Dependence Structure Analysis Of Meta-level Metrics in YouTube Videos: A Vine Copula Approach

Vikram Krishnamurthy, Fellow, IEEE      Yan Duan

Abstract—This paper uses vine copula to analyze the multivariate statistical dependence in a massive YouTube dataset consisting of 6 million videos over 25 thousand channels. Specifically we study the statistical dependency of 7 YouTube meta-level metrics: view count, number of likes, number of comments, length of video title, number of subscribers, click rates, and average percentage watching. Dependency parameters such as the Kendall’s tau and tail dependence coefficients are computed to evaluate the pairwise dependence of these meta-level metrics. The vine copula model yields several interesting dependency structures. We show that view count and number of likes are in the central position of the dependence structure. Conditioned on these two metrics, the other five meta-level metrics are virtually independent of each other. Also, Sports, Gaming, Fashion, Comedy videos have similar dependence structure to each other, while the News category exhibits a strong tail dependence. We also study Granger causality effects and upload dynamics and their impact on view count. Our findings provide a useful understanding of user engagement in YouTube.

I. INTRODUCTION

Since its establishment in 2005, YouTube has over 1 billion registered users, who spend millions of hours and generate billions of views everyday; more than 300 hours of video content are uploaded every minute [1]. While YouTube is a social media site, is is also a social networking site. Classical online social networks (OSNs) are dominated by user-user interactions. However YouTube is unique in that the interaction between users in the YouTube social network is incentivized using the posted videos. These interactions include:

1) Commenting on users’ videos and commenting on other users comments which is very similar to users interaction on blog posting sites such as Twitter.
2) Subscribing to YouTube channels provides a method of forming relationships between users.
3) Users can directly comment on a YouTube channel.
4) Users can also interact by embedding videos from another users channel directly into their own channel to promote exposure or form communities of users.

In this paper we view YouTube as a big-data time series. We statistically analyze a massive YouTube dataset of more than 6 million videos over 25 thousand channels to understand how various YouTube meta-level metrics (such as view count, subscriber count, number of likes, etc) affect each other. YouTube is a useful source of data: with efficient web crawlers, meta-level metrics data such as “number of views” can be collected for millions of videos. Careful analysis of YouTube data can reflect users’ preference and lead to useful outcomes in optimal caching [3], recommendation systems [4],[5] and targeted advertising [6]. For example [3] studies how popularity prediction of YouTube can be used to design efficient caching algorithms in 5G telecommunication systems.

A. Context. Copulas

With the rapid growth of YouTube data sets, there is strong motivation for adequately understanding and modeling the dependencies present in the resulting multivariate random processes. The modeling of a multivariate distribution function can be decomposed into considering the marginal distributions and then determining the underlying dependence structure, the so-called copula. One of the most popular copula classes, especially for high-dimensional data, are vine copulas.

In this paper we focus on the correlations amongst YouTube video metrics and their dependence relationships, which are determined using vine copula models. Vine copula models [8] have several attractive features: since they do not impose constraints on marginal distributions, they can capture unusual correlations such as extreme co-movements [9]. Furthermore, with vine copula, one can separate the task of estimating marginal distributions from that of estimating dependence between random variables. A short review of bivariate copulas and vine copula model is provided in the Appendix.

B. Main Results

This paper uses vine copula to determine the multivariate dependency structures in a YouTube dataset consisting of more than 6 million videos over 25 thousand channels. The dataset consists of the daily samples of meta-level video metrics from April, 2007 to May, 2015. Specifically we consider 7 YouTube meta-level metrics: view count, number of likes, number of comments, length of video title, number of subscribers, click rates, and average percentage watching.

The vine copula model is useful for visualizing the interdependencies of the YouTube metrics across different channels.
Also, the Kendall’s Tau yields the relative importance of these metrics. Based on the vine copula analysis of the YouTube dataset, we conclude several important facts regarding dependence structures (Sec.III-E describes these dependencies in more detail) of YouTube metrics.

1) We found that “number of views” and “number of likes” are in the central position of the dependence structure. Conditioned on these two metrics, the other five YouTube metrics are virtually independent of each other (see Figure 2). So for example, conditioned on the number of likes, the number of comments (negative or positive) is statistically independent of the number of subscribers (which at first sight is somewhat counter-intuitive). Surprisingly, this dependency structure is true for all categories of videos that we considered, namely, gaming, sports, fashion and comedy.

2) Regarding the various categories of videos; “News” videos have stronger statistical inter-dependence when their values are large (i.e. positive extreme co-movements). For “Gaming” videos, “annotation clicks rate” is less dependent on number of views compared to other categories, Users watching “Comedy” videos are more likely to give comments. For a video with certain popularity, users are more likely to finish a video in “Fashion” category than in other categories.

3) Finally, given the dependency structure from the vine copula, we dig further into the dynamics of YouTube. We use Granger causality to determine the causal relation between viewcounts and subscribers for channels in YouTube. We also study the upload dynamics of YouTube and find an interesting property that for popular gaming YouTube channels with a dominant upload schedule, deviating from the schedule increases the views and the comment counts of the channel.

The vine copula models in this paper are of interest in interactive advertising: based on the meta-level metrics, which channels to advertise on? Also, knowledge of the dependency structure allows YouTube content providers to maximize the number of views. The interaction between users in the YouTube is incentivized using the posted videos. In addition to the social incentives, YouTube also gives monetary incentives to promote users increasing their popularity. As more users view and interact with a users video or channel, YouTube will pay the user proportional to the advertisement exposure on the users channel. Therefore, users not only maximize exposure to increase their social popularity, but also for monetary gain. Finally, in 5G wireless networks adaptive caching of content is crucial: the vine copula gives a useful multivariate model in various fields: copula data by using probability integral transformation can be transformed into uniformly distributed data, or the so-called “copula data” by using probability integral transformation.
edge represents the bivariate copula connecting one pair of marginals; see appendix for an illustrative example.

Formally a regular vine structure is defined as follows (see Appendix for background):

**Definition.** A graphical structure $V$ is a regular vine of $m$ elements if:

1) $V = (T_1, ..., T_{m-1})$ and all trees are connected
2) The first tree $T_1$ has node set $N_1 = 1, ..., m$ and edge set $E_1$; then for the next trees $T_i$, $i = 2, ..., m-1$, $T_i$ has the node set $N_i = E_{i-1}$
3) Proximity condition: If two nodes are connected by an edge in the $i + 1^{th}$ tree, then the two edges in $i^{th}$ tree corresponding to these two nodes share a node.

The following steps (see appendix for details) construct a regular vine copula (R-vine) specification to a dataset:

1) Construct the R-Vine structure by choosing an appropriate unconditional and conditional pair of metrics to use for vine copula model.
2) Choose a bivariate copula family for each pair selected in step (1).
3) Estimate the corresponding parameters for each bivariate copula.

### III. Dependence Structure of YouTube using Vine Copula

In this section, we use the vine copula model to unravel the dependence structure of a massive YouTube data set.

#### A. YouTube Dataset

The dataset contains daily samples of metadata of YouTube videos from April, 2007 to May, 2015, and has a size of around 200 gigabytes. Table I summarizes the dataset.

| Videos | 6 million |
| Channels | 26 thousand |
| Average number of videos (per channel) | 250 |
| Average age of videos | 275 days |
| Average number of views (per video) | 10 thousand |

The data-set comprises seven meta-level metrics and five categories of YouTube videos. The 7 meta-level metrics of YouTube:

- Number of views
- Number of likes
- Number of comments
- Video title length
- Number of subscribers
- Annotation click rates
- Average percentage of watching

The YouTube videos we consider belong to five categories:

- Sports
- Fashion
- Gaming
- Comedy
- News

For each category, 6000 samples including data of seven metrics are chosen from the YouTube dataset, which has different number of videos for different categories. For the reader’s convenience, the abbreviations of long metrics names are also provided. All of above information is summarized in Table III.

Table III conducts a preliminary statistical analysis on the dataset in terms of mean, median, kurtosis and skewness. It can be concluded that apart from “title length” and “average percentage of watching”, the other five metrics of YouTube videos have very high kurtosis and skewness, which imply heavy tails and high asymmetry of those distributions. Therefore, these five metrics of YouTube videos are not normally distributed. It is this property that motivates the use of vine copula: the vine copula does not require the marginals to be normally distributed, it serves as a useful multivariate model for the dependence analysis of YouTube dataset.

#### B. Modelling Marginal Distributions

The marginal empirical distribution of each YouTube metric can be estimated from the data:

$$F_N(x) = \frac{1}{N+1} \sum_{i=1}^{N} I(X_i \leq x).$$

where $N$ is the number of data samples, $X_i$ is the $i^{th}$ data point, while $I(\cdot)$ is the indicator function. The probability integral transform (PIT) is then applied to the estimated marginal distribution to obtain uniformly distributed samples in [0,1], which is referred as “copula data” and required by the vine copula model. In order to test whether the copula data is uniformly distributed in [0,1], the Kolmogorov-Smirnov statistic [16] is used to do the goodness-of-fit test (i.e. how close the transformed copula data is to the reference uniform distribution) [17].

The high p-values in Table IV indicate that the copula data generated from PIT is uniformly distributed.

#### C. Measure of Dependence

For each pair-copula, instead of the classical correlation coefficient, the Kendall rank correlation coefficient (all called Kendall’s tau) [18] is a useful measure of dependence. The classical correlation coefficient for two random variables $X, Y$:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

only detects linear dependencies, and it is not preserved by copulas. Therefore is not considered. Unlike the classical correlation coefficient, the Kendall’s tau is preserved by copulas [19].

Given a dataset with $N$ observations and seven meta-level metrics, there are $\binom{7}{2} = 21$ pairs of metrics in total; an estimate of dependency for $k^{th}$ pair of metrics $(X, Y)$ is obtained by the empirical Kendall’s tau coefficient:

$$\tau_{em}(X, Y) = \frac{A_N^k - B_N^k}{N(N-1)/2}.$$  

Kendall’s tau is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities. The rank is high when the variables have a similar rank, and is low when the ranks are dissimilar.
| Number of Videos | Categories | Metrics | Abbreviation |
|------------------|------------|---------|--------------|
| 85657            | News       | Number of Views | Views |
| 90586            | Fashion    | Number of Likes | Likes |
| 60438            | Sports     | Number of Comments | Comm |
| 431803           | Gaming     | Video Title Length | TitLen |
| 58820            | Comedy     | Number of Subscribers | Subs |
| Null             | Null       | Annotation Clicks | Anno |

**TABLE II: YouTube Video Data Categories, Metrics and Abbreviations**

**TABLE III: Summary of YouTube data**

(A): News

| Stats | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|-------|-------|-------|------|--------|------|------|--------|
| Min   | 1     | 0     | 0    | 4      | 0    | 0    | 0      |
| Max   | 5775400 | 12111 | 9035 | 120 | 9793 | 35.78 | 99.03 |
| Median| 715   | 2     | 1    | 50    | 0    | 0    | 56.40  |
| Mean  | 25540 | 124.50| 71.43| 50.66 | 26.62| 0.27| 50.77 |
| Kurtosis | 9.59 | 11.44 | 7.83 | 0.23 | 9.81 | 10.41 | 0.74 |
| Skewness | 1.61 | 1.04 | 1.27 | 0.21 | 2.15 | 1.25 | -0.43 |

(B): Comedy

| Stats | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|-------|-------|-------|------|--------|------|------|--------|
| Min   | 1     | 0     | 0    | 4      | 0    | 0    | 0      |
| Max   | 4355984 | 134103 | 22932 | 120 | 23338 | 408.764 | 99.04 |
| Median| 1385  | 19    | 9    | 36     | 1.00 | 0.0  | 40.17  |
| Mean  | 50264 | 487.7 | 95.64| 38.69  | 54.36| 2.096| 40.43  |
| Kurtosis | 14.32 | 13.46 | 11.82 | 0.041 | 13.30| 5.82 | 0.042  |
| Skewness | 5.05 | 3.18 | 2.91 | 0.096 | 3.68 | 2.00 | 0.57 |

(C): Sports

| Stats | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|-------|-------|-------|------|--------|------|------|--------|
| Min   | 1     | 0     | 0    | 3      | 0    | 0    | 0      |
| Max   | 5516714 | 12361 | 5963 | 120 | 1589 | 109.1267 | 92.07 |
| Median| 1772  | 11    | 6    | 54     | 1.00 | 0.0  | 39.71  |
| Mean  | 17298 | 110.3 | 42.29| 55.25  | 13.60| 1.198| 40.34  |
| Kurtosis | 12.15 | 12.30 | 6.14 | 0.21 | 2.30 | 1.26 | 0.05  |
| Skewness | 3.10 | 1.52 | 2.14 | 0.10 | 1.30 | 0.90 | 0.066 |

(D): Fashion

| Stats | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|-------|-------|-------|------|--------|------|------|--------|
| Min   | 1     | 0     | 0    | 3      | 0    | 0    | 0      |
| Max   | 6900910 | 22846 | 5094 | 120 | 7553 | 176.8490 | 98.53 |
| Median| 1287  | 14    | 8    | 45     | 2.00 | 0.0  | 28.73  |
| Mean  | 30666 | 128   | 36.18| 45.05  | 36.76| 1.0688| 29.70  |
| Kurtosis | 4.44 | 5.63 | 4.85 | 0.037 | 4.70 | 3.65 | 0.03  |
| Skewness | 1.83 | 1.87 | 1.79 | 0.062 | 1.30 | 1.62 | 0.049 |

(E): Gaming

| Stats | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|-------|-------|-------|------|--------|------|------|--------|
| Min   | 0     | 0     | 0    | 2      | 0    | 0    | 0      |
| Max   | 3233762 | 28314 | 9731 | 120 | 7405 | 91.4822 | 98.36 |
| Median| 341   | 6     | 4    | 54     | 0.00 | 0.0  | 26.72  |
| Mean  | 7380  | 121   | 39.22| 54.37  | 12.02| 0.6450| 30.21  |
| Kurtosis | 12.16 | 4.94 | 5.99 | 0.027 | 13.17| 2.52 | 0.063 |
| Skewness | 3.21 | 1.96 | 2.29 | 0.021 | 3.82 | 1.40 | 0.11  |
Fig. 1: The scatter plots (lower matrix) and empirical Kendall’s tau (upper matrix) for the copula data of “News” YouTube videos. The empirical Kendall’s tau is estimated from data using Table II, while the scatter plots provide qualitative information regarding the pair-wise dependence. For instance, plots of “titlen” and “AvgPer” with other metrics are scattered everywhere, this implies weak dependence of these two metrics on the other metrics. On the other hand, scatter plots of “Views”, “Likes” and “Subs” show distinct patterns of co-movements with each other, this implies their strong dependence on each other, and might be in central position in the dependence structure.

where \( A_k^N \) and \( B_k^N \) are the concordance and disconcordance number \([20]\) of \( k^{th} \) pair that has \( N \) data points. The concordance number and the disconcordance number count the data points that move in the same direction and opposite direction, respectively.

In contrast to the empirical Kendall’s tau, the theoretical Kendall’s tau, which is computed using bivariate copulas\([21]\):

\[
\tau_{th}(X, Y) = 4 \int_{u=0}^{u=1} \int_{v=0}^{v=1} C(u, v)dC(u, v) - 1. \tag{4}
\]

Here, \( C \) is the bivariate copula to be selected in the vine model, \( u \) and \( v \) represent the copula data of \( X \) and \( Y \), respectively.

The empirical Kendall’s tau and scatter plot for the copula data of “News” videos are shown in Fig.1. In order to compute empirical Kendall’s tau, 6000 “News” videos are randomly selected from 85657 videos over 3217 YouTube “News” channels, its theoretical Kendall’s tau values shown in Table V(A) are computed from the vine copula model. The empirical Kendall’s tau, which is computed from data, is required in the R-vine structure selection step for the pair copula construction of vine copula (reviewed in the Appendix); while theoretical Kendall’s tau is computed as the measure of dependence given by the vine copula model.

D. Summary of Main Conclusions

Based on the vine copula analysis of the YouTube dataset, we conclude several important facts regarding dependence structures of YouTube videos. (Sec. III-E describes these dependencies in more detail.)

1) Conditional dependence of the seven meta-level metrics of YouTube videos is insignificant, compared with a much stronger unconditional dependence.

2) Amongst the seven meta-level metrics of YouTube, “number of views” and “number of likes” are in the central position of the dependence structure. The other five metrics are much more dependent on these two metrics than on each other.

3) The “News” category has a different dependence structure from the other four categories. Furthermore, the metrics of “News” videos have stronger inter-dependence when their values are large (i.e. positive extreme co-movements), the other categories don’t possess this property.

4) For “Gaming” video category, “annotation clicks rate” is least dependent on number of views among all the five categories.

5) For “Comedy” video category, users are more likely to subscribe to the channel for a popular video than other categories.

6) Users watching “Comedy” videos are more likely to give comments.

7) For a video with certain popularity, users are more likely to finish a video in “Fashion” category than in other categories.

8) “Length of video title” is independent of the other metrics.

E. Results and Discussion

We are now ready to discuss the main conclusions on our dependency analysis of YouTube video meta-level metrics.

Once the empirical marginal distributions of YouTube video metrics are obtained using vine copula described in Sec. III-B, the three steps outlined in the Appendix were applied to model the multivariate distributions for YouTube video metrics. Table V summarizes the estimated parameters: Kendall’s tau \( \tau \), copula parameters and lower/upper tail dependence \( \lambda_L/\lambda_U \).
(defined in (19) in the Appendix), for the vine copula model of all five video categories. General dependence analysis for the five categories and category-specific conclusions are shown as follows.

1) General Dependence Analysis for All Five Categories: Fig.2 and Fig.3 are the main graphical dependency results of our analysis. They illustrate the dependence structure for all five video categories, with pair-wise theoretical Kendall’s tau values and bivariate copula names.

First, as described in the Appendix, a vine structure uses multiple dependence trees to describe the dependence structure, which consists of two types of trees:

- Unconditional dependence tree: the first tree of the vine
- Conditional dependence trees: all the trees of a vine excluding the unconditional dependence tree

However, in Fig.2 and Fig.3, we only use the first trees (i.e. unconditional dependence trees) to represent the dependence structure for YouTube video metrics. Unconditional/conditional dependence trees describe the dependence of two or more random variables when another random variable is present, and are induced through the decomposition of vine structure. For instance, in (15), given three random variables $X_1$, $X_2$, and $X_3$, copula density $c_{12}$ is the joint density of ($X_1$, $X_2$), and $c_{13|2}$ is the conditional density of $X_1$, $X_3$ given $X_2$. So $c_{12}$ describes the unconditional dependence of $X_1$, $X_3$, while $c_{13|2}$ describes the conditional dependence of $X_1$, $X_3$ given $X_2$. Table V shows that the unconditional dependencies for the seven metrics (Kendall’s tau values of 0.3-0.6) are much stronger than the conditional dependencies (Kendall’s tau values of 0.01-0.2). Therefore, the conditional dependence is insignificant for YouTube video metrics, and can be ignored.

Second, in Fig.2, it can be observed that for “Comedy”, “Sports”, “Fashion”, “Gaming” video categories, “number of views” and “number of likes” are in the central position of the dependence structure of the seven YouTube video metrics. More specifically, “number of comments”, “number of subscribers” and “annotation clicks” have a strong unconditional dependence on these two video metrics, with high Kendall’s tau values between 0.3 and 0.65. The exceptions are “title length” and “average percentage of watching”: “title length” shows little dependence on other metrics with Kendall’s tau values that are no more than 0.1, while “Average Percentage of Watching” shows a moderate dependence on other metrics with Kendall’s tau values between 0.14 and 0.25.

Third, given the central roles of “number of views” and “number of likes” in the dependence structure, we should assign dominant weights to them if we want to assess a video using these seven video metrics. Note that “number of views” and “number of likes” are important indicators for the popularity of a video. Since “title length” is statistically independent of “number of views” and “number of likes” (considering its Kendall’s tau values that are less than 0.1), one can conclude that changing video title length will not increase the popularity of video. As for “average percentage of watching”, which is an important indicator for video quality, surprisingly it is not highly dependent on the YouTube video viewcount.

2) Category-specific Dependence Analysis: By comparing the dependence structures of different categories in Fig.2, we can see that “Gaming”, “Sports”, “Fashion” and “Comedy” videos possess similar dependence structure. That is, “number of views” and “number of likes” are in the central position of the dependence structure, while other metrics are in the border of the dependence structure. The exception is the “News” category. As shown in Fig.3, “number of subscribers” for the “News” category also plays the same central role as “number of likes” and “number of views”, although the overall dependence of “News” category turns to be weaker (i.e. smaller Kendall’s tau values) than the other categories. This can be explained by the fact that, all the other four categories possess a strong characteristic of entertainment, which “News” doesn’t have much, thus a big difference of dependence structure should be expected.

Furthermore, by comparing the theoretical Kendall’s tau values among different video categories, more category-specific conclusions can be made. The “Gaming” video category has the smallest Kendall’s tau value (0.22) for the (Anno, Views) pair, while “Comedy” category has the largest $\tau$ for both the (Comm, Likes) pair and the (Subs, Views) pair. The largest $\tau$ value for the (AvgPer, Likes) pair belongs to the “Fashion” category. Therefore, YouTube viewers watching “Gaming” videos are more likely to accept the suggestions from the channel and watch more related videos; while “Comedy” channels should be a better place for video makers raising more subscribers and create active online community; if business looks for channels to put advertisements on, “Fashion” channels should have a better chance to make viewers look at their break-in advertisements.

Finally, from Table V, it can be observed that for unconditional dependence structures of video categories including “Comedy”, “Sports”, “Fashion”, “Gaming”, there are few values for lower and upper tail dependence. This implies that for these categories, YouTube video metrics don’t have many co-movements (i.e. small dependence) when their values are very large or very small. The exception is again the “News” category: with large values of upper tail dependence for four pairs of metrics shown in Table V(A), the “News” video metrics are expected to have stronger inter-dependence when the values of these metrics are large.

IV. Validation of Vine Copula Model Using the White Test

How good is the vine copula model of Sec. III in terms of determining the dependency structures in YouTube? Goodness-of-fit tests including the misspecification test, the information matrix ratio test, and the Rosenbatt’s transform test, have been introduced for the vine copulas over the last decade [22]. Considering that we used regular vine (R-vine) to model the dependence structure, below we use one type of misspecification test (we call it “White test”) for the goodness-of-fit test of R-vine. Proposed by White and enhanced by Schepsmeier, this test is shown to have excellent performance and power behavior [23].

The theorem of White states that, under the correct model, the negative Hessian matrix and outer product of gradient for
Fig. 2: The first dependence (unconditional dependence) trees of the vine copula for the “Comedy”, “Sports”, “Fashion”, “Gaming” YouTube video data, along with their theoretical Kendall’s tau values and pair-wise copula families. Second and further dependence trees are not provided, since we have shown in Table V that conditional dependence is not significant in vine structure of YouTube videos. It can be observed that “number of views” and “number of likes” are in the central position of the dependence structure.
| Category   | Views | Likes | Comm | TitLen | Subs | Anno | AvgPer |
|------------|-------|-------|------|--------|------|------|--------|
| News       | 0.79  | 0.97  | 0.72 | 0.91   | 0.98 | 0.86 | 0.91   |
| Sports     | 0.93  | 0.82  | 0.92 | 0.71   | 0.96 | 0.86 | 0.87   |
| Fashion    | 0.83  | 0.99  | 0.76 | 0.99   | 0.87 | 0.88 | 0.85   |
| Gaming     | 0.76  | 0.96  | 0.87 | 0.88   | 0.99 | 0.70 | 0.90   |
| Comedy     | 0.93  | 0.98  | 0.98 | 0.89   | 0.92 | 0.98 | 0.78   |

TABLE V: Estimated Parameters for the Vine Copula Model: bivariate copula types discussed in the Appendix, theoretical Kendall’s tau $\tau$, bivariate copula parameters, lower/upper tail dependence $\lambda_L/\lambda_U$ (defined in [19] in the Appendix)

(A): News

| Tree | Dependence Pair | Copula | $\tau$ | Parameters | $\lambda_L$ | $\lambda_U$ |
|------|-----------------|--------|--------|------------|-------------|-------------|
| 1    | (Views, Comm)   | S-Joe  | 0.098  | [1.20, null] | 0.22        | 0           |
| 1    | (Views, Likes)  | Joe    | 0.35   | [1.98, null] | 0           | 0.58        |
| 1    | (Views, Subs)   | Joe    | 0.35   | [1.95, null] | 0           | 0.57        |
| 1    | (Views, Avgper) | Joe    | 0.18   | [1.42, null] | 0           | 0.37        |
| 1    | (Views, Avgper) | S-BB8  | 0.16   | [1.77, 0.80] | 0           | 0           |
| 2    | (Titlen, Views) | BB8-270 | -0.081 | [-1.27, -0.91] | 0           | 0.55        |
| 2    | (Comm, Views)   | S-BB7  | 0.11   | [1.10, 0.16] | 0.12        | 0.013       |
| 2    | (Likes, Views)  | BB8    | 0.17   | [1.35, null] | 0           | 0.33        |
| 2    | (Anno, Views)   | Tawn-180 | 0.017  | [1.28, 0.020] | 0.015       | 0           |
| 2    | (Subs, Avgper)  | BB8-90 | -0.121 | [-1.72, -0.72] | 0           | 0           |

(B): Comedy

| Tree | Dependence Pair | Copula | $\tau$ | Parameters | $\lambda_L$ | $\lambda_U$ |
|------|-----------------|--------|--------|------------|-------------|-------------|
| 1    | (AvgPer, Views) | Tawn   | 0.14   | [1.78, 0.20] | 0.17        | 0           |
| 1    | (AvgPer, Views) | Tawn   | 0.12   | [0.26, null] | 0.07        | 0           |
| 1    | (Comm, Views)   | BB8    | 0.63   | [4.40, 0.98] | 0           | 0           |
| 1    | (Subs, Views)   | BB8    | 0.62   | [4.53, 0.96] | 0           | 0           |
| 1    | (Likes, Anno)   | BB8    | 0.30   | [1.95, 0.96] | 0           | 0           |
| 2    | (AvgPer, Comm)  | Frank  | -0.17  | [-1.58, null] | 0           | 0           |
| 2    | (Titlen, Views) | Frank  | -0.041 | [-0.37, null] | 0           | 0           |
| 2    | (Comm, Views)   | Tawn   | 0.12   | [1.30, 0.33] | 0.15        | 0           |
| 2    | (Views, Subs)   | Clayton| 0.13   | [0.3, null] | 0           | 0.10        |
| 2    | (Subs, Anno)    | BB8    | 0.0075 | [1.3, 0.84] | 0           | 0           |

(C): Sports

| Tree | Dependence Pair | Copula | $\tau$ | Parameters | $\lambda_L$ | $\lambda_U$ |
|------|-----------------|--------|--------|------------|-------------|-------------|
| 1    | (AvgPer, Views) | Frank  | 0.19   | [1.80, null] | 0           | 0           |
| 1    | (Comm, Views)   | BB8    | 0.58   | [3.94, 0.96] | 0           | 0           |
| 1    | (Likes, Views)  | BB8    | 0.55   | [4.19, 0.90] | 0           | 0           |
| 1    | (Subs, Views)   | BB8    | 0.53   | [3.47, 0.95] | 0           | 0           |
| 1    | (Views, Anno)   | BB8    | 0.31   | [2.65, 0.79] | 0           | 0           |
| 1    | (Titlen, Anno)  | Tawn-180 | 0.045  | [1.29, 0.099] | 0.06        | 0           |
| 2    | (AvgPer, Comm)  | Gaussian| -0.052 | [-0.082, null] | 0           | 0           |
| 2    | (Comm, Views)   | Tawn   | 0.035  | [1.11, 0.22] | 0           | 0.53        |
| 2    | (Likes, Subs)   | Tawn   | 0.17   | [1.38, 0.54] | 0           | 0.25        |
| 2    | (Subs, Anno)    | BB8    | 0.12   | [1.46, 0.88] | 0           | 0           |
| 2    | (Views, Titlen) | Student t | 0.0036 | [0.0051, 9.92] | 0.0073      | 0.0073      |
Fig. 3: The first dependence (unconditional dependence) trees of the vine copula for the “news” category, with its theoretical Kendall’s tau values and pair-wise copula families. It can be observed that different from the other four categories, in the dependence structure of the “News” category, “number of subscribers” is as important as “number of views” and “number of likes”. Although the overall inter-dependence of video metrics in “News” are weaker than in the other four categories.

(D): Fashion

| Tree | Dependence Pair          | Copula | $\tau$ | Parameters | $\lambda_L$ | $\lambda_U$ |
|------|--------------------------|--------|--------|------------|-------------|-------------|
| 1    | (Likes, Comm)            | BB8    | 0.50   | [3.25, 0.95] | 0           | 0           |
| 1    | (Views, TitleLen)        | Survival BB7 | 0.11   | [1.13, 0.12] | 0.15        | 0.003       |
| 1    | (Likes, AvgPer)          | Survival BB8 | 0.25   | [2.06, 0.85] | 0           | 0           |
| 1    | (Views, Likes)           | BB8    | 0.53   | [3.25, 0.97] | 0           | 0           |
| 1    | (Views, Subs)            | BB8    | 0.55   | [3.46, 0.95] | 0           | 0           |
| 1    | (Views, Anno)            | BB8    | 0.32   | [2.68, 0.81] | 0           | 0           |
| 2    | (Comm, Views|Likes)        | Tawn   | 0.10   | [1.25, 0.30] | 0           | 0.13        |
| 2    | (TitleLen, Likes|Views)  | BB8    | -0.053 | [-1.22, -0.81] | 0           | 0           |
| 2    | (AvgPer, Views|Likes)  | Tawn-90 | -0.029 | [-1.17, 0.093] | 0           | 0           |
| 2    | (Likes, Subs|Views)   | BB8    | 0.27   | [3.12, 0.67] | 0           | 0           |
| 2    | (Subs, Anno|Views)   | Tawn   | 0.068  | [1.17, 0.33] | 0           | 0.1         |

(E): Gaming

| Tree | Dependence Pair          | Copula | $\tau$ | Parameters | $\lambda_L$ | $\lambda_U$ |
|------|--------------------------|--------|--------|------------|-------------|-------------|
| 1    | (AvgPer, Likes)          | BB8    | 0.16   | [2.22, 0.65] | 0           | 0           |
| 1    | (Comm, Likes)            | Joe    | 0.50   | [2.82, null] | 0           | 0.72        |
| 1    | (Likes, Views)           | BB8    | 0.51   | [2.96, 0.99] | 0           | 0           |
| 1    | (Subs, Views)            | Joe    | 0.44   | [2.42, null] | 0           | 0.67        |
| 1    | (Views, Anno)            | BB8    | 0.22   | [1.59, 0.97] | 0           | 0           |
| 1    | (TitleLen, Anno)         | Gaussian | 0.037 | [0.06, null] | 0           | 0           |
| 2    | (AvgPer, Views|Likes)  | Student t | -0.034 | [-0.054, 7.19] | 0.016      | 0.016       |
| 2    | (Comm, Views|Likes)  | Student t | 0.13   | [0.21, 16.5] | 0.003      | 0.003       |
| 2    | (Likes, Subs|Views)  | Gumbel  | 0.17   | [1.21, null] | 0           | 0.23        |
| 2    | (Subs, Anno|Views)   | Frank   | 0.046  | [0.41, null] | 0           | 0           |
| 2    | (Views, TitleLen|Anno) | Tawn-180 | 0.014  | [1.11, 0.045] | 0.016      | 0           |
Fig. 4: The scatter plots for original and simulated “Sports” YouTube copula data. It can be observed that, the original YouTube copula data and the copula data simulated from vine copula model share a dependence pattern. This qualitatively shows that the vine copula model captures the dependence patterns accurately, such conclusion is validated by the high p-values of White test in Table VII.

Table VII: p-values of White test for YouTube copula data

| Categories | News  | Sports | Fashion | Gaming | Comedy |
|------------|-------|--------|---------|--------|--------|
| p-values   | 0.674 | 0.685  | 0.591   | 0.714  | 0.666  |

The p-values of the White test can be computed based on (4) and (5), the reader can refer to [23] for more details. Table VII lists the p-values of the White test for the vine copula model, given the YouTube dataset. Based on the high p-values of White test for all five YouTube video categories considered, the null hypothesis can not be rejected. Thus the vine copula model obtained from Sec.III-E is a reasonable model for the multivariate probability distribution of YouTube video metrics. Fig. 4 compares the pairwise scatter plots of YouTube copula data, along with the simulated data sampled from our vine copula model. A common dependence structure pattern can be observed in both plots, which show that the vine copula model captures the dependence structures of YouTube data accurately.
The number of subscribers will also increase the view count of new videos or increase the views of old videos to increase results. Illustrate the importance of channel owners to not...

Entertainment channels satisfy Granger causality while only widely as illustrated in Table 4. For example, if different channel categories are accounted for then the channels satisfy the hypothesis. For approximately (p-value = 0.05) relationship between the subscriber dynamics and the view count. Similar results also hold for the coefficients a_i. Hence, as expected, the previous day subscriber count and the previous day view count most influence the current subscriber count.

Given the above causal relation between subscriber count and view count, a natural question is: Is there a causal relationship between the subscriber dynamics and the view count dynamics? This is modeled using the hypothesis in (9). We use the Wald test with a confidence of 0.95 (p-value = 0.05) and found that approximately 55% of the channels satisfy the hypothesis. For approximately 55% of the channels...
that satisfy the AR model [8], the view count “Granger causes” the current subscriber count. An interesting property is that the percentage of channels that satisfy Granger causality vary widely depending on the channel category. For example, 67% of Sports channels and 60% of the Gaming channels satisfy Granger causality, while 80% of Sports channels satisfy Granger causality. These results illustrate the importance of channel owners to not only maximize their subscriber count, but to also upload new videos or increase the views of old videos to increase their channels popularity (via increasing their subscriber count).

B. Scheduling Dynamics

From the dataset, we selected video channels with a dominant upload schedule, as follows: The dominant upload schedule was identified by evaluating the periodogram of the upload times of the channel and comparing the highest value to the next highest value. If the ratio defined above is greater than 2, we say that the channel has a dominant upload schedule. From the dataset containing 25 thousand channels, only 6500 channels contain a dominant upload schedule. Some channels, particularly those that contain high amounts of copied videos such as trailers, movie/TV snippets upload videos on a daily basis. These have been removed from the analysis so as to concentrate on those channels that contain only user generated content.

We found that channels with gaming content account for 75% of the 6500 channels with a dominant upload schedule and the main tags associated with the videos were: “game”, “gameplay” and “videogame”. We computed the average views when the channel goes off the schedule and found that on an average when the channel goes off schedule the channel gains views 97% of the time and the channel gains comments 68% of the time. Thus channels with “gameplay” content with periodic upload schedule benefit from going off the schedule. This suggests that by deliberately going off schedule, gameplay channels can increase their view count.

VI. Conclusions

YouTube video metrics across thousands of channels form a big-data time series. In this paper, we conducted a data-driven dependence analysis for seven meta-level metrics of YouTube videos among five different categories, based on a YouTube dataset of over 6 million videos across 25 thousand channels. To unravel the dependency structures of these meta-level metrics, we constructed a vine copula model on the YouTube video categories: conditional dependence is insignificant in YouTube video metrics, number of views and number of likes are in the central position of dependence structure, and “News” category possess a stronger tail dependence than the other YouTube video categories.

Finally, we also studied the dynamics of YouTube in two ways. First, using Granger causality, we unravelled the causal dependence between subscribers and view count. Second, by studying the upload dynamics, we found that periodic content providers can increase their view count by going off schedule.

REFERENCES

[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Erfron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, “Big data: astronomical or genomical?” PLoS Biol, vol. 13, no. 7, pp. e1002195, 2015.
[2] D. G. Popescu, P. Crama, “Ad revenue optimization in live broadcasting” Management Science, vol. 62, no. 4, pp. 1145–1164, 2015.
[3] W. Hoiles, O. N. Gharehshiran, V. Krishnamurthy, N.-D. Dao, and H. Zhang, “Adaptive Caching in the YouTube content distribution network: A revealed preference game-theoretic learning approach,” IEEE Transactions on Cognitive Communications and Networking, vol. 1, no. 1, pp. 71–85, 2015.
[4] R. Zhou, S. Khemmarat, and L. Gao, “The impact of YouTube recommendation system on video views,” in Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. ACM, 2010, pp. 404–410.
[5] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston et al., “The YouTube video recommendation system,” in Proceedings of the Fourth ACM Conference on Recommender Systems. ACM, 2010, pp. 293–296.
[6] J. R. Zhang, Y. Song, and T. Leung, “Improving video classification via YouTube video co-watch data,” in Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access. ACM, 2011, pp. 21–26.
[7] V. Krishnamurthy, A. Prempri and S. Bhatt, “Opportunistic advertisement scheduling in live social media: A multiple stopping time POMDP approach,” arXiv:1611.00291, 2016.
[8] F. Durante and C. Sempi, “Copula theory: an introduction,” in Copula theory and its applications. Springer, 2010, pp. 3–31.
[9] P. K. Trivedi and D. M. Zimmer, Copula modeling: an introduction for practitioners. Now Publishers Inc, 2007.
[10] H. Joe, “Dependence modeling with copulas”, CRC Press, 2014.
[11] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, “I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system,” in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM, 2007, pp. 1–14.
[12] J. Cheng, C. Dale, and J. Liu, “Statistics and social network of YouTube videos,” in IWQoS’08. 16th International Workshop on Quality of Service. IEEE, 2008, pp. 229–238.
[13] X. Cheng, J. Liu, and H. Wang, “Accelerating YouTube with video correlation,” in Proceedings of the First SIGMM Workshop on Social Media. ACM, 2009, pp. 49–56.
[14] T. Holfeld, R. Schatz, E. Biersack, and L. Plissonneau, “Internet video delivery in YouTube: From traffic measurements to quality of experience,” in Data Traffic Monitoring and Analysis. Springer, 2013, pp. 264–301.
[15] S. Krishnapa, M. Zink, C. Griiowod, and P. Halvorsen, “Cache-centric video recommendation: an approach to improve the efficiency of YouTube caches,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 4, p. 48, 2015.
[16] F. J. Massey Jr., “The kolmogorov-smirnov test for goodness of fit,” Journal of the American Statistical Association, vol. 46, no. 253, pp. 68–78, 1951.
[17] W. H. Lilliefors, “On the kolmogorov-smirnov test for normality with mean and variance unknown,” Journal of the American Statistical Association, vol. 62, no. 318, pp. 399–402, 1967.
[18] M. G. Kendall, “Rank correlation methods,” 1948.
[19] G. G. Venter, “Tails of copulas,” in Proceedings of the Casualty Actuarial Society, vol. 89, no. 171, 2002, pp. 68–113.
[20] C. Genest and A.-C. Favre, “Everything you always wanted to know about copula modeling but were afraid to ask,” Journal of Hydrologic Engineering, vol. 12, no. 4, pp. 347–368, 2007.
Archimedean copula includes several bivariate copulas such as Joe, Frank and Clayton copulas used in this work. Although there are over forty different types of bivariate copulas that are widely used, it is a harder problem to develop higher dimensional copulas. The pair-copula construction based on vine structure is a popular method of constructing a high dimensional copula [28].

B. Pair Copula Construction

For convenience, we illustrate the construction process for three random variables scenario, based on which a general n-variable case can be easily obtained. The density of a 3-dimensional copula function \( C \) is defined as:

\[
C_{123}(u_1, u_2, u_3) = \frac{\partial C_{123}(u_1, u_2, u_3)}{\partial u_1 \partial u_2 \partial u_3}.
\]

Then by (1), if we let \( u_1 = F_1(x_1), u_2 = F_2(x_2), u_3 = F_3(x_3) \), the density of joint distribution \( F(x_1, x_2, x_3) \) can be represented as:

\[
f(x_1, x_2, x_3) = f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \
\cdot c_{123}(F_1(x_1), F_2(x_2), F_3(x_3)).
\]

One possible decomposition from the chain rule of probability density gives:

\[
f(x_1, x_2, x_3) = f_1(x_1)f_2(2, x_2 | x_1)f_3|_{12}(x_3 | x_1, x_2).
\]

Note that similar to (12), the conditional probability densities can be represented as:

\[
f_{3|12}(x_3 | x_1, x_2) = c_{123}(F_1(x_1), F_2(x_2), F_3(x_3)).
\]

By combining (13) and (14), the multivariate probability density can be represented as multiplications of bivariate copula densities and marginal probability densities (recall the definitons of \( c_{12} \) and \( c_{13|2} \) in Sec. [III-E1]):

\[
f(x_1, x_2, x_3) = f_1(x_1)f_2(x_2)f_3(x_3) \
\cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \
\cdot c_{13|2}(F_1(x_1), F_3(x_3)).
\]

In order to obtain a systematic way to represent these decompositions, Bedford and Cooke [30] proposed the so-called “Regular Vine” (or simply “R-Vine”) graphical structure, hence the name “vine-copula”. “Vine” refers to a nested set of tree structures, where the edges of the \( i^{th} \) tree are the nodes of the \( i + 1^{th} \) tree. The formal definition of a regular vine structure is provided below.

Definition. A graphical structure \( V \) is a regular vine of \( m \) elements if:

1) \( V = (T_1, ..., T_{m-1}) \) and all trees are connected
2) The first tree \( T_1 \) has node set \( N_1 = 1, ..., m \) and edge set \( E_1 \); then for the next trees \( T_i, i \in 2, ..., m-1 \), \( T_i \) has the node set \( N_i = E_{i-1} \)

APPENDIX

Recalling the definitions of copula and Sklar theorem in Section [II], this appendix gives a short outline of vine copulas and their construction (which was used to analyze the YouTube dataset).

A. Bivariate Archimedean Copula

One widely used copula type is the “Archimedean” copula:

Definition. Bivariate Archimedean Copula is a \([0,1]^2 \rightarrow [0,1] \) function \( C \) with the form:

\[
C(u, v) = \phi^{-1}(\phi(u) + \phi(v)).
\]

where \( \phi(t) \) is called the generator function, which is a continuous, strictly decreasing function such that \( \phi(1) = 0; \theta \) is the single parameter to be estimated.

Bivariate archimedean copulas can be easily extended to arbitrarily high dimensions, but only bivariate ones are used for current copula applications, since it is just not a good idea that only one parameter is used to capture high dimensional dependence.
3) Proximity condition: If two nodes are connected by an edge in the \( i + 1 \)th tree, then the two edges in \( i \)th tree corresponding to these two nodes share a node.

In copula applications, the trees are used to represent the dependence structure of multiple variables: each edge represents the bivariate copula connecting one pair of marginals. For instance, the regular vine structure for three variables is shown in Fig. 6.

Once the uniformly distributed marginals are obtained, we can construct the vine copula model for a multivariate distribution. Dissmann et al. \[31\] summarizes an inference procedure for pair-copula construction of vine copula. The following steps are required to construct a regular vine copula (R-vine) specification to a given dataset:

1) Construct the R-Vine structure by choosing an appropriate unconditional and conditional pair of metrics to use for vine copula model.
2) Choose a bivariate copula family for each pair selected in step 1.
3) Estimate the corresponding parameters for each bivariate copula.

1) R-Vine Structure Selection: Dissmann et al. \[31\] suggested a sequential tree-by-tree selection method, where the first tree of the regular vine is selected to be the one with strongest dependencies, then the following trees are selected based on the same criteria. Such a heuristic approach is not guaranteed to capture the global optimum point. However, since the first tree of R-vine (unconditional dependence tree) has the greatest influence on the model fit, the sequential method has a reasonable trade-off between efficiency and accuracy. A maximum spanning tree (MST) algorithm is proposed for R-vine structure selection, where the spanning tree with the maximum sum of absolute empirical Kendall’s tau is selected \[31\]. To illustrate the dependence structure, the first trees of Vine copulas for the all the five categories of YouTube videos are shown in Fig. 2 along with their bivariate copula families and pairwise empirical Kendall’s tau values.

2) Copula Families Selection: After the R-vine structure is selected, the next step is to choose a suitable bivariate copula function for each dependence pair (edge of dependence tree) to fit the data. Due to its quick computation, the independence copula (simple product of two distributions) is first tested to determine if it is suitable to model dependence of two distributions.

According to Genest \[22\], under the null hypothesis of independence, Kendall’s tau is approximately normally distributed with mean of zero. With empirical Kendall’s tau value \( \tau \), we can choose independence copula using \( N \) observations if:

\[
\sqrt{\frac{9N(N-1)}{2(2N+5)}} |\tau| < 2. \tag{16}
\]

If the above independence test fails, then the Akaike Information Criterion (AIC) \[33\] is applied to choose the desired copula families. By computing the AIC for all possible families, the copula with smallest AIC will be selected:

\[
AIC = 2k - 2 \ln(M). \tag{17}
\]

where \( k \) is the number of estimated parameters of a copula (either 1 or 2 for bivariate copulas), and \( M \) is the maximum value among the likelihood functions for all copula candidates \[34\]. In this work, the bivariate copula families are chosen from 48 possible candidates.

3) Parameter Estimation: Finally, the corresponding parameters for the bivariate copula families \( \theta \), which includes \( \binom{n}{2} \) parameters for \( n \) metrics, can be estimated by using maximum likelihood estimation given data points \( U = (x_1, x_2, ... x_N) \):

\[
\theta^* = \arg\max L(\theta | U) = \arg\max \sum_{i=1}^{N} \sum_{j=1}^{n-1} \sum_{e \in E_j} \log C_{\theta_e}(F(x_{e1}), F(x_{e2})).
\]

where \( F(x_{e1}) \) and \( F(x_{e2}) \) are the two marginal distributions (nodes of a dependence tree and can be estimated from \( U \)) connected by copula (edge of a dependence tree) \( e \), while \( \theta_e \) is the copula parameter for edge \( e \). The associated likelihood function is the sum of log-likelihood of all bivariate copulas (i.e all edges \( E_j \) of tree \( j \)), for all \( n-1 \) trees of a vine over all \( N \) observations.

In addition to the regular dependency measure Kendall’s tau, the tail dependence measure can also be computed. The upper and lower tail dependence coefficients \( \lambda_U \) and \( \lambda_L \) are both defined using copula \[35\]:

\[
\lim_{u \to 0} \frac{1 - 2u + C(u, u)}{1 - u} = \lambda_U, \tag{19}
\]

\[
\lim_{u \to 0} \frac{C(u, u)}{u} = \lambda_L.
\]

\( \lambda_U \) and \( \lambda_L \) denote the probabilities that describe the simultaneous co-movements in the upper and lower tails of a bivariate distribution.