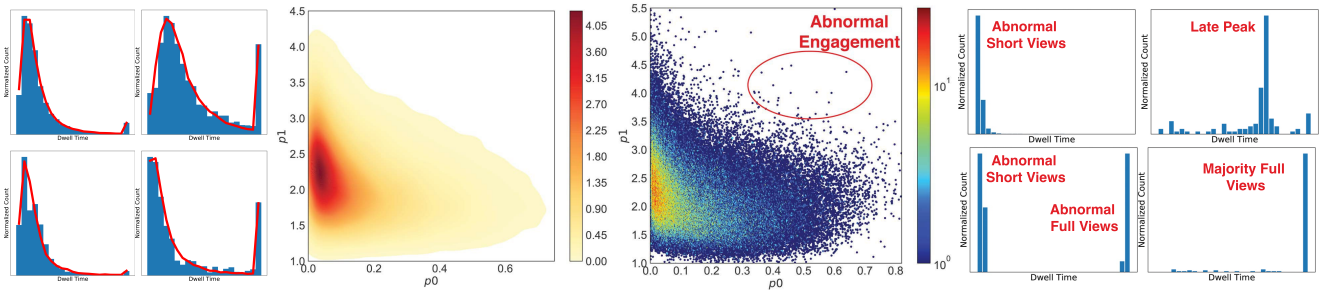
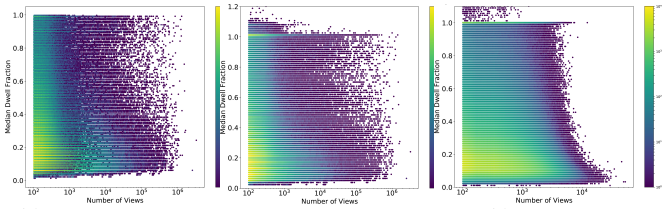


(a) Individual modeling (b) Group modeling (c) Anomaly detection (d) Dwell time anomalies

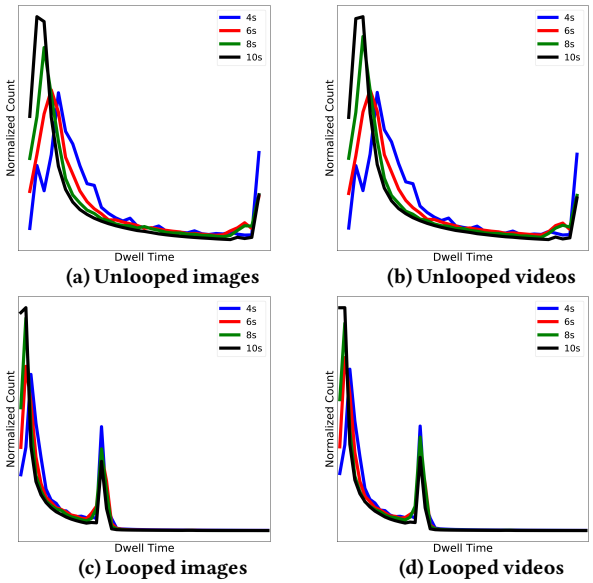
Figure 1: Our work discusses (a) state-of-the-art parametric models for individual sample dwell times which closely mirror empirical data, (b) flexible copula modeling of aggregated multivariate parameter fits, (c) utilization of aggregate models for detecting dwell time engagement anomalies which (d) reflect abnormal behaviors radically inconsistent with most samples.



(a) Unlooped Media (b) Looped Media (c) Viewers

Figure 2: Median dwell time ratios vs. number of views on (a) unlooped and (b) looped media, and (c) viewers show outliers which exhibit excessively high dwell times compared to normal engagement patterns of similar view-count peers.

- **C3. Anomaly Detection:** We Temonsrrate lo



(a) Unlooped images (b) Unlooped videos

(c) Looped images (d) Looped videos

Figure 3: Aggregated dwell time ratio statistics for varying media types and durations inform our modeling choices: treat images and videos similarly, and unlooped and looped content distinctly.

24-hour observation of engagement with all samples[1]. Table 1 details several key summary statistics of our dataset. All content samples and viewers have 100+ associated views/data-points, enabling us to draw reasonably reliable inferences about engagement.

4 INITIAL OBSERVATIONS

Before delving into details, we conduct several exploratory analyses to motivate and direct our approach and give intuition for our subsequent modeling choices.

Firstly, we aim to understand dwell time behavior across the entire dataset, to determine patterns and anomalies in dwell times across different content and viewers. Since different content samples have varying durations, we normalize all dwell times with respect to these in order to compare them. Henceforth, when we mention "dwell time," we consider instead the dwell time *fraction* or *ratio*. Thus, dwell time ratios of views on unlooped media must lie in $(0, 1]$, whereas dwell time ratios on looped content can lie on $(0, \infty)$.

Figure 2 shows quantized heatmaps of median dwell ratio of unlooped/looped media (2(a) and 2(b), respectively) and viewers (2c) versus view count, with brighter colors indicating logarithmically increasing density and darker colors denoting sparsity.Intuitively, sparsity increases towards the right of each plot due to skewed view count distributions, and towards the top of each plot, as few entities have high dwell ratios. Additionally, there are sparse entities in all plots which have very low dwell ratios. In all cases, we observe well-defined regions of high density.This suggests the following key observation, which motivates our modeling and anomaly detection goals.

Key Observation ((In)Consistencies in Visual Multimedia Dwell Times). *There exist patterns and anomalies in content and viewer dwell time engagement on visual multimedia.*

[1]Due to privacy reasons, we obscure certain sensitive details (timeframes and certain axes values) while communicating our insights.

Next, we consider collective differences between unlooped and looped media, and their implications for dwell time modeling. Figure 3 shows the collective dwell ratios across our entire dataset, for unlooped images and videos in 3(a-b), and their looped counterparts in 3(c-d). The stark differences in distribution shape is apparent; unlooped dwell ratios are effectively censored at 1.0, where they achieve a second peak after a tapered drop. However, while looped dwell ratios exhibit a similar decay and noticeable peak at the first view "completion," (near mid-plot) they show a decreasing but nonzero probability afterwards due to differences in feasible view duration across the media types. This suggests the following:

Observation 1 (Looped/Unlooped Media Dwell Time Disparity). *Looped and unlooped media require characteristically different dwell time models, due to the differences in support over dwell time ratios of* $(0, 1]$ *and* $(0, \infty)$*, respectively.*

Lastly, we consider the effect of different media type (image and video) on dwell ratios. By comparing Figures 3(a)/(c) with 3(b)/(d), we can observe that images and videos actually admit very similar dwell times. Despite videos being intuitively "richer" than images, the plots mirror each other. Moreover, since we observe no significant differences in the "stickiness" across the collective media type splits, we hypothesize that a significant portion of users' decision to engage with content may actually occur *before* the user accesses the content, for example due to self-selection and preferences towards certain content. Our major takeaway regarding media types is thus

Observation 2 (Image/Video Dwell Time Parity). *Dwell time similarities across image and video engagement suggest that they can be modeled characteristically similarly.*

Given these observations, we next discuss our proposed parametric models for dwell time distributions of individual content samples and viewers; parametric models are appealing due to their conciseness and interpretability over nonparametric alternatives.

5 INDIVIDUAL DWELL TIME MODELING

How can we parametrically model the dwell time distributions of multimedia content and viewers? In this section, we first propose "dwell processes" to generatively model the multimedia content for both looped and unlooped media. Following this, we posit the same contributions for viewers. In both cases, we give the intuition behind our modeling approaches, discuss efficient parameter inference procedures and validate against alternatives using goodness-of-fit metrics.

5.1 Multimedia Content Modeling

5.1.1 Looped Content. We begin by discussing modeling of looped content. Views on such content are unbounded, and dwell time ratios can range from $(0, \infty)$. Given our earlier insights regarding long-tailed dwell times from collective analysis in Figure 3, we consider several suitable distributions that may be able to model such shapes. In our preliminary analyses, we observed that the tails of many samples matched quite closely with Log-logistic distribution, defined as

Definition 5.1 (Log-logistic (LL) Distribution). Let T be a non-negative continuous random variable, such that $T \sim LL(\alpha, \beta)$. The PDF and CDF of T are given by

$$f_{LL}(t; \alpha, \beta) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{(1 + (t/\alpha)^{\beta})^2} \qquad F_{LL}(t; \alpha, \beta) = \frac{1}{1 + (t/\alpha)^{-\beta}}$$

where $t \in [0, \infty)$, and α(scale), $\beta > 0$(shape) are the parameters.

Note that the *LL* distribution admits the same support as our use-case for looped content, but does not do so for unlooped content. We propose using the original, unmodified *LL* distribution as the core of our LM-DP (**L**ooped **M**edia **D**well **P**rocess), which can be written generatively as

Definition 5.2 (Looped Media Dwell Process (LM-DP)). Sample each dwell time ratio $t_i \sim LL(\alpha, \beta)$.

Use of *LL* distribution over alternatives is justified for several reasons. *LL* is widely used in survival modeling and has a hazard function implying that the longer a view has persisted, the longer it will continue to do so [1]. Also, it has demonstrated success in modeling other real-world temporal phenomena [8, 13] besides visual multimedia dwell times, and as we will show below, it outperforms other candidate distributions in this task.

Inference of LM-DP. Inference of α and β cannot be computed in closed form. As a result, we infer parameters using the Nelder-Mead simplex method [15], which maximizes likelihood via iterative approximations, while converging quickly and accurately.

Validation of LM-DP. We validate the model both qualitatively (visually) in terms of empirical versus simulated dwell time probabilities, and quantitatively via the Kolmogorov-Smirnov (KS) 2-sample test. In Figure 5, we illustrate the strong match in empirical dwell time distributions and our superimposed model fits across several looped media samples of varying exposure durations, viewer counts and dwell time behaviors. For brevity, we show results only on 6 users, but most others exhibited similar quality of fit. Observe that LM-DP is able to well-approximate the peak and decay corresponding to view drop-offs reasonably well despite differences in distribution shapes across the samples, thus suggesting the appropriateness of our modeling choice.

To analyze the goodness-of-fit quantitatively, we perform KS tests comparing dwell times that were (a) empirically observed, with (b) those simulated by LM-DP using parameters inferred from MLE for each content sample. We compared LM-DP with four other alternative distributions which have previously been used for dwell time modeling in other contexts. These are CL-LN (Log-normal) [28], CL-IG (Inverse Gaussian) [11], CL-WB (Weibull) [16] and CL-G (Gamma) [14]. Figure 4(a) shows the sorted p-values reported across KS tests over samples reflecting the rejection probability for the null hypothesis H_0 that the empirical data and our simulated data are drawn from the same distribution. Assuming H_0 is true, the p-values should be uniformly distributed, manifesting as the 45° line. We observe that our proposed LM-DP using *LL* performs the best, with the CL-LN model the next closest, CL-IG/CL-WB and CL-G demonstrating significantly worse performance. Figure 4(b) further shows the percentage of samples that were fitted "successfully" (KS $p < .05$) given their view counts. Again, we observe that LM-DP outperforms competitors, modeling the vast majority of samples successfully (over 90% for

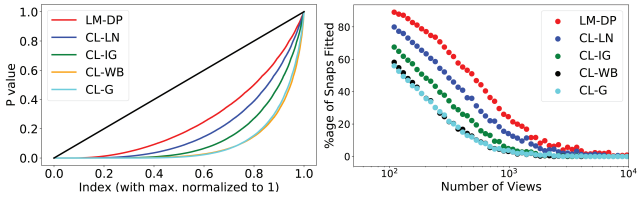


Figure 4: LM-DP outperforms alternatives:(a) sorted p-values from KS tests; the closer a model curve to the 45° line, better the fit. (b) %age of samples where model fits were successful($p < .05$).

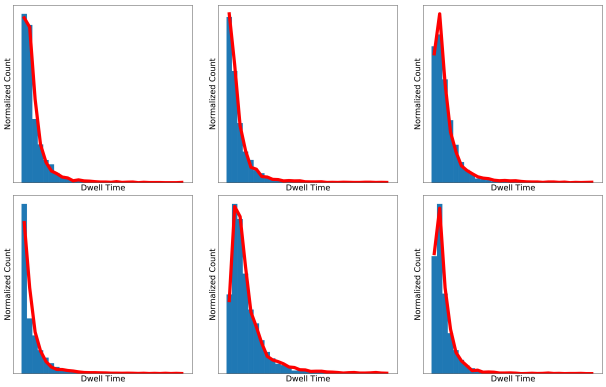


Figure 5: Proposed LM-DP (red) visually matches empirical dwell times (blue) across several looped media samples of varying patterns.

Table 2: % of instances where proposed models outperforms alternatives (higher is better, >50% implies superior performance).

LM-DP	CL-LN	CL-IG	CL-WB	CL-G
NLL	54.5%	82.4%	94.6%	93.7
KS	78.9%	86.2%	84.7%	86.7
UM-DP	**CU-LN**	**CU-IG**	**CU-WB**	**CU-G**
NLL	53.6%	78.2%	84.1%	85.5
KS	73.2%	86.7%	88.9%	90.2
V-DP	**CV-LL**	**CV-IG**	**CV-WB**	**CV-G**
NLL	93.6%	82.6%	99.1%	99.9%
K-S	52.8%	54.1%	81.1%	84.1%

samples with ≈ 100 views). Note that since KS tests and p-values are highly sensitive given large sample sizes (i.e. H_0 would be rejected even for minute differences between empirical and simulated data), the percent of successful fits decreases in all cases with high view count; however, given the skewed distribution of view counts, high view count cases constitute only a small fraction of the population. Table 2(LM-DP) further demonstrates the aggregated percentage of samples for which LM-DP outperforms the alternatives, according to both KS test p-values and negative log-likelihood (NLL) of the fitted models; note that these differences are significant and persistent across hundreds of thousands of samples. Additionally, since all the models have same number of parameters, model complexity metrics are proportional to NLL and hence we do not explicitly mention them.

5.1.2 Unlooped Content. Unlike for looped media, unlooped views can have a maximum dwell ratio of 1.0 given viewing constraints (discussed in Observation 1). We observe that no typical continuous value distributions are able to handle this constraint on support natively. Therefore, we propose our UM-DP (**U**nlooped **M**edia **D**well **P**rocess) which significantly augments the LM-DP to handle this constraint. The model can be written generatively as

Definition 5.3 (Unlooped Media Dwell Process (UM-DP)). Sample each dwell time ratio t_i as

(1) $c_i \sim Bernoulli(\theta)$

(2) $t_i \sim \begin{cases} \delta_1(\cdot) & \text{if } c_i = 1 \quad \text{[complete view]} \\ TLL(\alpha, \beta) & \text{if } c_i = 0 \quad \text{[truncated view]} \end{cases}$

where $f_{TLL}(t; \alpha, \beta) = f_{LL}(t; \alpha, \beta)/Z$ is the PDF of right-truncated *LL* distribution on $t_i \in (0, 1)$, $Z = F_{LL}(t = 1; \alpha, \beta) - F_{LL}(t = 0; \alpha, \beta)$ for normalization, and $\delta_1(\cdot)$ denotes a point mass at 1.0.

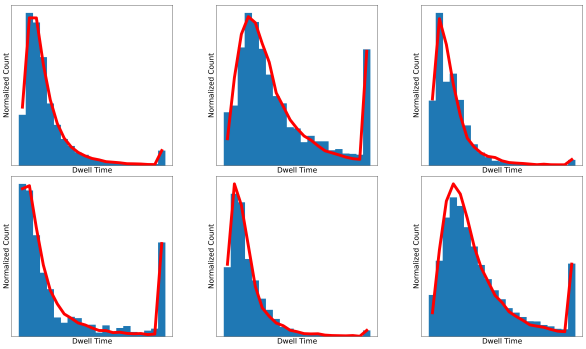


Figure 6: Our proposed UM-DP (red) visually matches empirical dwell time probabilities (blue) across unlooped media samples with varying viewing patterns.

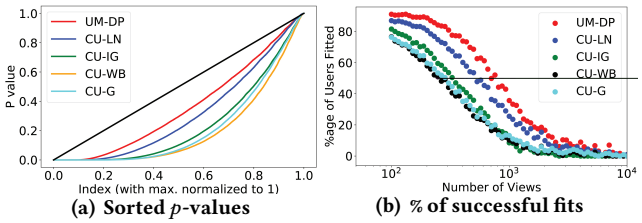
The main idea behind UM-DP is that it considers separately the cases where (a) viewers make a preemptive choice to consume the complete media content (due to friendship, self-selection, etc.), and (b) viewers are less invested and drop off when they lose interest. Intuitively, this reflects a dichotomous choice in media consumption: sometimes, we "exploit" the media which we highly suspect to be interesting given factors like interest in the poster, subscriptions , fascination with a content thumbnail, etc., and other times we "explore" other content whom we give attention to in a fickle way. We note that UM-DP bears resemblance to *hurdle models*, which are often used to model over-inflation of 0s in ecologicaldata[21]; such models pose a "hurdle" via a Bernoulli probability, which when overcome allows a non-zero sample to be generated from an auxiliary process. Our UM-DP places such a hurdle of probability θ on $P(t = 1.0)$ to model complete views, and with probability $1 - \theta$ we sample from the auxiliary TLL distribution such that $t < 1.0$ for truncated views.

Inference of UM-DP. We infer parameters for UM-DP by maximizing the log-likelihood. The overall log-likelihood is given by

$$\ell(\theta, \alpha, \beta) = \sum \theta \log P(t_i = 1.0) + (1 - \theta) \log f_{TLL}(t_i; \alpha, \beta)$$

We can infer θ by maximum likelihood by taking the proportion of empirically observed complete views, i.e. $\hat{\theta} = \sum \mathbf{1}(t_i = 1.0)/n$ over n total views. After filtering the complete views, we can estimate the α, β parameters for TLL on the truncated views by maximizing likelihood heuristically as in the LM-DP case.

Validation of UM-DP. Again, we validate the model both qualitatively and quantitatively. Figure 6 illustrates parity between empirical data and superimposed model fits across unlooped media samples of varying durations, viewer counts and dwell time behaviors. We observe that our proposed UM-DP is able to well-approximate both the completed views (far right), and maintains good performance in modeling the peak/decay corresponding to viewer drop-off despite differences in distribution shapes.



(a) Sorted p-values **(b) % of successful fits**

Figure 7: UM-DP outperforms alternatives:(a) sorted p-values from KS tests; the closer a model curve to the 45° line, better the fit. (b) %age of samples where model fits were successful($p < .05$).

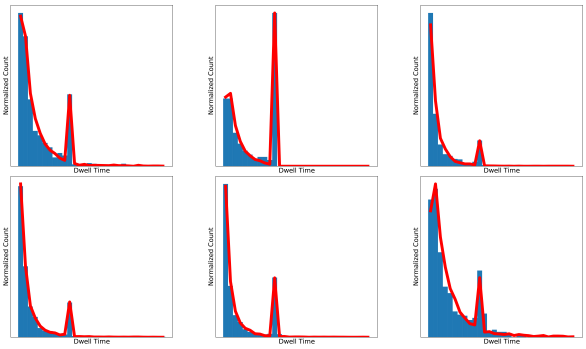


Figure 8: Our proposed V-DP (red) visually matches empirical dwell times (blue) across several looped media samples with varying viewing patterns.

Quantitatively, we again used KS tests and NLL to compare performance of UM-DP with alternatives: CU-LN (Log-normal), CU-IG (Inverse Gaussian), CU-WB (Weibull) and CU-G (Gamma). Technically, we used the truncated variants of these models in the same context proposed in our UM-DP formulation. Figure 7 shows the sorted p-values across samples in (a) and percentage of samples correctly fit against view count in (b); again, we observe that the proposed UM-DP fits the majority of samples well (around 90% with $\approx$ 100 views) outperforms the other models, with the CU-LN model the next closest, and CU-IG/CU-WB demonstrating significantly worse performance. Table 2(UM-DP) further shows that UM-DP outperforms the alternatives over aggregated percentage of samples better fit by both KS tests and log-likelihood comparisons across fitted models.

5.2 Viewers

Modeling viewers has a distinct set of challenges. Most notably, we must model viewers across time that they spend on looped and unlooped media both. Given the differences in support over dwell time ratios over the two, this is non-trivial. Moreover, we must account for differences in inherent propensities of viewers to watch unlooped and looped media. The alternatives to accounting for these complexities in a single joint model are undesirable, as they would result in having individualized models for each user across multiple content types and exposure durations, greatly increasing model complexity and requiring many more samples for inference.

To overcome these challenges, we propose V-DP (**V**iewer **D**well **P**rocess), which aims to unify the modeling of these heterogeneous phenomena. At the core of V-DP is the Log-normal distribution, which we observed closely matched the tails of many viewers' dwell time ratios. The Log-normal distribution is defined as

Definition 5.4 (Log-normal Distribution (LN)). Let T be a non-negative continuous random variable, such that $T \sim LN(\mu, \sigma)$. The PDF and CDF of T are given by:

$$f_{LN}(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}} \qquad F_{LN}(t; \mu, \sigma) = \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

where $t \in (0, \infty)$, $\mu \in (-\infty, \infty)$ and $\sigma > 0$ are the mean and standard deviation of $\log T$, and Φ indicates the standard normal CDF.

Like LL, the LN distribution is also commonly used in survival analysis [7]. Both distributions have very similar shapes; however, LL typically has heavier tails. Intuitively, this disparity in distributions between content-centric and viewer-centric modeling makes sense as viewers have more associated "outgoing" views than contents

have "incoming" ones, and proportionally more of those views tend to be short. This would explain why viewer dwell time ratios exhibit more probability in the head of the distribution with lighter tails, making *LN* a more suitable option for the viewer modeling task than *LL*. Given this, we propose the V-Dp as follows.

Definition 5.5 (V-Dp). Sample each dwell time ratio t_i as

(1) $l_i \sim Bernoulli(\psi)$

(2) $c_i \sim Bernoulli(\theta)$

(3) $$t_i \sim \begin{cases} LN(t;\mu_L,\sigma_L) & \text{if } l_i = 0 \quad \text{[LM view]} \\ \delta_1(\cdot) & \text{if } l_i = 1, c_i = 1 \quad \text{[UM comp. view]} \\ TLN(\mu_U,\sigma_U) & \text{if } l_i = 1, c_i = 0 \quad \text{[UM trunc. view]} \end{cases}$$

where $f_{TLN}(t;\mu,\sigma) = f_{LN}(t;\mu,\sigma)/Z$ is the PDF of right-truncated *LN* distribution on $t_i \in (0,1), Z = F_{LL}(t=1;\alpha,\beta) - F_{LL}(t=0;\alpha,\beta)$ for normalization, and $\delta_1(\cdot)$ denotes a point mass at 1.0.

Our proposed V-Dp is a mixture of viewing processes between both looped and unlooped content. The unlooped content has a max dwell time ratio of 1.0, and thus we sample views to this content in a manner similar to Um-Dp, with the exception of using *TLN* distribution. Looped content has views with unbounded dwell time ratios, and thus we sample these views in a manner similar to Lm-Dp, but using *LN* distribution. The mixture proportions are determined by a parameter trading off propensity for looped versus unlooped media. Note that here, we model views to content with different exposure durations in the same, dwell time ratio model. Technically, though we describe the unlooped and looped views using a single *LN* variant each, we are actually observing the *convolution* of the underlying varying duration distributions.

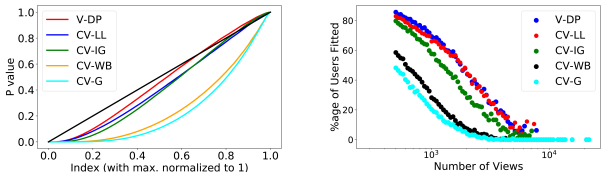
Inference of V-Dp. We aim to maximize the log-likelihood in inferring parameters for V-Dp. The log-likelihood of is given by

$$\ell(\psi,\theta,\mu_U,\sigma_U,\mu_L,\sigma_L \mid t) = \sum \psi[\theta \log P(t_i = 1.0) + (1-\theta)\log f_{TLN}(t_i;\mu_U,\sigma_U)] + (1-\psi)\log f_{LN}(t_i;\mu_L,\sigma_L)$$

Consider n as the total number of views, and n_U and n_L as number of views on unlooped and looped content (such that $n_U + n_L = n$). Then, we have $\hat{\psi} = n_U/n$, and similarly if we consider the number of complete views on unlooped snaps as n_U^C, then $\hat{\theta} = n_U^C/n_U$. To infer parameters μ_L, σ_L for looped media views, we can use closed form estimators. To infer the *LN* parameters μ_U, σ_U for unlooped snaps, we maximize the *TLN* log-likelihood using Nelder-Mead.

Validation of V-Dp. We validate V-Dp both qualitatively and quantitatively. Figure 8 shows several example fits of V-Dp on sample viewers; observe that our formulation allows a flexible fitting of various, complex distributional shapes which represent engagement with highly heterogeneous content using few parameters. Moreover, we compare V-Dp with other candidate models, which as in Um-Dp and Lm-Dp, differ from V-Dp in the central parametric distribution used. The candidate models CV-LL (Log-Logistic), CV-IG (Inverse Gaussian), CV-WB (Weibull) and CV-G (Gamma), differing in replacement of *LN* distribution to respectively mentioned ones.

Quantitatively, we evaluate V-Dp's goodness-of-fit by using KS tests and NLL. We plot the sorted p-values of V-Dp and alternatives in Figure 9(a), demonstrating that V-Dp's p-value curve is closest to the ideal and fits better than alternatives, with CV-LL coming in at a close second. Figure 9(b) shows the percentage of viewers that are successfully fit with V-Dp; here too, we observe that V-Dp fits for majority of the viewers (over 95% with $\approx$ 100 views) with CV-LL performing roughly on par at lower view counts, but trailing behind as view count increases. Again, decrease in fit performance at high view count is encountered by all models due to KS test sensitivity with large sample sizes. Table 2(V-Dp) lists the aggregate percentage of cases where V-Dp performs better than other candidate models in both KS p-values and NLL; NLL suggests significantly better fit performance using V-Dp over the competitors.



(a) Sorted p-values (b) % of successful fits

Figure 9: V-Dp outperforms alternatives:(a) sorted p-values from KS tests; the closer a model curve to the 45° line, better the fit. (b) %age of samples where model fits were successful ($p < .05$).

6 AGGREGATE DWELL TIME MODELING

Given parametric individual fits for each individual content or viewer sample, how can we identify patterns, normative behaviors and anomalies in dwell times of many content or viewer samples, respectively? How common is it to watch over 80% of an image or video? How common is it for a viewer's dwell times to be narrowly distributed around 5% and so on? To answer the above questions, we need to model the parameters in *aggregate*, across many samples. However, modeling the joint distribution of multivariate data is in general not trivial, posing challenges in dependency estimation, inference and curse of dimensionality. In this work, we propose to flexibly model joint distributions of parameters across many content and viewer samples respectively, using a powerful statistical tool known as a *copula* [20]. Copulas allow for scalable, parametric, approximate inference of multivariate distributions. This second level of parametricity in our modeling is advantageous, as it helps us better interpret inter-parameter dependency estimation, enables quick normality scoring and likelihood estimation, and moreover is generative, letting us actually simulate high-quality, realistic dwell time data.

6.1 Copula Modeling

6.1.1 Bivariate Modeling. Copulas are statistical tools, that explicitly model the dependency structure between given univariate marginals to estimate bivariate joint distributions. Copulas have been extensively used in finance [4], healthcare [18] and hydrology research [9]. We can define a bivariate copula as follows:

Definition 6.1 (Bivariate Copula). A bivariate copula C is a dependency function, defined as $C : [0,1]^2 \rightarrow [0,1]$. Given two random variables U and V and their marginal CDFs F_U and F_V, a copula $C(F_U(u), F_V(v))$ models the joint CDF, admitting a joint PDF of

$$f_{U,V}(u,v) = f_U(u) \cdot f_V(v) \cdot c(F_U(u), F_V(v))$$

where c and f denote copula and marginal densities.

Technically, copulas are defined on uniform marginal CDFs. We can transform any random variable Y to uniformity by using probability integral transform (PIT) or vice-versa (inverse transform sampling). Various parametric forms of copula exist and can be used to capture different dependencies (positive, negative, independent) between different types of random variables. While bivariate copulas have demonstrated great empirical success in capturing dependencies via a variety of parametric forms, the number of generalized

multivariate parametric copulas (for > 2 variables) are highly limited and inflexible in preserving pairwise dependencies, resulting in poor estimation. Given that some of our proposed models are multivariate, we seek a better option: to model multivariate dependencies parametrically while also allowing for flexible pairwise dependency modeling in high dimension, we propose the use of Vine copulas.

6.1.2 Multivariate Modeling. Vine copulas leverage the flexibility of parametric bivariate copulas to preserve bivariate statistical dependencies in higher-dimensional joint distributions. The dependency structure is modeled by the composition of (a) a set of bivariate copula families, (b) the associated copula dependency parameters, and (c) a nested tree structure to model the decomposition of joint distribution into the bivariate copula and marginal densities, as follows [6]

Definition 6.2 (Vine copula). A vine copula on n random variables $X_1 \dots X_n$ has a joint PDF defined by

$$f_{X_1 \dots X_n}(x_1 \dots x_n) = \prod_{} \prod_{} \prod_{} c_{i,i+j|i+1,\dots,i+j-1} \cdot \prod_{} f_k(x_k)$$

where c and f denote associated copula and marginal densities.

Different tree structures have been proposed to model these dependencies ; in this work, we use canonical vines (C-vines). C-vines decompose marginals and bivariate copula densities such that every tree has a one-to-many structure:

Definition 6.3 (C-vine). A set of linked trees $\mathcal{V} = (T_1, T_2, T_{n-1})$ is a C-vine on n elements if

(1) T_1 is a tree with nodes $N_1 = 1, \dots, n$ and a set of edges E_1 between a selected node $a \in T_1$ and all other nodes $b \in T_1$.
(2) For $i = 2, \dots, n-1$, T_i is a tree with E_{i-1} nodes and edge set E_i such that a single node in T_i is connected to all other nodes in T_i, and no other edges exist.

Inference. To select the appropriate C-vine structure, we use the procedure as mentioned in [6]; specifically the node with maximum absolute Kendall's τ-correlation to other nodes is selected as central node for each level tree. Given the structure, we maximize log-likelihood to infer bivariate copulas and the associated parameters.

6.2 Multimedia Content

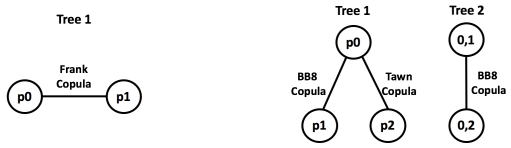
Below, we discuss how we conducted modeled aggregate modeling for looped and unlooped media.

6.2.1 Looped Content. Since our Lm-Dp produces only 2 parameters for each content sample, a bivariate copula suffices to model the two-parameter dependency. To do this, we used Lm-Dp to to fit parameters for all looped media samples, and subsequently applied the PIT using the empirical CDFs for both α and β describing the dwell ratio scale and shape. We then selected the bivariate copula (shown in Figure 10(a)) which best maximizes the log-likelihood across a variety of parametric forms discussed in [19], and inferred parameters using 30% of the samples; we call this model Lm-Am.

6.2.2 Unlooped Content. We obtained 3 parameters from Um-Dp, θ, α, β which describe view completion rate and truncated view dwell time ratio scale and shape. Given the multivariate setting, we inferred parameters for a Vine copula (shown in Figure 10(b)), training on 30% of the samples as in the looped case. We call this model Um-Am.

6.3 Viewers

In modeling individual viewer dwell ratios, our V-Dp produced 6 parameters for each viewer: $\psi, \theta, \mu_L, \sigma_L, \mu_U, \sigma_U$, denoting propensity to view unlooped media, propensity to complete unlooped views, and mean and standard deviation of the log dwell ratios for looped and truncated unlooped views, respectively. Using a sample of $100K$ instances , we estimated and fitted a C-vine. We call this model V-Am.



(a) Lm-Am (Looped Media) (b) Um-Am (Unlooped Media)

Figure 10: Bivariate and C-vine copula structures can model joint densities parametrically. (a) and (b) show our Lm-Am and Um-Am dependency structures, respectively.

Table 3: Pearson correlation coefficients between parameters in original and simulated data.

Correlations	(p0, p1)	(p1, p2)	(p0, p2)
Original	−0.32	0.08	0.43
Simulated	−0.31	0.10	0.41

6.4 Validation

To evaluate the performance of C-vine modeling in our usecase, we consider the following aspects:

- **Q1. Dependency preservation**: How well does C-vine approximate the original data dependencies?
- **Q2. Training size:** How is C-vine modeling performance influenced by training size?
- **Q3. Temporal consistency**: How robust is C-vine modeling for similar data from two different time-frames?

Given space constraints, we show experimental results only on Um-Am, noting that those for Lm-Am and V-Am are similar.

6.4.1 Dependency preservation. Here, we determine if dependency structure in original data is well approximated by the C-vine model. To this end, we compare generated random samples from the simulated data on $[0, 1]^n$ ($n = 3$ for Um-Am, used here) to the PIT-representation of training samples. We report the pairwise Pearson correlations in Table 3, and show heatmaps of the pairwise dependencies in Figure 11. We observe that correlations between all parameter pairs and density estimates are closely approximated.

6.4.2 Training size. We also study the effects of training size on C-vine modeling performance. We experimented by training the C-vine model using random samples of varying sizes from the entire set, and sampling instances from the fitted models. To comparing samples from multivariate distributions, we use kernel-based Maximum Mean Discrepancy (MMD) test as proposed in [10] (the KS test is only suitable for univariate samples), to test the null hypothesis H_0: simulated samples and original data samples are from same distribution. We present the MMD test statistic for data simulated using models with various training sizes in Table 4; results show that we are not able to reject H_0 in any case. Even when using only 10% training data, we observe that the C-vine model closely approximates the original data distribution. Notice that the MMD statistic decreases as training size increases, showing closer approximation towards the original data.

6.4.3 Temporal consistency. We next aim to validate that aggregate models produced from C-vine are temporally consistent, in that they closely match across data taken from different time periods

(we expect the underlying behavior does not shift significantly). We fit another C-vine model using dwell time engagement data from a different month than the data discussed here. We then compared the simulated data from both C-vine models and evaluated similarity between the two. To this end, we perform an MMD test between the two samples, obtaining a test statistic of 0.032, which does not let us reject H_0, and thus we can say they are drawn from same distribution. We observe this visually in Figure 12, where samples generated from both C-vines produce similar dependency structures for each pair.

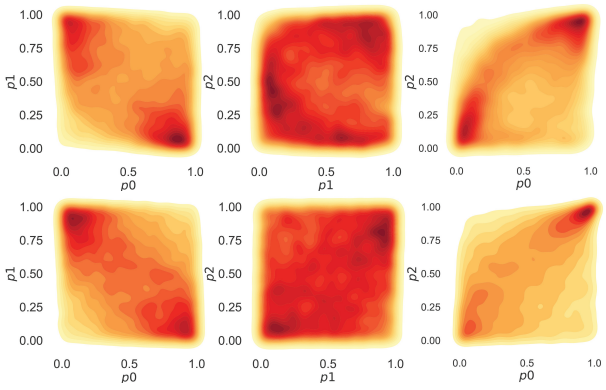


Figure 11: Aggregate C-vine models closely approximate real data. Pairwise dependency heatmaps between original data (top) and simulated data (bottom) are visually close.

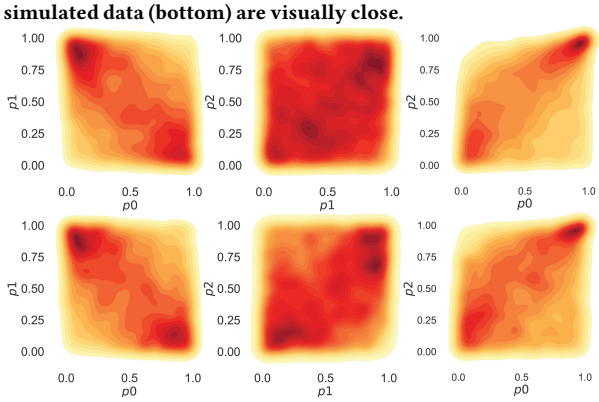


Figure 12: C-vine models are robust and consistent over time. Pairwise dependency heatmaps between simulated data from aggregate models trained on two different months (top and bottom) are visually close.

7 ANOMALY DETECTION

In the previous section, we introduced parametric copula-based models for aggregate content and viewer modeling, demonstrating success in modeling the vast majority of samples while preserving complex interactions between parameters. A natural line of evaluation is determining effectiveness of such models in detecting anomalous engagement samples(i.e. samples which have extremely low-likelihood according to the aggregate models); thus, we pose the following questions.

Table 4: MMD test statistics between original data and model-simulated data (lower is better).

Training Size	10%	20%	30%	40%	50%
MMD-Statistic	0.037	0.034	0.0334	0.333	0.30
Reject H_0	No	No	No	No	No

- **Q1. Robustness to contamination**: Can our aggregate models robustly detect anomalies under contamination?
- **Q2. Qualitative efficacy**: Do our aggregate models detect real engagement anomalies on real data?

7.1 Robustness to contamination

We first study the performance of our aggregate model in detecting anomalies present in the training data, known as contamination. This scenario is possible in unsupervised models, like ours, as anomalous samples are not labeled and are also involved in individual and aggregate modeling steps. Ideally, our models should be able to detect anomalies in training data with high precision. To evaluate performance in such settings, we analyze our model's ability to successfully detect simulated attacks, by means of injecting artificial, anomalous samples in the original data. We present results for only Um-Am given space constraints, but results for Lm-Am and V-Am were similar.

We consider 4 different contamination models (shown below) in which anomalies are generated (a) according to different attack models, and (b) constituting varying contamination ratios. **Model 1 (Complete Views)**: Anomalies have all fully complete views: dwell time ratios of 1.0, **Model 2 (Long Views)**: Anomalies have overly long views: dwell time ratios sampled uniformly between 0.8-1.0, **Model 3 (Short Views)**: Anomalies have overly short views: dwell time ratios Gamma-distributed such that most dwell time ratios < 0.2, and **Model 4 (Uniform Views)**: Anomalies have random-length views: dwell time ratios sampled uniformly on 0.0-1.0. Also, we consider varying contamination ratios of 1%, 2% and 5% anomalies in the training data. We evaluate detection capacity via AUROC, which is reliable in imbalanced class settings like ours. Results are shown in Table 5, and indicate extremely high detection performance. We observe an AUC of 0.9+ across most scenarios, noting that higher contamination results intuitively result in lower AUC due to increased model corruption.

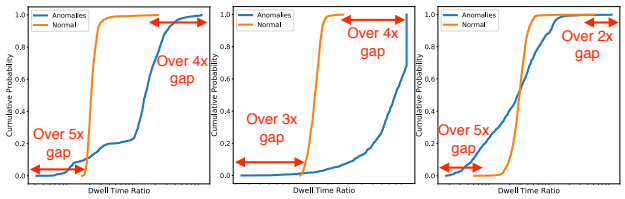
7.2 Effectiveness on real data

Next, we aim to evaluate whether our models can actually detect anomalous dwell time engagement in real data. To this end, we selected the 1000 most normal and anomalous samples according to log-likelihood, for each looped/unlooped content sample and viewer under Lm-Am, Um-Am and V-Am respectively, and compared the empirical CDFs of mean dwell times across these entities. Intuitively, if our aggregate models were not detecting anomalous engagement, the empirical CDFs would closely match. However, as Figure 13 shows, the curves are significantly different for normal and anomalous samples identified by each model; note that the x-axis is in log-scale, making the observed differences more significant. We observe clear differences throughout the range of the CDF, and moreover discover the biggest differences near the extremities, suggesting our model does detects engagement anomalies. At the lower extremity of dwell time ratios, we observe that the lowest anomalous samples dwell times were $3-5\times$ smaller than those of their normal counterparts. Likewise, at the upper extremity, the highest anomalous sample dwell times were $2-4\times$ larger.

Manual inspection of several observed anomalous dwell time behaviors indicated significant abnormalities: (1) One anomalous viewer had over 5000 views/day, with mean dwell ratio < 0.03, and was adding more than 200 friends/day from an already staggering 3900, (2) several anomalous looped media samples with over 500 views had mean dwell ratio of $10-15\times$ the duration, and (3) several unlooped media samples with 100-300 views had mean dwell ratios

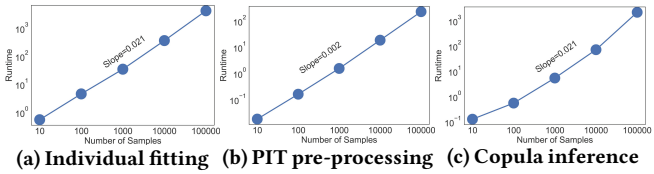
Table 5: Anomaly detection performance (AUROC) under various anomaly contamination %ages (higher is better).

Attack Model	1%	2%	5%
Model 1 (Full Views)	0.99	0.98	0.96
Model 2 (Long Views)	1.0	0.99	0.99
Model 3 (Short Views)	1.0	0.99	0.98
Model 4 (Uniform Views)	0.94	0.92	0.84



(a) Looped content (b) Unlooped content (c) Viewers

Figure 13: Our aggregate models detect real dwell time anomalies. The subplots show huge disparities in the mean dwell time ratio distributions between anomalous and normal (a) unlooped media, (b) looped media and (c) viewer samples.



(a) Individual fitting (b) PIT pre-processing (c) Copula inference

Figure 14: Our model inference is scalable: (a-c) show that individual fitting, copula preprocessing via integral transform, and copula inference are all near-linear in sample size.

of over 0.9; one sample with over 1000 views had a ratio of just 0.03. Figure 1(d) shows several examples of unlooped content anomalies discovered by Um-Am (others excluded for brevity). These anomalies could correspond to fake engagement, or possibly offensive or polarizing media. Overall, results demonstrate that our approach does empirically detect real-world anomalies across aggregate models, and could be additionally correlated with other features to discern abusive behaviors of various types.

8 SCALABILITY

We briefly discuss scalability in terms of both individual dwell process fitting and aggregate copula modeling. The major runtime cost in the former case is log-likelihood maximization for fitting parameters of the relevant dwell process.Figure 14(a) shows that this procedure exhibits empirically linear runtime in the number of training samples. The runtime costs in the latter case are incurred in conducting the PIT on original data samples for copula pre-processing, and selecting ideal copula structure and parameters. Figures 14(b) and 14(c) show that these steps admit linear and near-linear runtime respectively. Results are shown on Um-Dp/Um-Am.

9 CONCLUSION

In this work, we provide the first comprehensive analysis of modeling dwell time engagement on visual multimedia content. Studying such content is valuable, as its consumption constitutes a significant portion of daily online activity, and has valuable applications in behavior modeling and anomaly detection. We first discuss challenges and considerations in the modeling task, including content heterogeneity and behavioral diversity. Our first contribution constitutes the Lm-Dp, Um-Dp and V-Dp generative dwell time processes and inference procedures, which enable *individual modeling* of content-centric and viewer-centric dwell time engagement. We show that these models match empirical data visually and quantitatively according to KS tests and outperform alternatives in both log-likelihood and KS tests. Our next contribution posits the analog Lm-Am, Um-Am and V-Dp, which enable *aggregate modeling* of joint distributions across individual fits using parametric bi/multivariate copulas. We demonstrate the flexibility of such models in capturing high dimensional dependencies with limited training data, show that they closely match original data both visually and quantitatively according to MMD tests, and are temporally consistent. Our last contribution includes ramifications of our proposed models for *anomaly detection*, in both robustness to contamination (0.9+ AUROC in most experiments) and qualitative evidence in terms of anomalies detected on real engagement data.

REFERENCES

[1] Steve Bennett. 1983. Log-logistic regression models for survival data. *Applied Statistics* (1983).
[2] YouTube Offical Blog. 2017. You know whatś cool? A billion hours. https://goo.gl/zPNNT7. (2017).
[3] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A context-aware time model for web search. In *SIGIR*.
[4] Eric Bouyé, Valdo Durrleman, Ashkan Nikeghbali, Gaël Riboulet, and Thierry Roncalli. 2000. Copulas for finance-a reading guide and some applications. (2000).
[5] Liang Chen, Yipeng Zhou, and Dah Ming Chiu. 2015. Analysis and Detection of Fake Views in Online Video Services. *TOMM* 11, 2s (2015).
[6] Claudia Czado. 2010. Pair-copula constructions of multivariate copulas. In *Copula theory and its applications*. Springer, 93–109.
[7] Arabin Kumar Dey and Debasis Kundu. 2009. Discriminating between the log-normal and log-logistic distributions. *Commun. Stat. Theory Methods* (2009).
[8] A. F. Costa, Y. Yamaguchi, A. J. M. Traina, C. Traina, Jr., and C. Faloutsos. 2015. RSC: Mining and Modeling Temporal Activity in Social Media. In *KDD*. ACM, 269–278.
[9] A. C. Favre, S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée. 2004. Multivariate hydrological frequency analysis using copulas. *Water Resources* 40, 1 (2004).
[10] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. 2007. A kernel method for the two-sample-problem. In *NIPS*. 513–520.
[11] Charles N Haas, Josh Joffe, Mark S Heath, and Joseph Jacangelo. 1997. Continuous flow residence time distribution function characterization. *J. of Env. Eng.* 123 (1997).
[12] James Douglas Hamilton. 1994. *Time series analysis*. Vol. 2. Princeton Univ. Press.
[13] D. Juan, N. Shah, M. Tang, Z. Qian, D. Marculescu, and C. Faloutsos. 2017. M3A: Model, MetaModel and Anomaly Detection for Inter-arrivals of Web Searches and Postings. In *DSAA*. 341–350.
[14] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *WSDM*. ACM, 193–202.
[15] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. *J. of Opt.* 9, 1 (1998).
[16] Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *SIGIR*. 379–386.
[17] M. Marciel, R. Cuevas, A. Banchs, R. González, S. Traverso, M. Ahmed, and A. Azcorra. 2016. Understanding the detection of view fraud in video content portals. In *WWW*. IW3C2, 357– 368.
[18] José MR Murteira and Óscar D Lourenço. 2011. Health care utilization and self-assessed health: specification of bivariate models using copulas. *Empirical Economics* 41, 2 (2011).
[19] Thomas Nagler. 2018. VineCopula. https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf. (2018).
[20] R. B. Nelsen. 2007. *An introduction to copulas*. Springer Science & Business Media.
[21] Martin Ridout, Clarice GB Demétrio, and John Hinde. 1998. Models for count data with many zeros. In *IBC*, Vol. 19. IBS, 179–192.
[22] Neil Shah. 2017. FLOCK: Combating Astroturfing on Livestreaming Platforms. In *WWW*. 1083–1091.
[23] TIME. 2016. Here'ĂŹs How Much Time Snapchat Users Spend on the App. http://time.com/4272935/snapchat-users-usage-time-app-advertising/. (2016).
[24] TIME. 2016. Instagram Just Hit the 500 Million User Mark. http://time.com/money/4376329/instagram-users/. (2016).
[25] P. O. S. Vaz de Melo, L. Akoglu, C. Faloutsos, and A. A. F. Loureiro. 2010. Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. In *ECML-PKDD*.
[26] Songhua Xu, Hao Jiang, and Francis Chi-Moon Lau. 2011. Mining user dwell time for personalized web search re-ranking. In *IJCAI*. AAAI.
[27] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond clicks: dwell time for personalization. In *RecSys*. ACM, 113–120.
[28] P. Yin, P. Luo, W.-C. Lee, and M. Wang. 2013. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *KDD*. ACM.

10 REPRODUCIBILITY

10.1 Data collection

In this work, we collected a large dataset based on user engagement with Snapchat's Stories feature. We used publicly posted "My Story" snaps, which were made available to all Snapchat users (rather than the default "close friends" option) for privacy reasons. For our content modeling tasks, we collect viewing data associated with all such snaps posted in a 2 hours interval on a single day in May 2018. Furthermore, we collected all views associated with these Snaps (during the next 24h hour span). We filtered the snaps that have duration ≤ 4 seconds due to a discovered timer instability. Further, we truncated any views that appeared to persist longer than the media length; these are rare, but possible due to certain Snapchat app features like media long-presses.

For our viewer modeling task, we collected views to all Snaps over a 24 hour period. As previously mentioned, we conduct analysis only on Snaps with more than 100 views, and similarly filter out viewers who had viewed less than 100 Snaps during the associated timeframe. Due to privacy concerns and user agreements, the data is not shareable; however, we make our code developed for modeling and analysis available publicly.

10.2 Copula preliminaries

In this paper, we used copulas for *aggregate* modeling (Section 6). Copulas are a powerful statistical tool, which allow for scalable, parametric, approximate inference of multivariate distributions. As mentioned earlier, copulas are often used to model bivariate joint distributions, which is achieved by explicitly modeling the dependency structure given univariate marginals.

While bivariate copulas have demonstrated great empirical success in capturing dependencies via a variety of parametric forms, the number of generalized multivariate parametric copulas (for > 2 variables) are highly limited and inflexible in preserving pairwise dependencies, resulting in poor estimation. Therefore, we use *Vine copulas* which model multivariate dependencies parametrically by flexibly modeling independent pairwise dependencies in high dimensions, effectively as multi-level trees. To elucidate, we give an example of a C-vine copula (one type of vine structure, which we used in this work) as follows:

Example. Consider three random variables A, B, C and their corresponding marginal distributions f_A, f_B, f_C. In a C-vine, only one node is connected to all other nodes at each level tree. Thus, the first tree could be constructed by joining A with B and C, with the edges in the first tree becoming nodes in the second tree (see Figure 15). The associated joint distribution can be written as

$$\begin{aligned} f_V(a,b,c) = & f_A \cdot f_B \cdot f_C && \text{[Nodes in Tree 1]} \\ & c_{A,B} \cdot c_{A,C} && \text{[Edges in Tree 1]} \\ & c_{B,C|A} && \text{[Edges in Tree 2]} \end{aligned}$$

10.3 Implementation details

We used Jupyter notebooks and Python to run all experiments on a high-memory, single-node Google Cloud compute engine instance.

For individual dwell time modeling (Section 5), we used the Python `statsmodels` `GenericLikelihoodModel` module, which enables general likelihood function maximization via builtin optimizers (we used Nelder-Mead). We used `scipy.stats` for parametric models of the discussed distributions (Log-logistic, Gamma, Weibull, Log-normal). We fixed the location parameter to 0 for all distributions while fitting, due to the uninterpretability of negative values in dwell times. We evaluated model fits using the `KS_test` functionality in `scipy.stats` to obtain test-statistic and p-values.

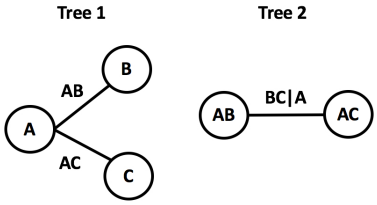


Figure 15: Illustration of C-vine copula structure: In Tree 1, marginals are used to fit bivariate copulas; these become nodes in Tree 2, between which another bivariate copula is fit.

For aggregated dwell time modeling (Section 6), we used the `VineCopula` package in *R* (well-documented and maintained). The C-vine dependency structure and parameter estimation was maximized according to log-likelihood. For computing the MMD test statistics to compare multivariate distributions, we used the `kmmd` package in *R*.

We make all our code available at https:/github.com/hemanklamba/ModelingDwellTime.

10.4 Generality

Our work presents the largest to-date modeling, evaluation and analysis of multimedia dwell times to-date, and inferences are drawn from *300 thousand* media samples, *24 million* viewers and *273 million* views. The data is a rich representation of multimedia engagement on the Snapchat platform in which visual multimedia is the predominant method of communication. Moreover, modeling inferences drawn were from sufficiently large sample sizes, suggesting that modeling inferences are consistent and accurate across a wide variety of user behaviors and content types. Though our work does not utilize data from other platforms for data availability and privacy reasons, we expect that the findings are representative of short-form visual multimedia at the least, but also have expected applicability to longer-form multimedia, given that most views to short-form content are even shorter, and decay with (super)exponential tails.