The holy grail, of course, is to determine which, if any, characteristics of a near-dup are predictive of longevity and virality; we suspect that, as in genetics, the answer will be multifactorial.

## 6. ACKNOWLEDGMENTS

The first author was partially supported by IBM Research, while on sabbatical leave there.

## 7. REFERENCES

[1] D. Arijon. *Grammar of the Film Language*. Silman-James Press, 1976.

[2] R. Colbaugh, K. Glass, and P. Ormerod. Predictability and Prediction for an Experimental Cultural Market. In *Advances in Social Computing*, volume 6007, pages 79–86. Springer Berlin, 2010.

[3] R. Crane and D. Sornette. Viral, Quality, and Junk Videos on YouTube: Separating Content From Noise in an Information-Rich Environment. In *Proceedings of the AAAI Symposium on Social Information Processing*, pages 18–20, March 2008.

[4] C. Gini. Measurement of Inequality of Incomes. *Economic Journal*, 31:124–126, 1921.

[5] J. L. Iribarren and E. Moro. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters*, 103(3):038702–1–038702–4, July 2009.

[6] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-Tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery nd Data Mining (KDD'09)*, pages 497–506. ACM, 2009.

[7] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, New York, NY, USA, 2003. ACM.

[8] M. Salganik, P. Dodds, and D. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, 2006.

[9] T. Smith and M. Waterman. Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[10] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! Modeling Contagion through Facebook News Feeds. In *Proceedings of the Third International Conference on Weblogs and Social Media*. AAAI Press, May 2009.

[11] J. Ziv and A. Lempel. A Universal Algorithm for Sequential Data Compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.