

Classification of abrupt changes along viewing profiles of scientific articles

Ana C. M. Brito¹, Filipi N. Silva², Henrique F. de Arruda³,
Cesar H. Comin⁴, Diego R. Amancio¹ and Luciano da F. Costa³

¹*Institute of Mathematics and Computer Science,
University of São Paulo, São Carlos, SP, Brazil*

²*Indiana University Network Science Institute, Bloomington, Indiana 47408, USA*

³*São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, Brazil*

⁴*Department of Computer Science, Federal University of São Carlos, São Carlos, SP, Brazil*

(Dated: October 12, 2020)

arXiv:2005.04512v3 [cs.DL] 9 Oct 2020

Abstract

With the expansion of electronic publishing, a new dynamics of scientific articles dissemination was initiated. Still substantially important, citations became a longer term effect. Nowadays, many works are widely disseminated even before publication, in the form of preprints. Another important new element concerns the views of published articles. Thanks to the availability of respective data by some journals, such as PLoS ONE, it became possible to develop investigations on how scientific works are viewed along time, often before the first citations appear. This provides the main theme of the present work. More specifically, our research was motivated by preliminary observations that the view profiles along time tend to present a piecewise linear nature. A methodology was then delineated in order to identify the main segments in the view profiles, which allowed several related measurements to be derived. In particular, we focused on the inclination and length of each subsequent segment. Basic statistics indicated that the inclination can vary substantially along subsequent segments, while the segment lengths resulted more stable. Complementary joint statistics analysis, considering pairwise correlations, provided further information about the properties of the views. In order to better understand the view profiles, we performed respective multivariate statistical analysis, including principal component analysis and hierarchical clustering. The results suggest that a portion of the polygonal views are organized into clusters or groups. These groups were characterized in terms of prototypes indicating the relative increase or decrease along subsequent segments. Four respective distinct models were then developed for representing the observed segments. It was found that models incorporating joint dependencies between the properties of the segments provided the most accurate results among the considered alternatives.

I. INTRODUCTION

Science can be understood as a social activity, conceived and applied by humans. As a consequence, communication plays a critical role in scientific development, allowing important results to be disseminated and used. Interchange is important not only between scientists working on related fields, but also between those deriving results and those applying these results. In the beginnings of science, communication proceeded mostly in terms of *letters* (see e.g. (Peat 2002)), which were exchanged between scientists in order to share their most recent results. Letters gave rise to proceedings, journals and, more recently, World Wide Web-based dissemination. The study of how scientific articles are read and cited is of great importance because such knowledge provides insights about the efficiency at which science is disseminated.

Many of the existing studies in scientometrics, the area aimed at studying how science unfolds, consider citations as the main indicator of usage and interest of scientific articles. Until recently, this was one of the few available objective measurements of scientific dissemination (Amancio et al. 2012, Bollen et al. 2005, Waltman 2016). Yet, with the introduction of the Internet and the WWW, other statistics became available, such as the number of views, shares and downloads of articles published online. Indeed, before the Internet and the WWW, it was very difficult to count how many times a journal or article was taken from the shelves and read. The availability of these new indicators paved the way to many interesting investigations in scientometrics, motivating the new area of *altmetrics* (Sud and Thelwall 2014).

Among the new scientific indicators, the number of views has some particularly interesting features. First and foremost, it takes place at a relatively high speed, involving little delay: once published online, a work starts being viewed almost immediately. Contrariwise, the first citation of a work can take months or even years to take place. Given that views tend to be faster, they can provide insights about current trends, allowing predictions to be made. Views also tend to take place in larger numbers than citations, therefore providing a potentially more complete sample that can lead to more accurate statistical analysis. Studying views is also intrinsically important as a means to better understand its relationship with citations.

One of the few limitations intrinsic to views is that they provide a somewhat weaker

indication of the use of the knowledge in the visualized work. Indeed, some views can be the consequence of actions of Web crawlers or surveys, without a direct implication that the reported knowledge has been somehow transferred or applied. All in all, views-based scientometrics has potential for contributing substantially to our knowledge about how scientific information is disseminated.

Being mostly based to online publication, and by providing several statistics, the PLoS ONE journal Note1 represents a good resource for performing scientometric/altmetric studies focusing views as main indicators. In particular, the number of views of each article is provided along time in a month-by-month fashion. Figure 1 illustrates some view profiles for 6 randomly chosen articles.

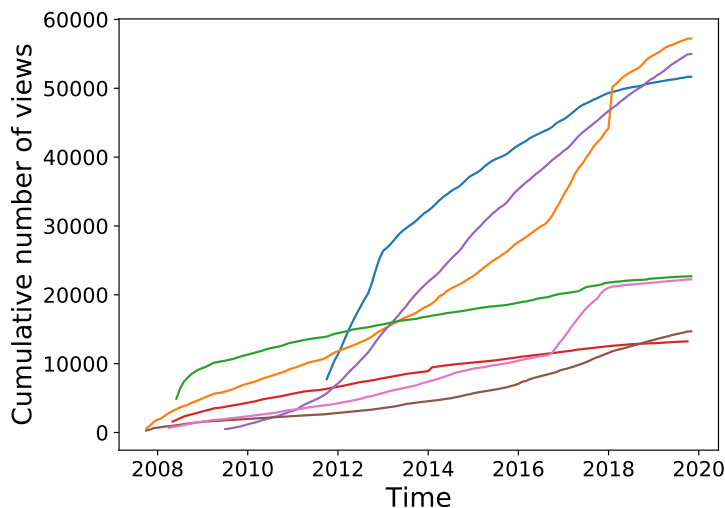


FIG. 1. Profiles of 6 randomly chosen PLoS ONE articles showing, in each case, the cumulative number of views along successive months. Interestingly, several of these profiles present approximate polygonal shape, with sharp turns followed by relatively straight segments.

Interestingly, some of the patters in Figure 1 suggest a polygonal organization, including relative increases and decreases of views along subsequent segments. This is even more interesting than the former effect, because it shows that some sort of events occur that are capable of changing substantially the visibility of a given article along time. The study about how frequent these patterns are, and their possible organization into respective categories, constitute the main question addressed in the present work.

More precisely, viewing profiles are obtained for almost every article published in PLoS ONE. These signatures are then analysed by using state of the art numerical methods,

namely segmented regression (Muggeo 2003), capable of fitting contiguous straight line segments along each signature, to match its original profile. The slope and transition points of these segments, corresponding to abrupt slope changes, can then be identified and used for statistical analysis and derivation of statistical models. Indeed, the current approach focuses on the derivation of these models as the means for trying to understand the properties of the considered view profiles.

The proposed models are based on variables describing the inclination angle and the extension of each subsequent linear-piecewise portion. Conditional densities are estimated from the experimental data, allowing the generation of synthetic profiles. We consider independent densities as well as Markov-1 univariate and multivariate dependencies between variables. Then, by comparing the congruence between the original and synthetic profiles, we can estimate which of the considered models tend to be more accurate. The so-obtained model is then studied in order to identify and understand possible mechanisms producing the observed profiles.

Several interesting results are reported. First, the linear-piecewise approximation was found to be more congruent with the real-world view profiles than with synthetic profiles obtained from uniformly random events. This supports the hypothesis that the real-world profiles tend to present a polygonal structure, therefore being potentially well-represented and modeled by using this type of curve. The several parameters observed for each set of views, considering the same number of segments, revealed two relatively well-defined clusters corresponding to two main types of polygonal profiles exhibiting sequences of varying slopes. For instance, in the case of profiles containing 3 segments, one of the detected clusters is characterized by a relatively low initial slope followed by a higher slope, which is then reduced in the final segment. The other type of cluster presented opposite structure. Among the several types of statistical models considered in this work, we found that the use of conditional densities defined on joint pairs of previous parameters, more specifically by predicting the slope and interval at each current time in terms of the the previous instances considered jointly, led to the most accurate results. This suggests that each segment along the polygonal structures can be, to a good extent, estimated by the previous segment, corresponding to a memory-1 Markov dynamics that may underlies respective real-world effects.

The remaining of this paper is organized as follows. In Section II, we present the re-

lated works. In Section III, we describe the dataset, the employed clustering algorithm, the methodology for curve characterization, and the employed statistical model. In Section IV the obtained results are discussed. Conclusions and perspectives for future works are presented in Section V.

II. RELATED WORKS

Altmetrics indicators have been employed to quantify the quality of some aspects of science (Bornmann 2014, Erdt et al. 2016, Galligan and Dyas-Correia 2013, Sud and Thelwall 2014). Some analyses account for the importance of articles (Huang et al. 2018) and others for measuring characteristics of scientists (Ioannidis et al. 2014, Larivière et al. 2013). Interestingly, many altimetric measurements are not found to be correlated to citation counts (Erdt et al. 2016), which indicates that these metrics can measure different and complementary information.

The Altimetric Attention Score (AAS) has been used to measure the importance of papers (Huang et al. 2018). This index measure mentions of papers in social media (e.g., Twitter and Facebook). Huang et al. (2018) found correlations between AAS and the number of citations for some journals. More specifically, by considering papers obtained from PLoS journals, the authors measured the Spearman correlation between a normalized AAS and a normalized count of citations. Interestingly, in the case of Medicine articles, this correlation was not found. In another study that considered highly cited papers, correlations were found in the comparison between metrics obtained from social networks and the number of citations (Thelwall et al. 2013).

Another important information that can be measured is the difference between the behavior of new and old articles. Due to the fast evolution of computer-related areas, computer science researchers attribute great importance to conference papers. By considering papers of journals and conferences, Thelwall (2019) compared the number of citations with *Mendeley Readership*. Mendeley Readership measures the number of users that include a given article to their account. They found that the number of Mendeley reads and the number of citations is correlated for both journals and conferences. However, in the case of old conference papers, a similar correlation was not found. Furthermore, Schlögl et al. (2014) compared citations, downloads, and Mendeley Readership of two information systems

journals, and found high Spearman correlations when comparing downloads with citation and downloads with Mendeley Readership. However, in the comparison between Mendeley Readership with citations, a moderate correlation was found.

One particularly important measurement is the number of views, for which some distinct aspects have been analyzed. For instance, the number of views can be linearly correlated with the age of a paper (Priem et al. 2012), in which older articles tend to have more views. Furthermore, in (de Winter 2015), the authors investigated the relationship between the number of article's views and the mentions of articles in *Twitter*. More specifically, their study suggests that views obtained from *tweets* are not related to views and citations. Other scholars also analyze the number of article views and downloads and found that the latter is much more correlated with citations than views (Wang et al. 2014). In an analysis regarding medical papers, the documents with high early views counts tend to be more cited than others (Perneger 2004). In a comparison among different altmetrics extracted from *PlumX*, the measurements of views and downloads were found to have the most extended life cycles (Ortega 2018). Their results also indicate that mentions in Twitter and blogs can impact the number of views and downloads.

Researchers have also been comparing the number of downloads with citations (Erdt et al. 2016). In some studies, a correlation between the number of downloads and citations was found (Brody et al. 2006, Jamali and Nikzad 2011, Perneger 2004, Watson 2009). A form of fast transmitting scientific results is publishing papers on *arXiv*, which is a preprint repository. In a study that took into account early citations (Shuai et al. 2012), the authors compared *arXiv* downloads and Twitter mentions and found that there is a correlation between tweets and downloads. By considering a given journal, in Moed (2005), the authors found that in the following three months after a citation, the number of downloads tends to increase. Additionally, for the same dataset, when downloads and citation distributions are compared, the older, the more similar (Moed 2005).

Another possibility of analysis is the comparison between altmetric indicators and the linguistic characteristics of the papers. In Chen et al. (2020), many distinct features were measured from papers of PLoS, which included text length, lexical diversity, lexical density, among others. Interestingly, for the majority of the considered characteristics, no significant correlation was found. However, for some PLoS journals, the title lengths and average sentence length were found to play an important role in the number of views and downloads.

This correlation was also found for Jamali and Nikzad (2011). However, Duan and Xiong (2017) did not find a correlation between the total number of downloads and title lengths.

III. MATERIALS AND METHODS

A. The Dataset

We extracted information from all the papers published in PLoS ONE up to 2016. For that, we employed a semi-automatic extraction of paper metadata using ALM API. This data was collected in November 2019. For each article, the dataset contains information about the number of views per month, as well as other social media features, including the number of tweets and shares. The latest properties still require some post-processing and further validation before use. Thus, we only focus on analyzing the number of views along time, which resulted in a total of 162,534 view profiles. The total number of publications views is divided into HTML, PDF, and XML. In order to be compatible with the way in which views are understood and exhibited in the PLoS ONE web site, we used the sum of all types, instead of only the HTML or PDF views.

It is important to observe that the consideration of data like in the PLoS ONE dataset will typically not account for views through other means such as pre-prints, department reports, etc. However, these views are expected to be relatively less frequent than to those recorded in the adopted database.

B. Agglomerative clustering

In order to identify groups of related articles regarding their viewing patterns, agglomerative clustering (Jain et al. 1999, Müllner 2011) will be applied in the experiments. An important choice concerns the linkage criterium to be used for aggregating datapoints, which often consists in adopting Ward’s approach. One disadvantage of this method is that it tends to identify groups even when the groups are not well-defined or do not exist, which are henceforth referred as false positives. The linkage method that is likely most suited for avoiding false positives is the *single-linkage* (Tokuda et al. 2020). This simple criterium, which links two groups based on the smallest distance between any two objects in those groups, only leads to the detection of clusters if there is a marked separation between the points belonging

to the clusters. For instance, for uniformly distributed data the single-linkage approach will result in a dendrogram containing similar cophenetic distances (Sokal and Rohlf 1962), thus indicating that clusters cannot be derived from the dendrogram. On the other hand, other common linkage criteria, in particular the Ward’s method, tend to result in dendrograms indicating clusters in the data.

The robustness of the single-linkage method to false positives has an important consequence. Clusters found by this method are unlikely to be caused by statistical fluctuation, missing data or noise. Thus, the identified clusters can be considered as being more statistically significant.

C. Curve characterization

Following (Muggeo 2003), a breakpoint is modeled as two straight lines joined at point ψ , that is,

$$y = \alpha x + \beta(x - \psi)_+ \quad (1)$$

where $(x - \psi)_+ = (x - \psi)I(x > \psi)$, with $I(A) = 1$ if condition A is true and $I(A) = 0$ otherwise. ψ is the position of the breakpoint, α is the slope of the line before the breakpoint and $\alpha + \beta$ is the slope after the breakpoint. Therefore, β is the difference in slopes between the two lines. This is illustrated in Figure 2. The term $(x - \psi)_+$ can be rewritten as

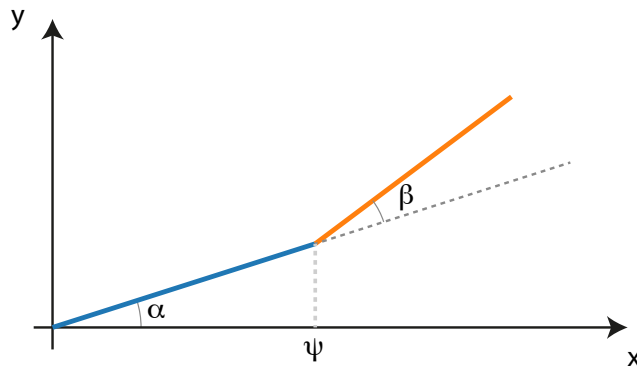


FIG. 2. Example of segmented line with one breakpoint.

$$(x - \psi)_+ = (x - \psi^s)_+ + (\psi - \psi^s)(-1)I(x > \psi^s) \quad (2)$$

which represents a first order Taylor expansion of $(x - \psi)_+$ at ψ_s . By defining $U^s = (x - \psi^s)_+$ and $V^s = -I(x > \psi^s)$, Equation 1 can be rewritten as

$$y = \alpha x + \beta U^s + \gamma V^s. \quad (3)$$

Thus, the piecewise linear curve is represented as a linear combination between variables x , U^s and V^s . Representing as x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n the values measured for, respectively, the x and y variables, we want to find β , γ and ψ_s which minimizes the sum of squared residuals for the model in Equation 3. Muggeo (2003) showed that this can be done according to the following procedure:

1. Set an initial value for ψ^s ,
2. Calculate $U^s = (x_i - \psi^s)_+$ and $V^s = -I(x_i > \psi^s)$ for each data point i ,
3. Fit the model in Equation 3 using linear regression,
4. Improve the breakpoint estimation according to $\psi^{s+1} = \gamma/\beta + \psi^s$,
5. Repeat 1-4 until convergence.

In case of multiple breakpoints, the process is similar. The breakpoints are modeled as

$$y = \alpha x + \sum_{k=1}^N \beta_k (x - \psi_k)_+ \quad (4)$$

where N is the number of breakpoints. β_k represents the difference in slopes between segments k and $k + 1$. Equation 4 can be rewritten as

$$y = \alpha x + \sum_{k=1}^N \beta_k U_k^s + \sum_{k=1}^N \gamma_k V_k^s. \quad (5)$$

The parameters of Equation 5 are found by least squares regression, and the breakpoints estimates are updated as $\psi_k^{s+1} = \gamma_k/\beta_k + \psi_k^s$.

The segmented regression requires the number of breakpoints to be known a priori. Muggeo and Adelfio (2010) defined a procedure for finding an appropriate number of breakpoints. First, the aforementioned segmented regression method is applied using a large number of candidate breakpoints. Then, during the optimization procedure a breakpoint is

removed if $\beta_k \approx 0$, which indicates that the slopes of segments k and $k+1$ are too similar, and if two breakpoints are too close to each other. Next, breakpoints that do not significantly contribute to the residual are removed using the least-angle regression algorithm (Efron et al. 2004).

For our analysis, the segmented regression was applied using the *segmented* package in the R language. The first point of each viewing profile was removed since it represents the number of article views along the first calendar month when the article was published. Thus, if the article was published at the end of the month, it will have an unreasonably low number of views for that month.

D. Statistical Model

To better understand the linear-piecewise nature observed in the studied view profiles, we propose a set of simple statistical models. Each of the models progressively incorporate more information about the data. In particular, it takes into account more relationships among parameters recovered from approximating the view profiles by using the segmented regression algorithm.

We start the analysis by applying the segmentation algorithm to all the curves, resulting in a set of piecewise linear curves. Each curve is given in terms of the breakpoints and slopes that estimates the original view profile. By considering only the curves with a fixed number N of segments, we extract the piecewise parameters given by the algorithm: l_i ($1 \leq i \leq N$), representing the length of the i^{th} segment between two breakpoints, and α_i ($1 \leq i \leq N$), the inclination of the i^{th} segment relative to the x-axis.

One of our goals is to produce synthetic profiles according to statistical models providing support to understand the real-world profiles. Here, we propose four types of models: null model, independent distribution model, Markov-1 univariate, and multivariate models.

For the models, we consider that l_i and α_i are random variables given by probabilities $P(\alpha_{i+1}|\alpha_i)$, $P(l_{i+1}|l_i)$, $P(l_{i+1}|l_i, \alpha_i)$, and $P(\alpha_{i+1}|l_i, \alpha_i)$ and employ the conditional probability equation, as follows

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \quad (6)$$

In our simplest model, which we call null model, α_i and l_i are generated following uniform distributions. Each α_i is drawn between $[0, 90)$ and the l_i in the range $[0, 1)$. The indepen-

dent distribution model takes into account independent densities. For samples obtained for a variable X , the probability density function (PDF) of the variable is estimated using kernel density estimation with a Gaussian kernel. The PDFs of the α_i and l_i variables were estimated from the articles data and used for generating synthetic values.

In the *Markov-1 univariate model*, given a sequential random variables X_i , the values of the next state $i + 1$ are drawn from the univariate conditional probabilities given by the current state X_i . More specifically, starting from an initial independent distribution of X_1 , the first state is generated. The next state is drawn from $P(X_{i+1}|X_i)$. To estimate $P(X_{i+1}|X_i)$, both the α_i and l_i parameters obtained for the data were partitioned into bins.

Finally, the Markov-1 multivariate model uses the combined joint probabilities for α_i and l_i to calculate α_{i+1} and l_{i+1} . More specifically, the initial independent joint distribution $P(\alpha_1, l_1)$ is used to generate the first states (α_1 and l_1 values), then the distributions $P(\alpha_{i+1}|\alpha_i, l_i)$ and $P(l_{i+1}|\alpha_i, l_i)$ are employed to calculate the next corresponding states.

IV. RESULTS AND DISCUSSION

In this Section, we discuss the adherence of the real data to segmented lines in Section IV A. The basics statistics of the real data are discussed in Section IV B. In Section IV C, we analyze the correlation between the parameters of the segmented lines describing the evolution of paper views. In Section IV D, we present a discussion on the clustering behavior of the segmented view profiles. Finally, in Section IV E, we provide an evaluation of the model aimed at reproducing the behavior of paper views along time.

A. Segmented regression adherence

The first step of the analysis is the application of the *segmented regression* algorithm to all the obtained view profiles. In order to do so, each view profile was normalized to the range $[0, 1]$ in both the time and number of views axes. Therefore, the cumulative number of views of an article for the most recent month in the dataset (September of 2019) is always 1. A preliminary analysis by visual inspection suggested that most of the profiles contain 5 or less linear segments. In order to check which curves best adhere to a structure of segmented lines, we computed the Root Mean Square Error (RMSE) (Hyndman and Koehler 2006)

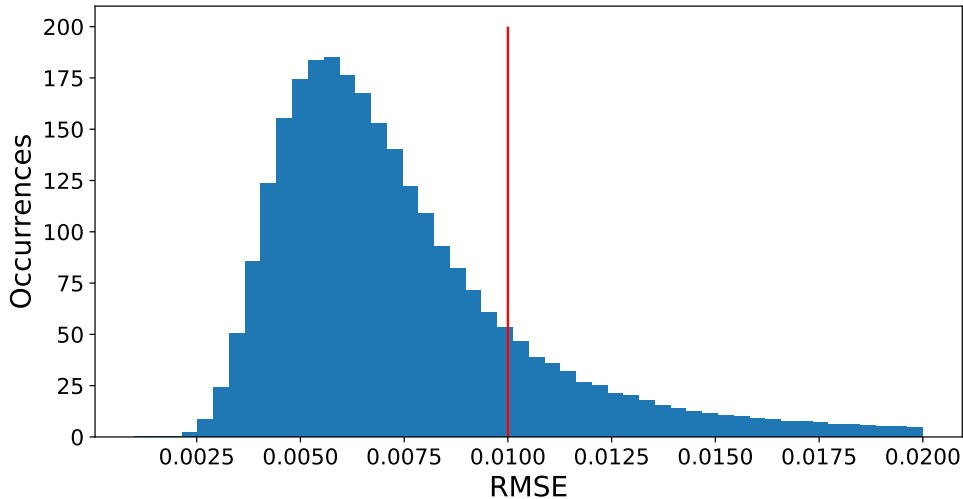


FIG. 3. RMSE distribution for the segmented regressions of the view profiles. The red line indicates the threshold RMSE value; only the curves on the left were selected to be used in the experiments. With the established threshold, roughly 80% of all view curves were analyzed.

between the curves derived from the algorithm and the original data. Figure 3 shows the distribution of the RMSE for the considered view profiles.

To test if the curves adhere to the segmented regression model, we chose a conservative threshold based on the obtained RMSE values. More specifically, we selected only the curves for which the regression resulted in an RMSE value lower than 0.01 (shown as a red line in Figure 3). Figure 4 shows two examples of views profiles (green dots) – characterized by a higher (a) and lower (b) adherence to the model – and their respective piecewise curves (red segments) together with the breakpoints (gray vertical lines). About 80% of the curves have been found to pass the RMSE test and were selected for further analyses.

In order to validate the hypothesis that real-world view profiles tend to present a polygonal structure, control synthetic profiles were generated. For each original profile, a synthetic one with the same number of points was created. Synthetic views were generated following a discrete uniform distribution $[0, h_M)$ for the number of views an article received in a month, where h_M is the largest number of views among all articles in the real data. We found a similar adherence compared to the original profiles when the *segmented* method was applied with the same parameters in synthetic profiles: about 78% of the lines passed the RMSE test. Despite this, there is a significant difference between the angle distribution in original and synthetic piecewise curves (see Figure S1 of the Supplementary Information (SI)). The

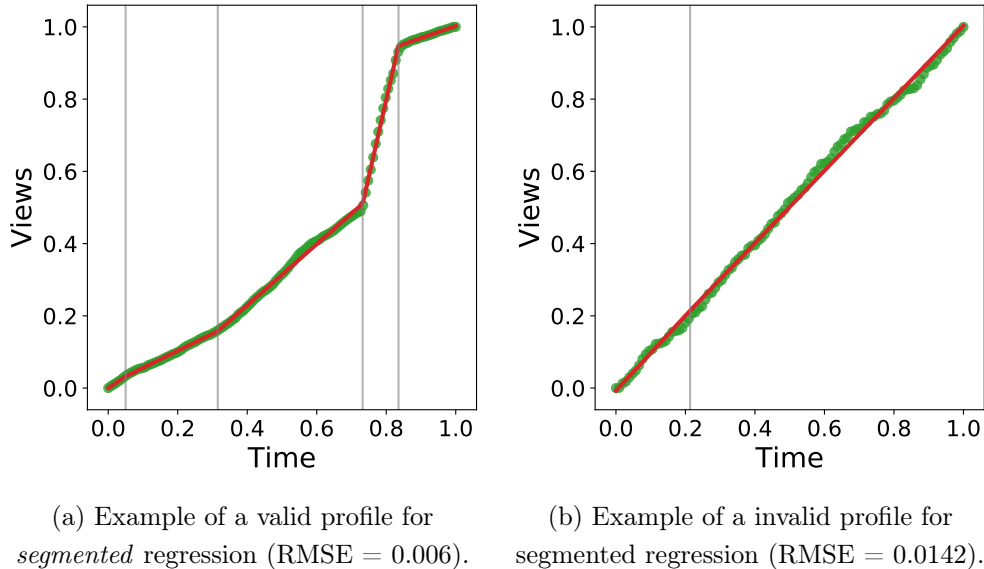


FIG. 4. Examples of view profiles and corresponding linear-piecewise curves. The *segmented* regression result is the red lines, the original values are the green points, and the breakpoints occur where there are the gray vertical lines.

latter is centered on 45-degree values. The differences between the above mentioned synthetic profiles and the real curves mean that the observed view curves for PloS ONE articles cannot be explained by this simple stochastic process.

B. Basic Statistics

We analyzed some basic statistics of the considered views curves, i.e., those that passed the segmented regression test described in Sections III C and IV A. Figure 5 shows the lifetime and cumulative views distributions of the views curves. The lifetime is the number of years since a certain article was published (up to 2016), and ranges predominantly between 4 and 7 years according to Figure 5(a). Also, most papers have from 1,500 to 3,500 views, as shown in Figure 5(b).

Most of the curves passing the RMSE test have been found to be modeled best by 5 linear segments. The proportion of curves found to be described by 2, 3, 4 and 5 intervals are: 0.1%, 4.5%, 26.4% and 69.0%, respectively. It is not a surprise since the segmentation algorithm chooses the best number of segments, and the tendency is to have more segments to better adhere to the regression and the original data. We also found that the lifetime average (and standard deviation) seems to increase with the number of segments of the

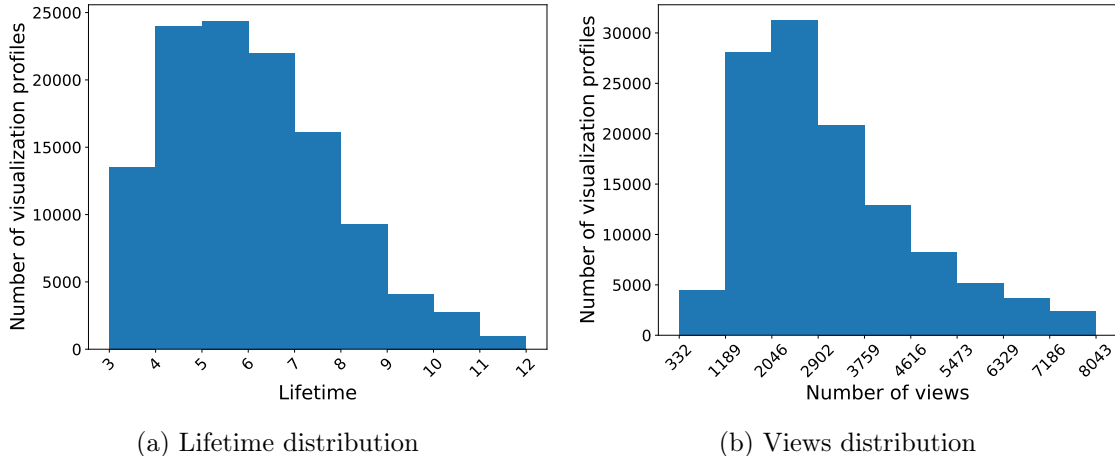


FIG. 5. Distributions of (a) lifetime (in years) and (b) number of views of the selected view profiles (after removing outliers). From a given profile, lifetime is the period since its publication to the moment data was collected, and views refer to the total paper views along its lifetime.

curves, this is shown in Figure S2 of the SI. This effect indicates that, as could be expected, the intricacy of the views profiles tend to increase with time.

From the segmented patterns, we extracted the parameters describing the curves according to the proposed segmentation method, more specifically, the angles α_i and segment lengths l_i . The respective parameter distributions are shown in Figure S4 of the SI. We found that the angles tend to become smaller along the consecutive segments, possibly reflecting the loss of visibility with time. The lengths of segments distributions become narrower with time, meaning that the monthly views of articles tend to change more rapidly as the article gets older.

C. Joint Distributions

An important step to understanding how subsequent segments are related is the analysis of the joint distributions among their parameters. In particular, we focus our analysis on the bivariate distributions of α_i and l_i for subsequent segments.

Figure 6 shows the joint density plots for l_i and l_{i+1} and their respective Pearson correlations, ρ , which displayed moderate negative values. In general, the first segment is small, as illustrated in Figure 6(a), which can be an effect of the time of response, given by outside factors, such as dissemination on the web, conferences, advertisements on magazines, posts on social networks, media coverage, the popularity of the topic, and citations from other

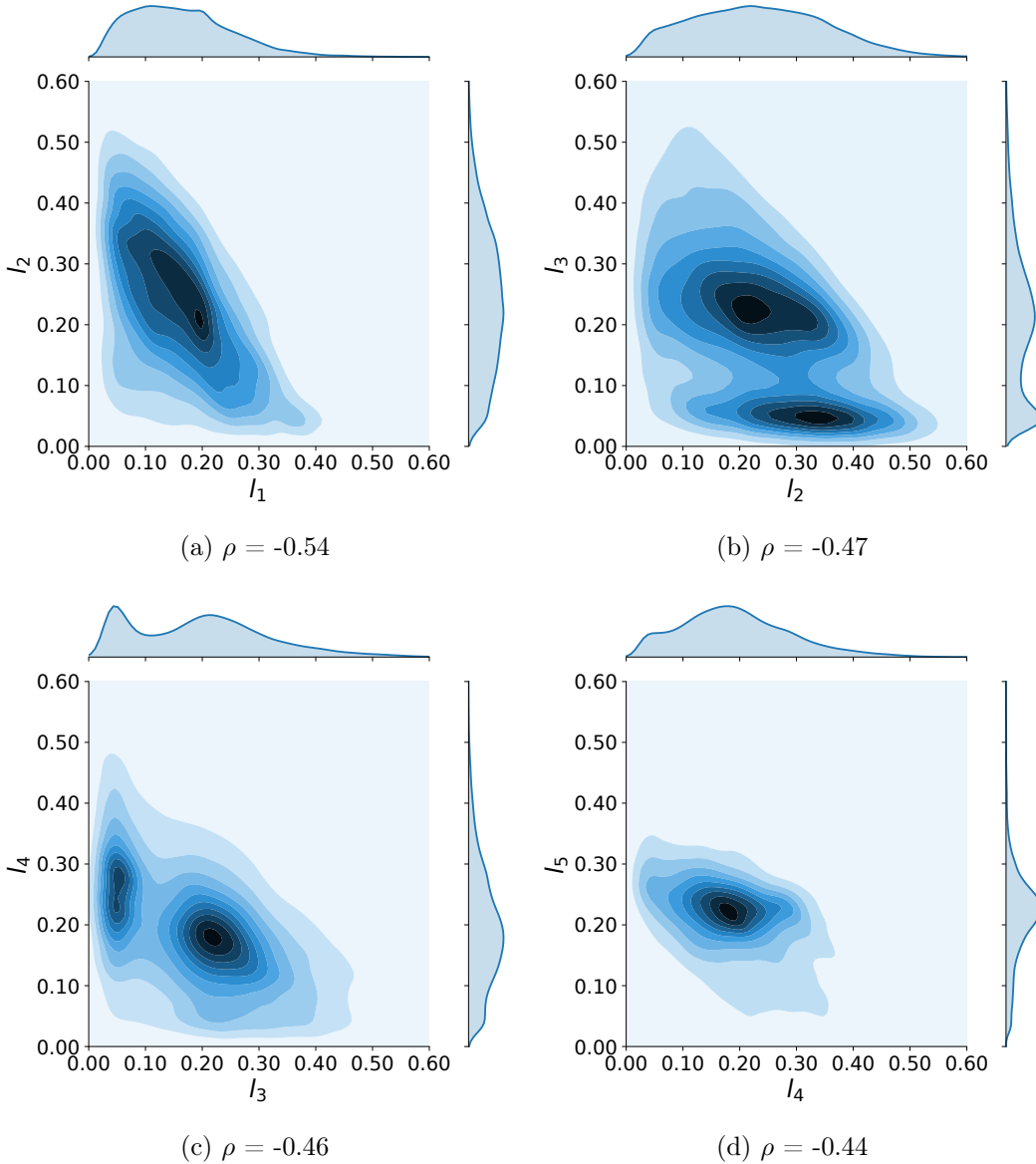


FIG. 6. Joint density plots and corresponding marginal probabilities of l_i and l_{i+1} for profiles with five segments. ρ is the Pearson correlation.

papers. Furthermore, similar outcomes were found for α_i and α_{i+1} (see Figure S3 of supplementary material). In this case, we observe correlations that are not particularly strong. Two groups seem to be present on all the plots. A more detailed analysis of the groups in the profiles is performed in Section IV D.

We also considered the relationships between α_i and l_i at subsequent segments. The ellipses plots shown in Figure 7 represent the correlations between the different parameter combinations of subsequent segments. In more detail, the colors and inclination indicate the

sign of the correlations, and negative correlations are plotted in blue while positive relations in red. The bigger the correlation magnitude, the stronger the color tone is, and more elongated are the ellipses. We found stronger correlations between α_i and l_i in the initial segments compared to those found for the subsequent segments. In general, given a linear segment i , the correlations are negative: high inclination angles occur jointly with short segments, and lower inclination angles tend to occur for longer segments. Contrariwise, in consecutive segments, the correlation is positive. This result indicates that short segments tend to be followed by low inclination angles and long segments by high inclination angles. A possible explanation of this effect could be related to a sudden increase in views following the dissemination of the paper. In general, this surge of interest is not sustained along time.

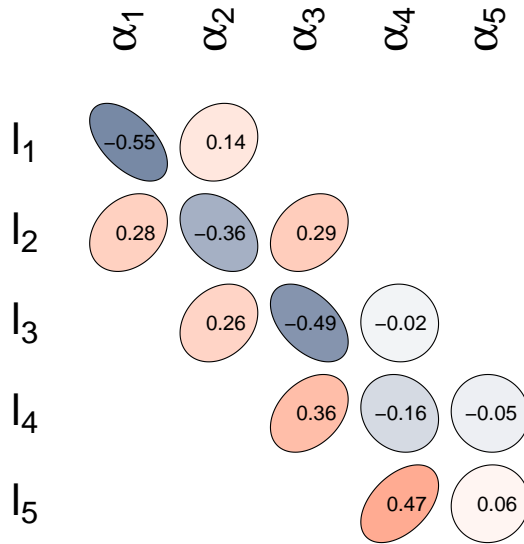


FIG. 7. Correlations between the variables of the piecewise curves shown as ellipses. The Pearson correlation coefficients were calculated by considering only the curves with five segments.

D. Clustering Analysis

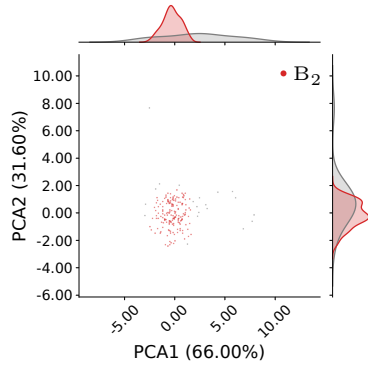
Before proposing a stochastic model to reproduce the observed view profiles, it is interesting to check if we can find different patterns of view profiles that could be understood as clusters. This type of analysis can give us insights regarding the proposed model. More

specifically, knowing about the existence of groups, it is possible to create separate models incorporating the singularities of the models. For that, a clustering analysis was performed in the measured segmented curves parameters. Groups were obtained by running a Hierarchical Clustering algorithm with the single-linkage criteria and considering Euclidean distances. We set the number of clusters to a maximum of three since other values led to less defined groups. For that analysis, each curve is represented by their set of segmented parameters α_i and l_i .

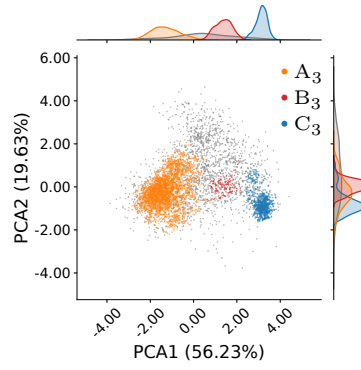
The number of parameters defining the curves is at least four, corresponding to the simplest case in which only one breaking point exists. Thus, for view purposes, we employed a Principal Component Analysis (PCA) projection to reduce the dimensionality of the set of parameters (Gewers et al. 2018). The panels in Figure 8 show the clusters of curves for each of the adopted number of segments. In each of these scatterplots, each point represents a curve in the projected space. We named the three groups as A_i , B_i , and C_i , where i refers to the number of segments. In the figure, the marginal distributions help to distinguish between the overlapping groups. Groups C tend to be more well-defined and separated from the others. The first two Principal Components accounts for 97.6% of the total variance in Figure 8(a), 75.86% in Figure 8(b), 58.86% in Figures 8(c) and 8(d), 43.24% in Figures 8(e) and 8(f). In the case of 2 and 3 segment clusters, only two principal components are enough to explain the variance of the data. However, for 4 and 5 segments, the third principal component is needed. This component is also shown in Figure 8(a) for these cases.

In general, the obtained clusters tend to overlap more as the number of segments increases. The marginal distributions are more separated when the curves have two or three segments in the PCA, their peaks being considerably distinct (Figures 8(a) and 8(b)). When the curves have four or five segments, the distributions become flatter and indicate a larger overlap among the groups (Figures 8(c) and 8(d)). One possible explanation for the change in the clustering structure is the progressive increase of view profile types with the number of segments.

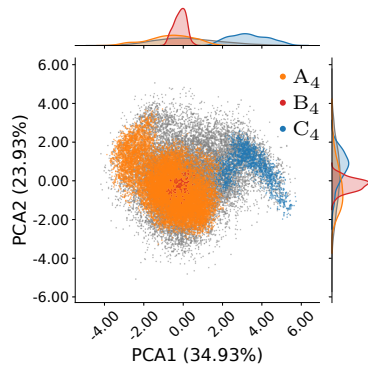
In the following analysis, we investigate and determine which are the typical features of the views of papers in groups $\{A_3, A_5\}$, $\{B_3, B_5\}$ and $\{C_3, C_5\}$. More specifically, the curve obtained for each view was assigned symbols “+” and “-” reflecting if the subsequent angle increased or decreased respectively to the previous angle of the previous segment. Note that this analysis starts on the second angle. The resulting lists of “+” and “-” are then compared



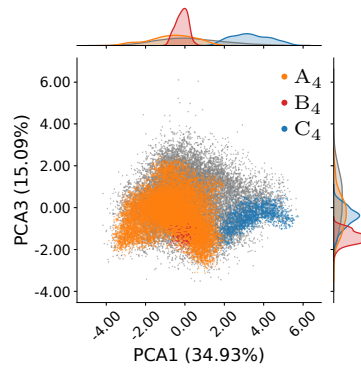
(a) The 2-segment profiles clusters.



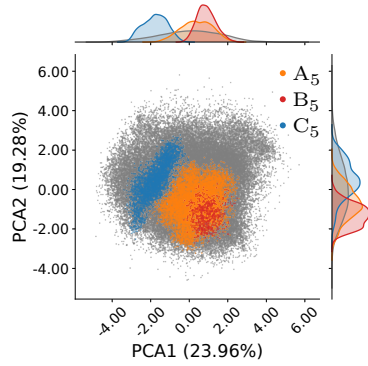
(b) The 3-segment profiles clusters.



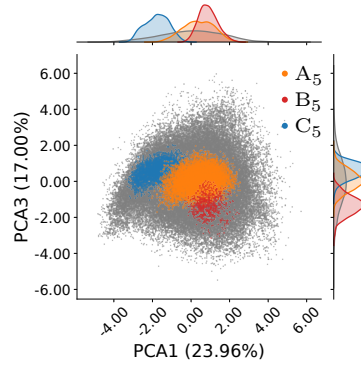
(c) The 4-segment profiles clusters (PCA1 and PCA2).



(d) The 4-segment profiles clusters (PCA1 and PCA3).



(e) The 5-segment profiles clusters (PCA1 and PCA2).



(f) The 5-segment profiles clusters (PCA1 and PCA3).

FIG. 8. Scatter plots showing the clusters of the view profiles and the respective marginal distributions. The points correspond to the PCA projection of the angle and length of the segments of the profiles. The colors (orange, blue, and red) refer to the three detected groups, while the gray points correspond to views not assigned to these 3 groups.

and organized into prototypes. For instance, in the case of 3 segments, we have: “--”, “-+”, “+-”, and “++”. The relative frequencies of each of these prototypes for papers in clusters A, B and C was then estimated, and the most frequent patterns are depicted in Figure 9. The signatures respective to the views with 3 and 5 segments are shown in Figures 9(a-c) and (d-f), respectively. In the case of the views with 3 segments, the most frequent group (a) is characterized by successive decreases of the view rates, indicating progressive assimilation by the respective research community. Contrariwise, the situations in the groups shown in (b) and (c) are characterized by a relative increase of views along the intermediate segment. This could have been implied by an event like dissemination in the news, meetings, or social networks. In the case of groups shown in (c), the increase in views occurs at a shorter time span than in (b).

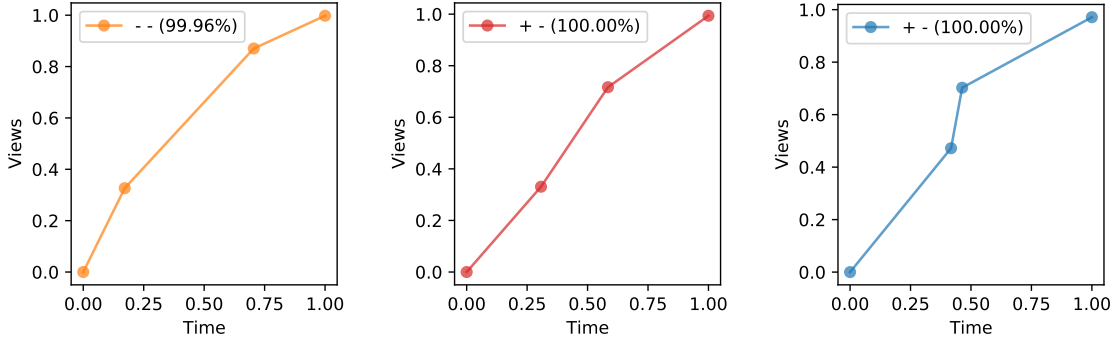
The results for the articles with 5 segments obtained for the least frequent group (e), similarly to the group (a), was characterized by successive reductions of the viewing rate (“- - - -”). The other two groups, (d) and (f), are characterized by respective prototypes “- - + -” and “- + - -”. Therefore, cases (e) and (f) share the same prototype as (b) and (c).

E. Models Adherence

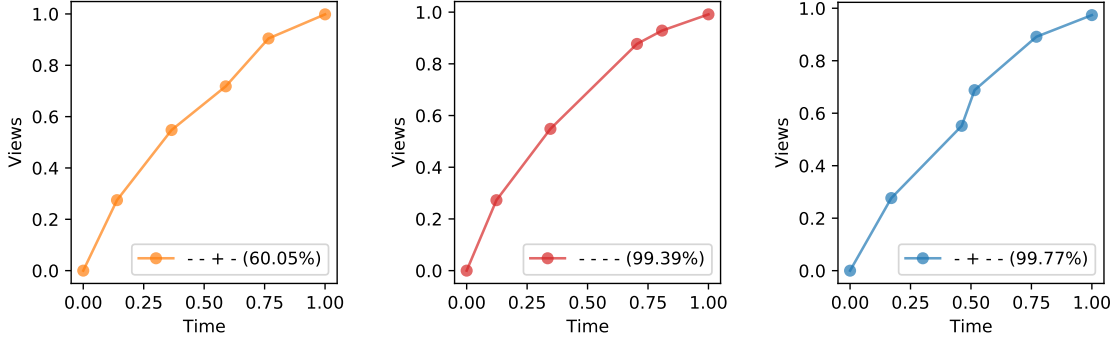
In this section, we check if the proposed Markov models can reproduce the joint distributions of the curves parameters (α_i and l_i) as expressed in terms of their principal component projections, shown in Figure 10. Observe that each line in this figure corresponds to one of the four considered models (see Section III D). The marginal densities are also depicted along the respective axes. For each of the four types of models, the cases corresponding to each of the three identified clusters were adjusted separately, and then combined when obtaining the principal component projection.

Then, the obtained density surfaces were compared by calculating absolute point-to-point differences (ε) between the original and synthetic 2d histograms of the PCA data (third column in Figure 10), and then adding all these values into the single error parameter ε , computed as

$$\varepsilon = \sum_x \sum_y |\pi_o(x, y) - \pi_s(x, y)|, \quad (7)$$



(a) Cluster A_3 (total of curves: 2677). (b) Cluster B_3 (total of curves: 143). (c) Cluster C_3 (total of curves: 804).



(d) Cluster A_5 (total of curves: 7970). (e) Cluster B_5 (total of curves: 661). (f) Cluster C_5 (total of curves: 3045).

FIG. 9. Average curves obtained for the most frequent prototype of each cluster. The legend indicates the percentage of curves characterized by the most frequent prototype. The x-axis represents time, and the y-axis represents the cumulative number of views along time.

where π_o and π_s are the surfaces corresponding to the original and synthetic data, respectively. Lower values of ε indicate more accurate models.

As expected, the Null model resulted in the worst approximation of the curves (as seen in Figure 10(c)), and the quality of the models increases as we incorporate additional statistical information. The independent distribution model, shown in Figure 10(d-f), better approximates the original profiles without considering conditional probabilities (i.e. memory). However, this model is unable to capture the medium scale details in the original distribution, and also broke the cluster.

The models based on Markov-1 take into account not only the independent conditional distributions but also the parameters of consecutive segments. The Markov-1 univariate synthetic profiles resulted in better approximations of the real view profiles (Figure 10(g-i)).

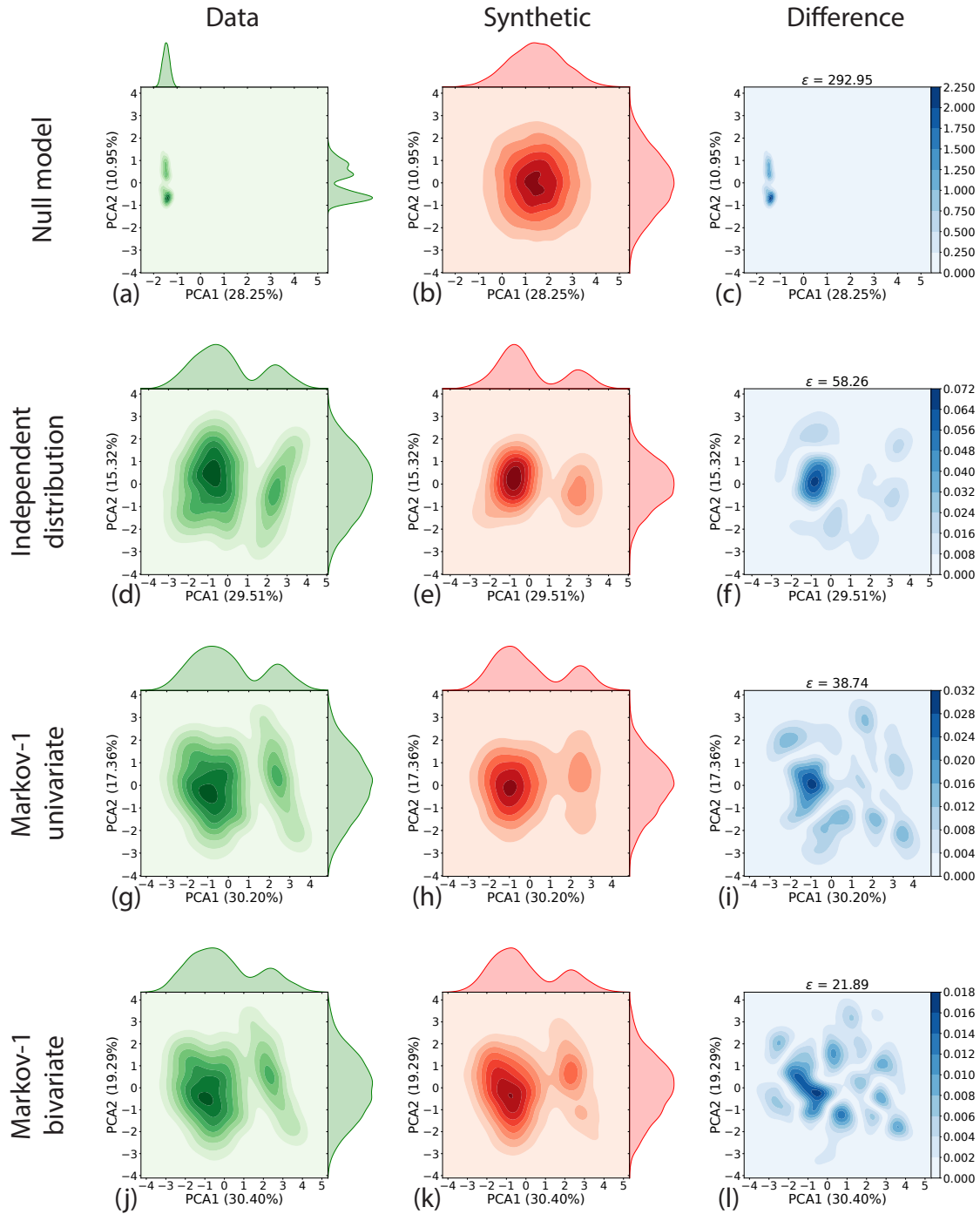


FIG. 10. Comparison of the original profiles and the synthetic profiles produced according to the considered models. The three columns correspond respectively to the original data distributions along the PCA axes, the synthetic distributions and the difference. Note that the differences in density at the center of plot (c) are too small compared to the differences caused by the peaks of the original profiles. Note that for each model, a new PCA projection is obtained since it incorporates both the real and synthetic data.

Finally, the Markov-1 multivariate model produced synthetic profiles more similar to the

original counterparts (Figure 10(j-l)). The bivariate distribution of α and l provide the best approximation of the original data among the considered models, as it yields the smallest ε value. This suggests that not only the angles of previous segments but also their lengths are important subsidies in predicting the parameters of the next segments.

All in all, the obtained results confirm to a good extent the initial hypothesis that the view profiles are not trivial or random and have intrinsic structure that were progressively reflected by modeling approaches taking into account additional information. This implies that there are interesting real-world effects and mechanisms implementing the types of observed structure. In particular, we have memory effects and time dependencies, in the sense that the properties of one segment tend to correlated with subsequent segments. These effects could be hypothetically related to tendencies such as a brief surge of interest, e.g. caused by media dissemination, would be followed by a longer period of less intense inclination.

V. CONCLUDING REMARKS

Science is inherently a collaborative endeavor. Therefore, the speed of dissemination of new ideas has a relevant impact in the development of novel theories and experiments. Traditionally, the impact of papers has been studied in terms of the number of citations, but since the World Wide Web became the main medium for publishing papers, the development of new data aggregation tools led to the definition of many alternative metrics (Sud and Thelwall 2014). One of the simplest of such metrics is the number of page views. Measuring page views is relatively simple and can usually be done with arbitrary granularity (hourly, daily, monthly, etc). Compared to citations, the number of views also tends to display a much lower delay to important events such as publication and conference presentation.

Here, we studied to what extent the monthly number of views for articles published in the PLoS ONE journal present a polygonal structure, which constitute the main question of the present work. A key observation regarding the number of views of the articles in the PLoS ONE dataset was used in the analysis: articles tend to display periods of relatively constant number of monthly views, with sharp changes in views between such periods. This hypothesis was investigated throughout the work by considering the cumulative number of article views. If the hypothesis is true, the cumulative number of views should be correctly represented by a piece-wise linear function.

A segmented least squares regression methodology (Muggeo 2003) was applied to identify breakpoints between linear segments in cumulative article views, and the length l_i and angle α_i of each segment were measured and used as parameters of four models for generating synthetic article views profiles. The models took into account progressively more information about the profiles, so as to allow the identification of the most relevant properties.

Several interesting results were obtained. It was verified that the segmented regression led to a lower RMSE than in the case of synthetic profiles generated from a model which took into account randomly generated number of monthly views. The result indicates that representing the cumulative number of article views by a piece-wise linear function led to a relatively low regression error. Thus, the profiles can be modeled by linear segments. Another important result was the observation of two view profiles, one corresponding to a relatively low initial slope followed by a higher slope and another presenting only slopes that decrease with time. In order to better interpret these groups, additional metadata is necessary and is a future development of this work. Regarding the synthetic models, it was found that curves generated from α_i and l_i sampled independently with the same distribution as the real data led to profiles that approximated well the real profiles. Taking into account conditional probabilities between subsequent segments and between α_i and l_i led to improved models. Although we found many interesting results, our study also includes some limitations. The analysis was performed only with a single dataset, obtained from PLoS ONE papers, and cannot be generalized for all journals.

For future developments, additional metadata about the papers can be taken into account in the analysis. In particular, it would be interesting to investigate how the views dynamics changes according to the subject area and authors institution. It would also be interesting to associate social network data with the observed profiles. For instance, verify if the identified breakpoints correlate with messages published by power users (Eysenbach 2011) in a social network about the article. It is also worth investigating how bibliometric networks can affect view patterns along time (Costa 2006, de Arruda et al. 2017). Finally, we could also analyze if other factors such as authors and topics visibility can affect the patterns of view profiles in papers (Amancio 2015, Corrêa Jr et al. 2017, Larivière et al. 2016, Lu et al. 2019).

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. C. H. Comin thanks FAPESP (grant no. 18/09125-4) for financial support. F. N. Silva acknowledges CAPES and FAPESP (grant no. 15/08003-4). H. F. de Arruda acknowledges FAPESP for sponsorship (grants 2018/10489-0 and 2019/16223-5). D. R. Amancio thanks FAPESP (grant no. 16/19069-9) and CNPq (grant no. 304026/2018-2). L. da F. Costa thanks CNPq (grant no. 307085/2018-0) and NAP-PRP-USP for support. This work has been supported also by the FAPESP grant 15/22308-2.

-
- D. R. Amancio. Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, 105(3):1763–1779, 2015.
- D. R. Amancio, O. N. Oliveira Jr, and L. F. Costa. Three-feature model to reproduce the topology of citation networks and the effects from authors’ visibility on their h-index. *Journal of Informetrics*, 6(3):427–434, 2012.
- J. Bollen, H. Van de Sompel, J. A. Smith, and R. Luce. Toward alternative metrics of journal impact: A comparison of download and citation data. *Information processing & management*, 41(6):1419–1440, 2005.
- L. Bornmann. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4):895–903, 2014.
- T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.
- B. Chen, D. Deng, Z. Zhong, and C. Zhang. Exploring linguistic characteristics of highly browsed and downloaded academic articles. *Scientometrics*, 122(3):1769–1790, 2020.
- E. A. Corrêa Jr, F. N. Silva, L. F. Costa, and D. R. Amancio. Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, 11(2):498–510, 2017.
- L. F. Costa. Learning about knowledge: A complex network approach. *Physical Review E*, 74(2):026103, 2006.
- H. F. de Arruda, F. N. Silva, L. F. Costa, and D. R. Amancio. Knowledge acquisition: A complex networks approach. *Information Sciences*, 421:154–166, 2017.
- J. C. de Winter. The relationship between tweets, citations, and article views for plos one articles. *Scientometrics*, 102(2):1773–1779, 2015.
- Y. Duan and Z. Xiong. Download patterns of journal papers and their influencing factors. *Scientometrics*, 112(3):1761–1775, 2017.
- B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.
- M. Erdt, A. Nagarajan, S.-C. J. Sin, and Y.-L. Theng. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, 109(2):1117–1166, 2016.

- G. Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4):e123, 2011.
- F. Galligan and S. Dias-Correia. Altmetrics: Rethinking the way we measure. *Serials review*, 39(1):56–61, 2013.
- F. L. Gewers, G. R. Ferreira, H. F. de Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. F. Costa. Principal component analysis: A natural approach to data exploration. *arXiv*, 1804.02502(1):1–33, 2018.
- W. Huang, P. Wang, and Q. Wu. A correlation comparison between altmetric attention scores and citations for six plos journals. *PLOS ONE*, 13(4):1–15, 04 2018. doi:10.1371/journal.pone.0194962. URL <https://doi.org/10.1371/journal.pone.0194962>.
- R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.
- J. Ioannidis, K. W. Boyack, H. Small, A. A. Sorensen, and R. Klavans. Bibliometrics: Is your most cited work your best? *Nature News*, 514(7524):561, 2014.
- A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
- H. R. Jamali and M. Nikzad. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2):653–661, 2011.
- V. Larivière, C. Ni, Y. Gingras, B. Cronin, and C. R. Sugimoto. Bibliometrics: Global gender disparities in science. *Nature News*, 504(7479):211, 2013.
- V. Larivière, N. Desrochers, B. Macaluso, P. Mongeon, A. Paul-Hus, and C. R. Sugimoto. Contributorship and division of labor in knowledge production. *Social Studies of Science*, 46(3):417–435, 2016.
- C. Lu, Y. Bu, X. Dong, J. Wang, Y. Ding, V. Larivière, C. R. Sugimoto, L. Paul, and C. Zhang. Analyzing linguistic complexity and scientific impact. *Journal of Informetrics*, 13(3):817–829, 2019.
- H. F. Moed. Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10):1088–1097, 2005.
- V. M. Muggeo. Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19):3055–3071, 2003.

- V. M. Muggeo and G. Adelfio. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166, 2010.
- D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.
- Note1. <https://journals.plos.org/plosone>.
- J. L. Ortega. The life cycle of altmetric impact: A longitudinal study of six metrics from plumx. *Journal of Informetrics*, 12(3):579–589, 2018.
- F. D. Peat. *From certainty to uncertainty: The story of science and ideas in the twentieth century*. Joseph Henry Press, 2002.
- T. V. Perneger. Relation between online “hit counts” and subsequent citations: prospective study of research papers in the bmj. *Bmj*, 329(7465):546–547, 2004.
- J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*, 2012.
- C. Schlögl, J. Gorraiz, C. Gumpenberger, K. Jack, and P. Kraker. Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics*, 101(2):1113–1128, 2014.
- X. Shuai, A. Pepe, and J. Bollen. How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one*, 7(11), 2012.
- R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, pages 33–40, 1962.
- P. Sud and M. Thelwall. Evaluating altmetrics. *Scientometrics*, 98(2):1131–1143, 2014.
- M. Thelwall. Mendeley reader counts for us computer science conference papers and journal articles. *Quantitative Science Studies*, pages 1–13, 2019.
- M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5), 2013.
- E. K. Tokuda, C. H. Comin, and L. d. F. Costa. Revisiting agglomerative clustering. *arXiv preprint arXiv:2005.07995*, 2020.
- L. Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016.
- X. Wang, C. Liu, Z. Fang, and W. Mao. From attention to citation, what and how does altmetrics work? *arXiv preprint arXiv:1409.4269*, 2014.

A. B. Watson. Comparing citations and downloads for individual articles at the journal of vision.
Journal of vision, 9(4):i-i, 2009.

SUPPLEMENTARY INFORMATION

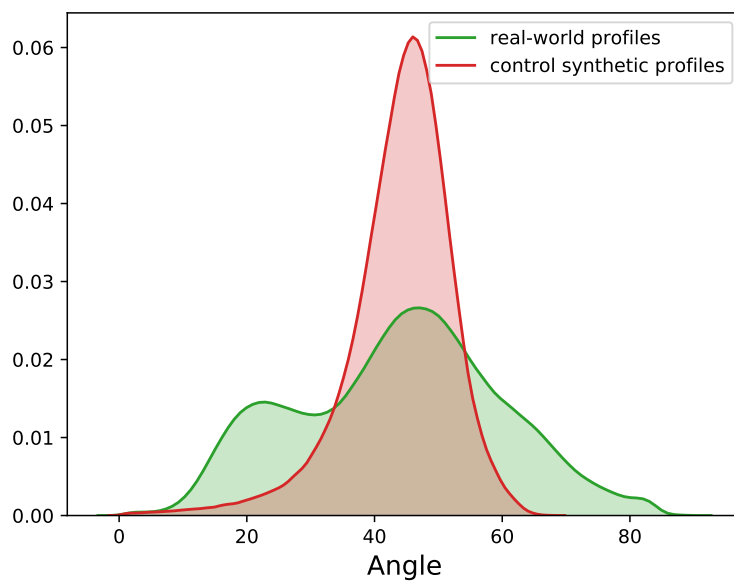
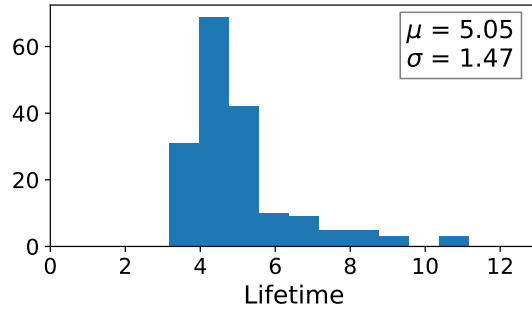
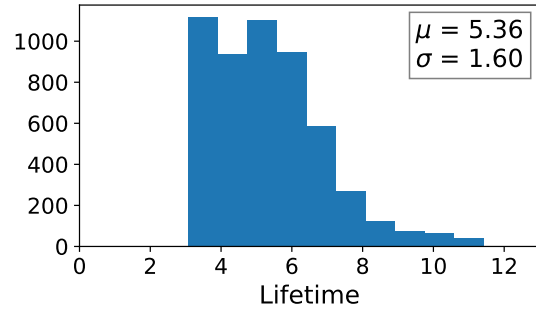


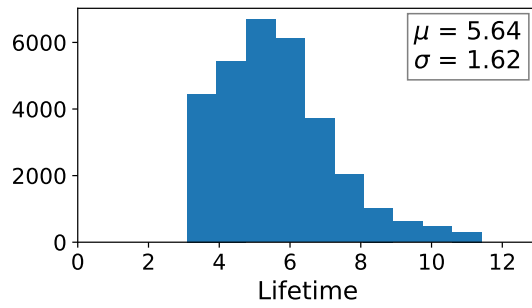
FIG. S1. Angles distributions of real-world (green) and control (red) synthetic profiles. The distributions are clearly different, and a higher deviation in angles is observed for the real data.



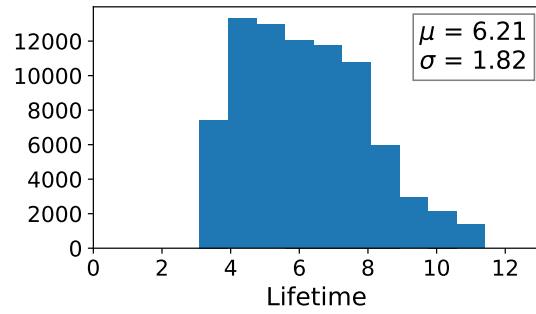
(a) Lifetime distribution for 2-segment profiles.



(b) Lifetime distribution for 3-segment profiles.



(c) Lifetime distribution for 4-segment profiles.



(d) Lifetime distribution for 5-segment profiles.

FIG. S2. Distributions of lifetime grouped by profiles with the same number of segments. Lifetime average and standard deviation are shown at the upper right corner.

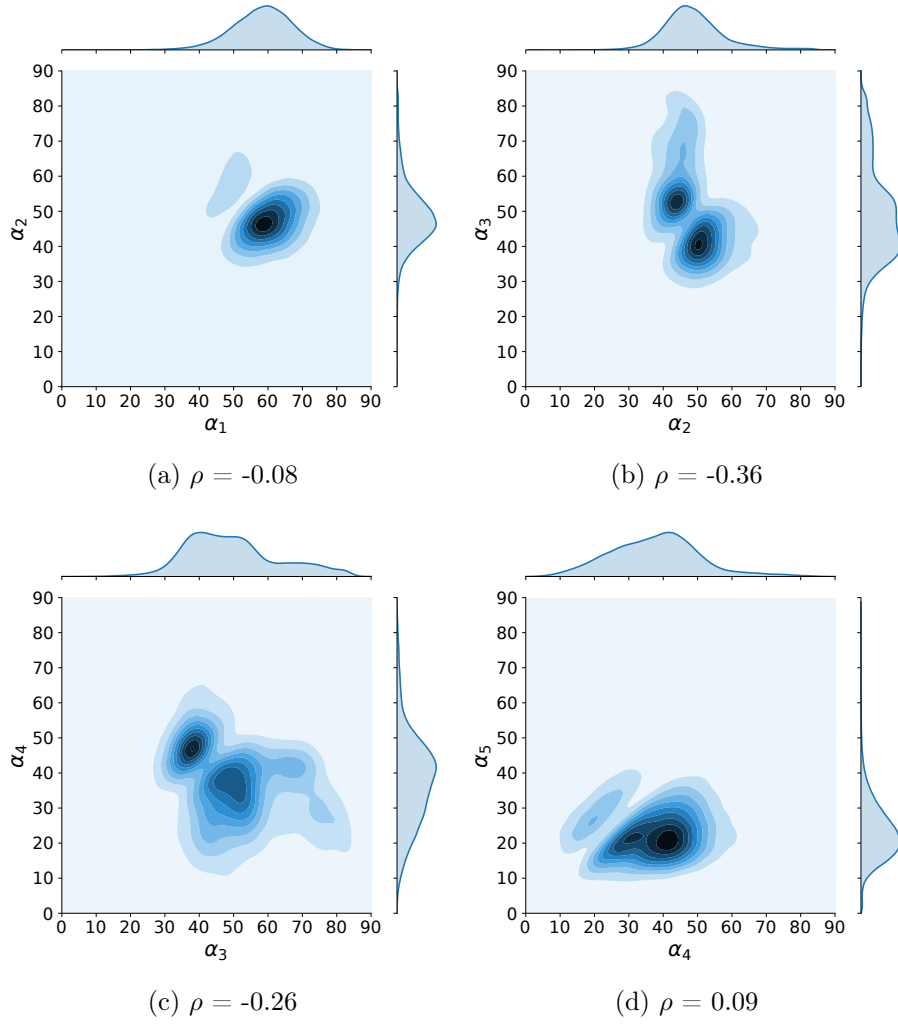


FIG. S3. Joint density plots and corresponding marginal probabilities of α_i and α_{i+1} for profiles with five segments. ρ is the Pearson correlation.

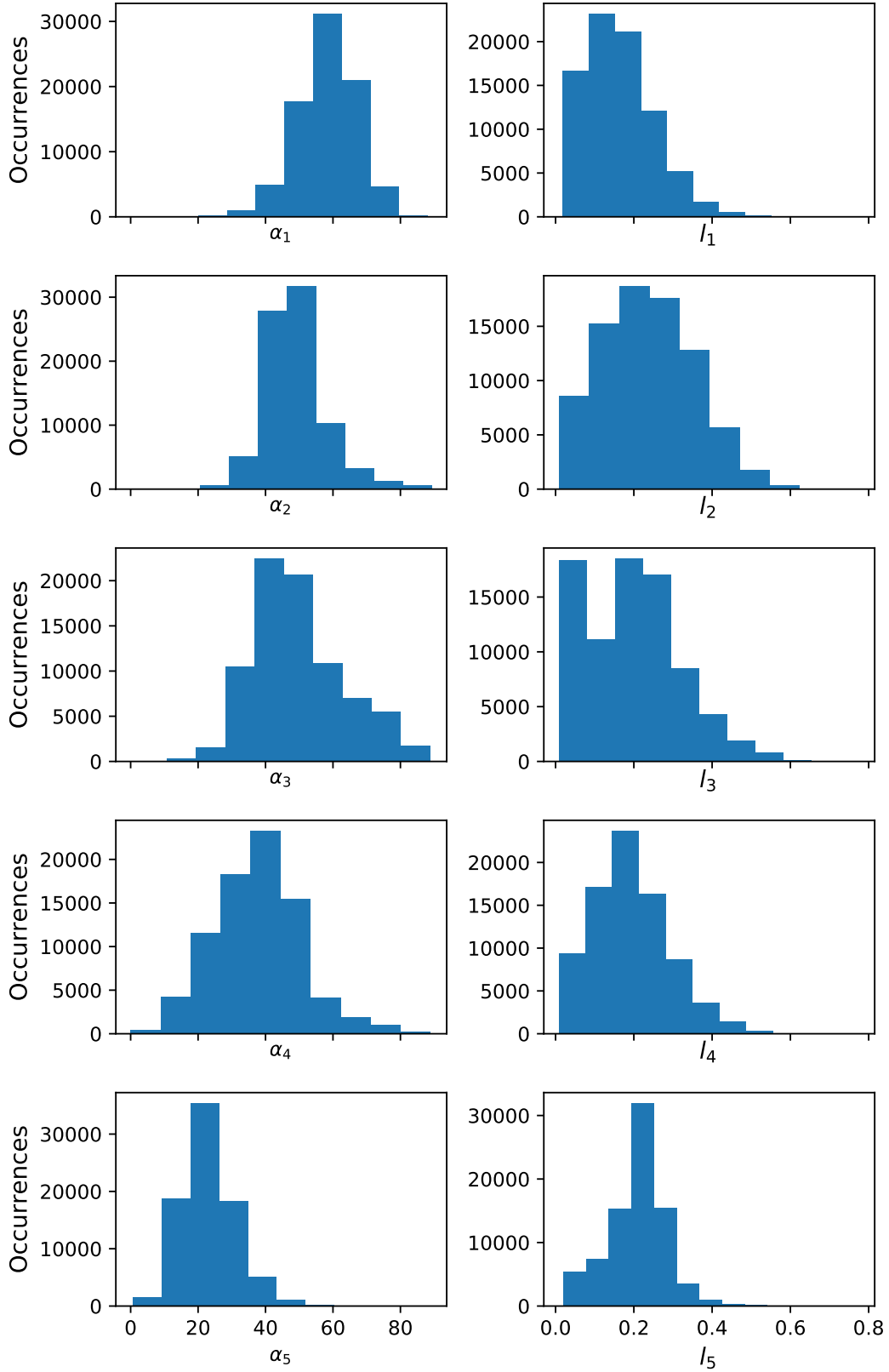


FIG. S4. Distribution of the segment parameters among all the obtained curves.