INCOPRO

Statistical Analysis Report

s.to

[redacted] | 26 November 2020

Contents

Statement of [redacted]

1. I have been instructed by the Motion Picture Association to provide a statistical analysis of the proportion of copyright protected content on the website identified in the summary table in the 'Results' section of this report (the **Target Website**).

2. I work as a Data Scientist and Statistician for INCOPRO Limited (**INCOPRO**) and have held this position since 18 December 2014. INCOPRO is a limited company incorporated in England and Wales which monitors the online environment for infringement of intellectual property rights. INCOPRO specialises in the identification of infringement of intellectual property and in the provision of reporting and other services to assist in the evaluation of action required to protect intellectual property online. Part of my role within INCOPRO is to perform statistical analysis and to develop systems to apply a statistical approach to the collection of data.

3. I have 17 years' professional experience in various roles with a strong quantitative and computational element, including 5 years as an interdisciplinary Post-Doctoral Research Associate at Imperial College London specialising in the mathematical modelling and statistical analysis of large data sets. I hold a BSc in Biological Sciences from Birkbeck College London, an MRes in Modelling Biological Complexity and Bioinformatics from University College London (UCL) and a Ph.D. in Mathematical Modelling from the University of London. I am a Fellow of the Royal Statistical Society.

4. In order to enable the analysis summarised in this report, I was supplied with data in the form of a website sample scrape undertaken by Incopro's Analysts (the **Analysts**).

5. I confirm that I have made clear which facts and matters referred to in this report are within my own knowledge and which are not. Those that are within my own knowledge I confirm to be true. The opinions I have expressed represent my true and complete professional opinions on the matters to which they refer.

[redacted]

Date: 26 November 2020

Executive Summary

6. I analysed the [redacted] on the Target Website using recognised methods of statistical sampling.

7. My conclusion is that [redacted]

Data Collection

8. On 10 November 2020, the Analysts used random sampling to scrape (or copy) 100 links to available TV content on the Target Website.

9. The Analysts produced the scraped sample, set out as a computer text file, with each line in the file representing a link to relevant content. Hereafter, I will refer to the computer file as the "sample".

Statistical Sampling and Analysis

10. Inferential statistics permits the drawing of conclusions as to the characteristics of the whole population based on the results of a random sample drawn from the whole population. By undertaking this sampling procedure, I wanted to be able to draw conclusions as to [redacted] made available by the Target Website based on a sample drawn from each of them. Taking a random sample ensures a good representative sample because it removes any human influence, and therefore subjectivity or bias from the process. In contrast, any deterministic method, however arbitrary or devoid of subjectivity, would run the risk of returning biased results. To evaluate [redacted] on the Target Website a random sample of the entries was drawn from each of them to be analysed manually.

11. The simplest method of drawing a sample from a population is sampling with replacement. This method was implemented in respect of all Target Website.

12. Sampling with replacement means that individual records from the population might appear more than once in the sample. This method ensures that the precision of the results depends on the size of the sample only. In contrast and counter-intuitively, the sampling rate (defined as the ratio of the sample size to the population size) plays no role in that precision. This is because sampling with replacement ensures that each time a record is drawn from the population, the probability that [redacted] remains constant throughout the process. In other words, the sampling rate is immaterial because the underlying population size is immaterial.

13. In essence, the process replicates a series of random experiments with a binary outcome [redacted] and recording the overall result. The canonical example of such a random experiment is throwing a coin and checking on which side it lands. The proportion of heads obtained by throwing a well-balanced coin is more likely to be

close to ½ with a large number of throws. Likewise, one is less likely to obtain an unbroken series of heads (or tails, for that matter) if that coin is thrown a large number of times. By sampling randomly with replacement one aims to replicate this kind of property and thereby harness the laws of probability. Of course, the "coin" that is virtually thrown might be unbalanced and the degree to which it is unbalanced is investigated through sampling.

14. The sample size of 100 was selected on the basis that this provides enough precision to the results to be able to confidently reach robust conclusions as to the proportion of content that is commercially available or otherwise protected by copyright made available by the Target Website.

15. Subsequently, the Analysts manually analysed the 100 selected entries for the Target Website to determine the number of entries that related [redacted] The Analysts used the following methodology:

- They analysed every record starting with the first.
- For each entry, they checked the [redacted]
- Where they were unable to identify an entry [redacted]

16. The list of content titles contained within the sample and analysed using the above method can be found in Appendix 1.0: List of Titles.

17. The results of this analysis enabled me to make inferences about [redacted] on the Target Website as a whole.

18. To reflect the fact that the entire population has not been analysed, the results of inferential statistics are typically reported by reference to a degree of confidence and/or precision: it is not valid to just return a single indicator, [redacted] It needs to be accompanied with complementary indicators describing the confidence and/or precision. Here, I have chosen to present the results in two different ways.

[redacted]

19. Firstly, by reference to a 99% confidence interval, which is the traditional way to report results from surveys and polls. This means that instead of a single value, a range of values is returned and there is a 99% chance that the corresponding population value is within this range.[2]

20. Secondly, by contrasting the probability that the population proportion is above a certain value to the probability that it is below that value. This allows me to assert, more directly, where the preponderance of evidence lies. In order to do this I define the following descriptive proportions:

- A 'large majority': 90%.
- A 'sizeable majority': 80%.
- A 'simple majority': 50%.

21. The obvious consequence of those definitions is that asserting that the large majority of the examined content of a website is infringing is a stronger statement than asserting that a sizeable majority is, which in turn is a stronger statement than asserting that a simple majority is.

22. It is important to note that the results obtained are not absolute because only a portion of the underlying population was analysed. The robust statistical methodology employed here is designed to mitigate as far as possible the likelihood of an erroneous result.

23. Both of the series of indicators were computed using another bespoke computer program written in "R", a computer language and environment widely used and recognised to perform statistical analysis, which has become a standard tool within the statistical community and with which I am very familiar. The raw results that the Analysts communicated to me were fed to that program to obtain the indicators presented below.

24. Within that program, I used specialised commands existing within the R environment implementing the calculations of the aforementioned indicators.[3] The mathematical formulas involved in these calculations are quite complicated and would otherwise have taken a long time to deploy using less specialised tools. I performed so-called sanity checks to verify the soundness of the computer program's output by using approximate formulas for the confidence intervals and comparing these results to the ones returned by the program. These checks were performed before the actual data presented below was fed to the program and were successful, which convinced me that the program could be confidently used with real data.

Results

25. In respect of s.to I was informed by the Analysts that they were able to match 100 entries from the analysed sample of 100 entries as [redacted] On that basis, I conclude with 99% confidence [redacted] comprised between [redacted] In addition, the probability that this [redacted] very close to 100% whereas the

[2] Strictly speaking that probability of 99% refers to the range itself rather than the population value: there is a 99% chance that the range comprises the real value.

[3] The confidence intervals were computed using the "exact" method in the R binom package. The second set of indicators was determined using Bayesian inference with a uniform prior on the interval [0,1]. The prior and posterior distribution was modelled with the Beta distribution as outlined in *Doing Bayesian Analysis* by John K. Kruschke (Chapter 6).

probability that [redacted] is very small: [redacted][4]. Therefore the preponderance of evidence overwhelmingly lies towards the fact that [redacted] on s.to [redacted]

26. The table below collates the results of the sample analysis for the Target Website. The preponderance of evidence lies towards the fact that [redacted] of the examined content of the Target Website is [redacted] overwhelmingly so in this case.

[4] In order to represent a very small number it is convenient and customary to use either scientific or E notation where either an 'E' or an 'e' represents "times ten raised to the power of". In this case, [redacted] in words would be [redacted] times 10 raised to the power of -3". Converting to standard notation [redacted] would therefore equal 0.00239% (there are 2 zeros following the decimal point and preceding the [redacted].

INCOPRO

Target Website	Categories Scraped	Sample size	No. of links matched to content that is commercially available	[redacted]	Lower bound of 99% confidence interval	Upper bound of 99% confidence interval	[redacted]	Definition of "the majority"
s.to	TV	100	100		94.84%	100%		90%

INCOPRO

Appendix 1.0: List of Titles

No.	[redacted]	No.	[redacted]
1		26	
2		27	
3		28	
4		29	
5		30	
6		31	
7		32	
8		33	
9		34	
10		35	
11		36	
12		37	
13		38	
14		39	
15		40	
16		41	
17		42	
18		43	
19		44	
20		45	
21		46	
22		47	
23		48	
24		49	
25		50	

000106

INCOPRO

No.			No.		
51	[redacted]		76	[redacted]	
52			77		
53			78		
54			79		
55			80		
56			81		
57			82		
58			83		
59			84		
60			85		
61			86		
62			87		
63			88		
64			89		
65			90		
66			91		
67			92		
68			93		
69			94		
70			95		
71			96		
72			97		
73			98		
74			99		
75			100		

000107

INCOPRO