AAEE2016 CONFERENCE
Coffs Harbour, Australia

Statistical Analysis of Global Online Watch Data

Ronald J Hugo and Alexandra Meikleham
Schulich School of Engineering, University of Calgary, Canada
Corresponding Author Email: hugo@ucalgary.ca

CONTEXT

Online video lectures have been available to students through platforms like YouTube for approximately 10 years. Content varies from full-length lectures that have been recorded during live lectures, to PowerPoint presentations with voiceover, to shorter segments recorded using a tablet and screen recording software. The statistical analysis of global online watch data of publicly-available engineering undergraduate content placed on YouTube offers an ability to examine how students from around the world use this resource. The variability in viewing behaviour provides information about both the students and the country from which the content is being viewed. This offers the opportunity to estimate and consider how this type of learning will grow with time in numerous countries around the world.

PURPOSE

To develop an understanding of how engineering students worldwide use publicly-available online video lectures and to draw inferences about how and why content viewing varies by both country and time.

APPROACH

From September 2013 to December 2015, three Mechanical Engineering undergraduate courses were developed for online delivery and uploaded to YouTube. The content included segments that vary in duration from 5 to 15 minutes, and were recorded using a tablet computer with screen recording software. This content has been used for blended / flipped delivery by the first author in his assigned teaching, but it has also been made available for public viewing at no cost. YouTube Analytics has been used to extract global online watch data including the number of views and number of watch minutes. Statistical analysis has been performed on this data in order to provide insight into how online content is used by students.

RESULTS

Data analysis reveals annual cycles in watch data that correlate to both academic schedules and weekly cycles. Rank-ordered plots of watch data reveal a decay rate of -1.32 by country and -0.60 by video. A change in decay rate slope may be attributed to the coherence or completeness of the total data set being examined. In the case of country data, this is most likely due to English proficiency and the percentage of internet adoption. In the case of video data, this is most likely due to the multi-year staggered nature by which courses have been posted. A scaling law is proposed that relates annual watch minutes per video segment per country to inverse engineering enrolment raised to the power minus 4/3. The magnitude of this relationship is approximately dependent on English proficiency index and the percentage of internet adoption, although exceptions to this trend are noted for certain countries.

CONCLUSIONS

Global online watch data reveals both annual and weekly study habits of students. Rank-ordered power-law plots reveal a change in slope that corresponds to the coherence or completeness of the total data set. Average internet bandwidth, provided that it is above 500 kbps, was found to have no noticeable influence on viewing habits. A scaling relationship is proposed that is able to collapse watch minute data in relation to engineering undergraduate enrollment, English proficiency, and household internet adoption. It is anticipated that watch data in certain countries will increase as the percentage of households with internet increases or as restrictions on YouTube are relaxed.

KEYWORDS

Online delivery, internet bandwidth, threshold content viewing.

Introduction

Over the past five years, a great deal of attention has been placed on the potential promise and evolution of online courses offered by universities throughout the world. Content delivery has either been in a blended format where students watch content online and have face-to-face meetings with instructors or in a completely online format where students never actually physically meet either the instructor or other students. A variety of motivating factors has driven the development of online learning, be it the high costs associated with operating a research-intensive university (Bowen, 2013), the ability to provide increased access and reduced time to degree for students in more economically disadvantaged areas (Lewin, 2013i), or by delivering programs in new and creative ways (Crow and Dabars, 2015; Selingo, 2013; DeMillo, 2015).

Of the programs that have been developed, a number of innovative structures exist. These include an online Master's degree offered by the Georgia Institute of Technology in collaboration with AT&T and Udacity (Lewin, 2013ii). In this program, Georgia Tech develops and delivers the content using Udacity as a hosting platform, and AT&T provides funding to help offset program costs. In return, AT&T is able to have input into the program's content and also have access to student transcripts. This enables AT&T to develop talent globally, have the ability to interview those that perform well, and hire in regions to meet their global business needs.

A second innovative program has been developed by the Minerva Schools at K.G.I. (Kamenetz, 2013). This program involves a flipped delivery model with students spending their four-year program living in apartments in San Francisco, Buenos Aires, London, Berlin, Bangalore, and Seoul. Experiential learning activities derived from the residential city augments the flipped delivery model while providing students with global awareness. In addition, without the need to maintain the traditional bricks-and-mortar infrastructure of a university, program costs are kept relatively low.

In addition to these innovative models are the more traditional massive open online courses (MOOC's) offered by organizations that include Coursera, edX, and Udacity (Selingo, 2014). All three have teamed with universities in delivering content, with the offerings provided by Udacity involving company-university partnerships that include Starbucks / Arizona State University (Perez-Pena, 2014) and AT&T / Georgia Tech (Lewin, 2013ii). Online offerings through Coursera and edX are more widely accessible, and represent a mechanism that combines both open education and global advertising for the university offering the course.

A parallel development has been the growing repository of online content that is freely available through YouTube. The early trendsetter in this movement was the MIT OpenCourseWare project (DeMillo, 2013) involving full-length lectures posted to YouTube. This model was later refined by Sal Khan and the Khan Academy using shorter duration video segments (Khan, 2012). Students tend to prefer shorter video segments to full-length lectures given that video segments are searchable, short to review, free to access without the need for login or program enrollment, and readily available on nearly any platform (desktop, laptop or mobile) or web browser. As such, YouTube content serves as a supplemental study aid to traditional live lecture delivery and it can assist students by providing them with an alternate presentation style on a given topic. Given the ubiquitous nature by which this content is available, it also serves as an ideal data source with which to examine the adoption and use of online content globally.

Background

Beginning in July 2013, a junior-level Mechanical Engineering course was made publicly available through YouTube that consisted of 144 video segments. In July 2015, a sophomore-level course consisting of 122 video segments was added, and in September 2015 a second junior-level course consisting of 132 video segments was uploaded. The content has been used to deliver courses in a blended delivery format by the first author at the University of Calgary, but it has also been made openly available for others to view through standard YouTube search procedures.

Methods

This section discusses the instruments and measures that were used to collect the data analyzed in this paper.

Instruments and Measures

1) YouTube Analytics Data

With online content delivered through YouTube, the Analytics package within YouTube was used to examine and compare viewing statistics. Data from July 2013 (when the first video lecture was uploaded) to August 2016 (when the data was extracted) was used.

2) Internet Connectivity

Average connection speed by country was obtained from Akamai (2016). The data for Quarter 1 2016 was retrieved. It should be noted that average internet connection speed increased over the time period that the YouTube analytics data was analyzed. As an example, average internet connection speed for Australia increased from 5,535 kbps in Quarter 3 2013 to 8,785 kbps in Quarter 1 2016. During this same time period, average internet connection speed for the United States increased from 9,524 kbps to 15,333 kpbs.

3) Number of Enrolled Undergraduates

The number of enrolled undergraduates by country was determined using data produced by the World Economic Forum and obtained from WEF (2015). Country profile data and the number of students by field of study were accessed. The number of students currently enrolled in tertiary education in Engineering, Manufacturing and Construction was extracted.

4) English Proficiency Index

The ability to communicate in English was quantified for each country using the English Proficiency Index as obtained from EF-EPI (2015).

5) Internet Adoption

Internet adoption quantifies the percentage of the population, as of July 1 2015, that is considered to be an internet user. The term internet user quantifies the percentage of the population who can access the internet at home, via any device type and connection. This data was obtained from Internet Live Stats (2015). As for Internet Connectivity described above, internet adoption also increased over the time period that the YouTube analytics data was analyzed.

Results and Discussion

Annual Watch Data by Country

Figure 1 shows ensemble averaged and normalized (to a peak value of one) annual watch data for Australia (left) and the United States (right). The data for Australia indicates a steady increase in watch minutes throughout the Autumn Session with a spike in content viewing during the June final exam period. The watch data for the Spring Session shows a similar distinct final exam period from the end of October into November; however, the watch data throughout the Spring Session remains relatively flat. The plot to the right in Figure 1 shows normalized US watch data and again two distinct semesters can be discerned; however, the semester timing is different given the differences between the Northern and Southern Hemispheres. In the US, the fall semester begins in early September and ends with final exams in early-to-mid December. Periods of reduced content viewing are evident during both semesters, specifically Halloween and the Thanksgiving long weekend during the Fall Semester and Spring Break during the Winter Semester. Spring Break, which takes place in the February to March timeframe, varies from institution to institution and thus is less pronounced than the US Thanksgiving long weekend which is observed simultaneously by all US institutions.

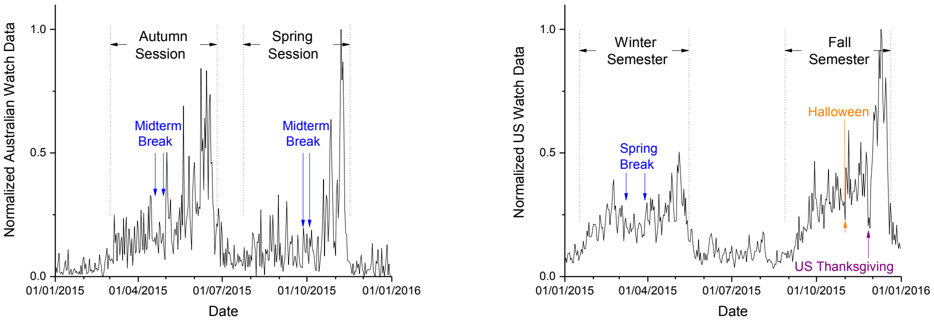

Figure 1: Annual Watch Data – Australia (left) and United States (right) (Normalized to a Maximum of 1.0)

Weekly habits of US students can be better examined by performing a conditional sampling experiment with the three years of annual watch data. Through this sampling approach, data is processed in a manner that averages data sets after they have been aligned by day of the week instead of by calendar date (1 January to 31 December). The result of this is shown in Figure 2 where red drop lines indicate the Friday of each week and it is observed that the lowest watch data typically occurs on a Friday. This provides support to the long-held suspicion that US students tend to do less homework on Fridays. The highest watch data, on the other hand, is not consistent but rather varies from week to week, covering most days other than Friday.

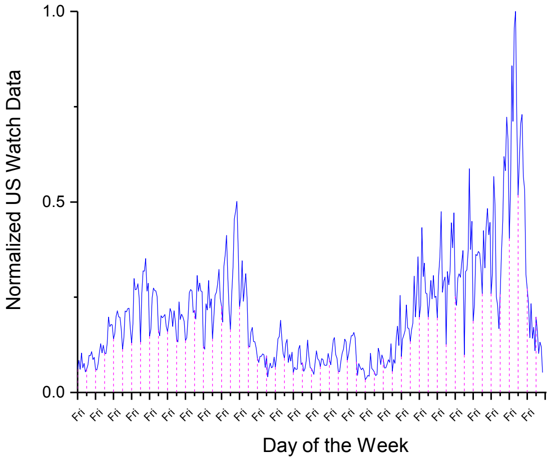

Figure 2: Conditionally Sampled US Watch Data – Red Drop Lines Denote Friday

Rank-Ordered Power-Law Relationship

Figure 3 and Figure 4 show rank-ordered plots of watch minutes by Country and Video segment, respectively. Rank-ordered plots have been used extensively for the analysis of a wide range of data sets, including income distribution, word frequencies in texts, city sizes, gross-domestic product, and internet traffic (Adamic & Huberman, 2002). Despite the organized manner by which the data collapses, open debate remains within the internet research community as to the degree of usefulness that this type of data representation provides (Clegg, Di Cairano-Gilfedder, & Zhou, 2010).

The log-log power-law plots reveal that watch minutes decrease at a more rapid rate by country, with a slope of -1.32 over country 1 to 60 (Pearson’s r = -0.99583 shown in Figure 3), than they do by video, with a slope of -0.60 over video 1 to 100 (Pearson’s r = -0.99435 shown in Figure 4). The coloured vertical lines in each of the power-law plots shown in Figure 3 and Figure 4 denote time series that are shown in Figure 5. By examining the time series, it becomes evident that data in the linear region (power-law region) has periods with significant DC offset, whereas data from the decay region has random spikes with mostly zero DC offset. Cristelli, Batty, and Pietronero (2012) proposed an explanation for a change in slope in the power-law relationship, attributing it to the coherence or completeness of the total data set being examined. This explanation may assist in interpreting the Video data in Figure 4 where data from three distinct courses posted over a two-year time period have been aggregated into one single data set. In applying the concept of coherence to the interpretation of the Country data in Figure 3, it is possible that country-specific factors have influenced watch data for each country, resulting in the observed change in slope. A number of potential country-specific factors will be considered later in this paper.

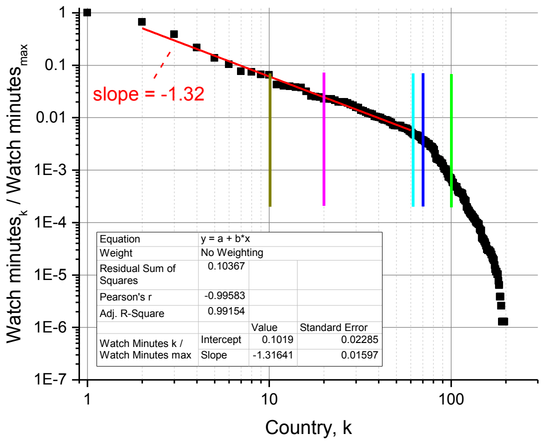

Figure 3: Rank-Ordered Power Law Relationship by Country

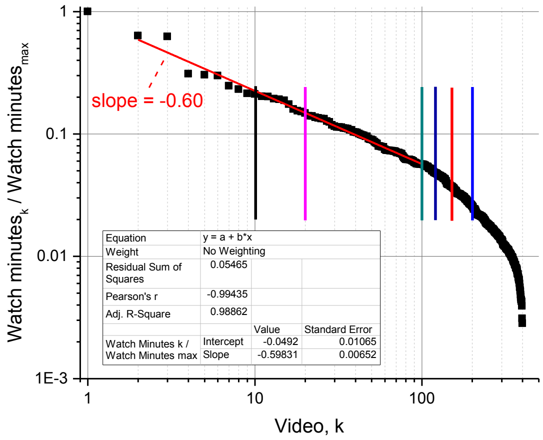

Figure 4: Rank-Ordered Power Law Relationship by Video

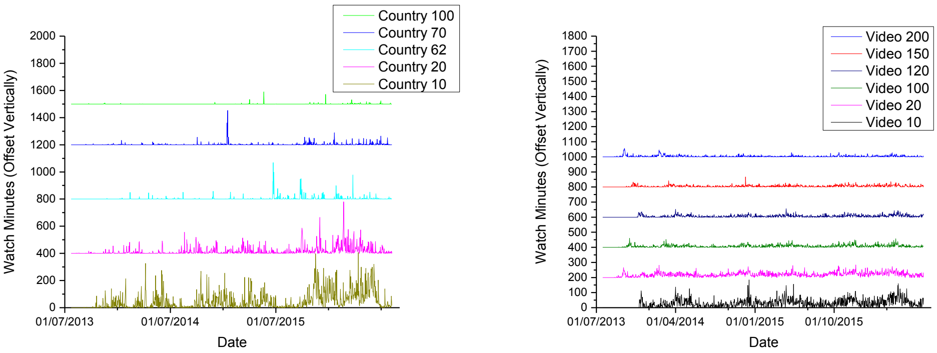

Figure 5: Country Watch Data (left) and Video Watch Data (right)

Internet Bandwidth and English Proficiency

Figure 6 shows viewer retention as a function of the number of watch minutes for each country. This plot demonstrates that, as the number of watch minutes increases, the retention rate statistically converges to 33%. Exceptions to this pattern of convergence appear to exist for India, the USA, and Canada. In the case of India, the lower internet bandwidth of 3,465 kbps may result in a reduction in viewer retention, while for the USA and Canada retention rates in excess of 33% are evident. The case of Canada can be explained as Canada is the home country for the authors of this paper and the YouTube content has been used for the delivery of four individual course offerings at the University of Calgary. This has artificially driven the Canadian retention rate data higher. The higher retention rate in the USA is more difficult to explain but may be a result of certain commonalities between the USA and Canada. These include similar accents between Canadians and Americans and common textbooks used in both countries. Another factor to consider is that students in the USA have become comfortable learning from YouTube given the success of the Khan Academy (Khan, 2012) and other forms of online content delivery.

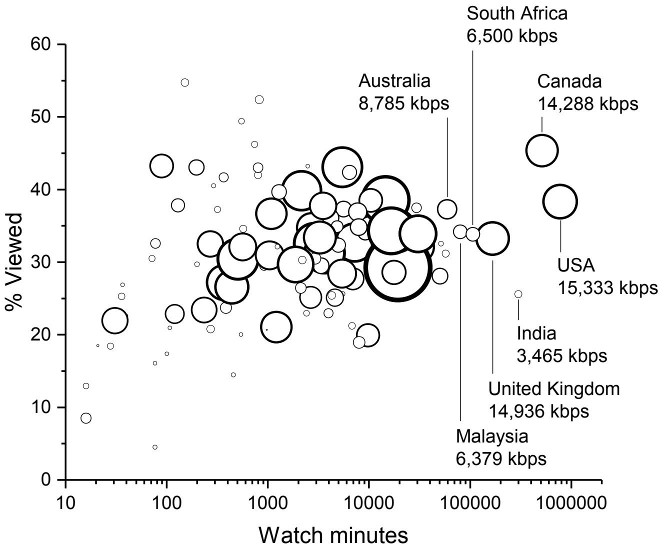

Figure 6: Viewer Retention with Watch Minutes and Internet Bandwidth (Bubble Size)

Figure 7 examines watch minute data as a function of total undergraduate enrolment in Engineering, Manufacturing, and Construction as extracted from WEF (2015). The plot to the left examines the relationship between watch minutes and undergraduate enrollment with the size of the bubbles corresponding to the average internet bandwidth (Akamai, 2016), while the plot on the right examines the same relationship but now with the bubble size corresponding to the English Proficiency Index (EF-EPI, 2015). The data sets reveal that, in general, the larger the engineering undergraduate enrollment in a country, the larger the total number of watch minutes in that country. The internet bandwidth data plotted on the left does not reveal any clear relationship between internet bandwidth and content viewing. This may be due to the fact that YouTube system requirements specify internet bandwidth to be 500 kbps or higher (YouTube, 2016), and all of the countries in the data set had average internet bandwidths that exceeded this minimum specification. The English proficiency data plotted on the right, however, does indicate that the higher the English Proficiency Index (the larger the bubble), the larger the number of watch minutes.

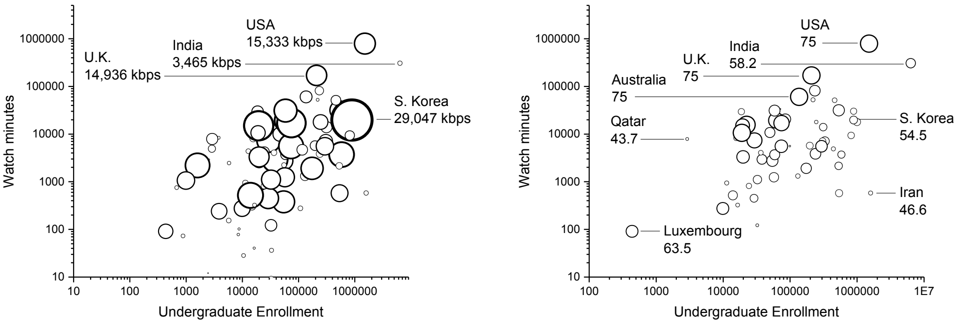

Figure 7: Watch Minutes with Undergraduate Enrolment
Average Internet Bandwidth (left) and English Proficiency Index (right)

Scaled Power-Law Relationship

The data plotted in Figure 7 does indicate a relationship between watch minutes and the total number of students enrolled in engineering programs within a given country. The use of watch minutes is, however, somewhat arbitrary given that the magnitude of watch minutes depends on the time period over which the data is collected. In an effort to report data in a more representative manner, the total number of watch minutes was divided by the number of years during which the data was collected over and the number of video segments upon which the data is based. Furthermore, taking inspiration from the study of turbulence in the field of experimental fluid mechanics, the horizontal axis has been converted into the equivalent of wavenumber by taking the average enrollment for the entire data set (350,000 students) and dividing by the enrollment for each country. The result of this is shown in Figure 8. With this representation, a larger country like India appears to the left in the plot, while a smaller country like Luxembourg appears to the right.

The plot to the left in Figure 8 shows the English Proficiency Index, while the plot to the right shows the percentage of internet users by country. The percentage of internet users reports the percentage of households in a country that have access to the internet, independent of device. Percentage of internet users is more meaningful than average internet bandwidth as it relates to conditions for the total population of a country, not only the conditions for those who have access to the internet. As an example, a country with one high-speed internet connection at 30,000 kbps would have a high average internet bandwidth, although the percentage of internet users would be close to zero.

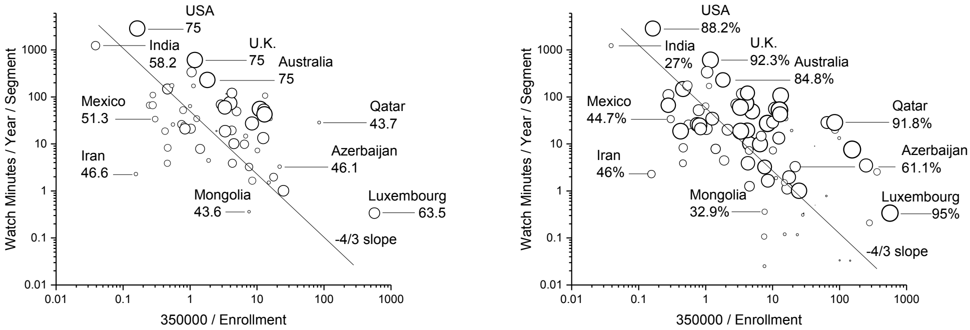

Figure 8: Annual Watch Minutes per Segment with Inverse Engineering Enrolment English Proficiency Index (left) and Percentage of Internet Users (right)

Examining the results of Figure 8, it was hypothesized that the number of watch minutes in a particular country would be most strongly dependent on the English Proficiency level of that country, and the percentage of households with access to the internet. Given this, the English Proficiency Index was multiplied by the percentage of households with access to the internet and this product was scaled to the size of the bubbles. The result of this process is shown in Figure 9.

Figure 9 shows that, in general, the largest bubbles tend to be above and to the right of the minus four-thirds constant-slope line. This results as the amount of content viewing is positively impacted by both the number of internet users in a population and the English proficiency index of that population. There are a number of exceptions to this trend including the United Arab Emirates and Qatar. Both countries have a high percentage of internet users, with U.A.E. being 91.4% and Qatar 91.8%. Given the high percentage of internet users, it is the English Proficiency Index that causes the bubble size to be small. It is speculated that the educational systems in both countries are such that those who attend tertiary-level institutions have a high English Proficiency, one that is well above that of the general public. Consequently, although the bubble size may be representative of the entire country, it is not necessarily representative of those studying Engineering in a tertiary-level institution.

With time, the number of internet users in each country will increase as part of national initiatives. India is the strongest example of this as the number of internet users in that country increased by over 100 million people between 2014 and 2015, and yet there are still 864.677 million people listed as non-users (Internet Live Stats, 2015), indicating strong growth potential. China is another country with significant growth potential, with 660.889 million non-users. However, government regulations in China restrict access to certain websites, including YouTube. As with China, Iran has also blocked YouTube content which explains the low watch data despite its high student enrolment in Engineering programs.

The second data set examined in Figure 9 is English proficiency. Although the English proficiency of a country may not increase rapidly over time, low English proficiency can be addressed through simple solutions such as the addition of video subtitles. Finally, although Figure 9 does not show as clean and collapsed of a data set as the rank-ordered plots in Figure 4, the data presented is reflective of actual conditions in each country. Hence Figure 9 provides insight into how watch minutes will change with time. It is more difficult to extract this type of information from the rank-ordered plots in Figure 4.

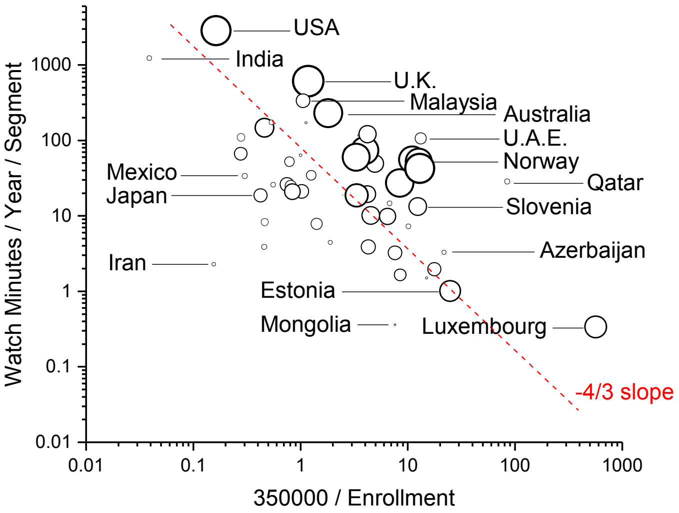

Figure 9: Annual Watch Minutes per Segment with Inverse Engineering Enrolment Bubble Size - Internet Users multiplied by English Proficiency Index

Conclusions

This paper examines the statistical nature of global watch data as derived from publicly-available lectures posted to YouTube. Ensemble averaged annual watch data reveals waveforms that relate to the annual cycles of academic schedules. Rank-ordered power-law plots reveal a change in slope that corresponds to the coherence or completeness of the total data set. Average internet bandwidth, provided that it is above 500 kbps, was found to have no noticeable influence on viewing habits. English proficiency, the percentage of a population with home access to the internet, and the number of students enrolled in Engineering programs was found to correlate with the total number of minutes of content viewed per year. A scaling relationship is proposed that is able to collapse watch-minute data in a manner that is reflective of the situation within each specific country.

The results indicate that the greatest growth in online viewing will come from India. This ultimately will be tied to the success of Digital India, an ambitious government initiative under Prime Minister Modi that seeks to provide internet access to all of India by 2019. Given that only 27% of the population had internet access in 2015, the magnitude of change underway in India is unprecedented. Thus, growth potential in India will be driven by increased household access to the internet, and by the extremely large number of students enrolled in Engineering undergraduate programs.

References

Adamic, L.A., Huberman, B.A. (2002). Zipf's law and the Internet. *Glottometrics*, *3*, 143-150.

Akamai (2016). Internet connection speeds and adoption rates by geography. Retrieved August 5 2016, from https://www.akamai.com/stateoftheinternet

Bowen, W.G. (2013). *Higher Education in the Digital Age*, Princeton University Press.

Clegg, R.G., Di Cairano-Gilfedder, C., Zhou, S. (2010). A critical look at power law modelling of the Internet. *Computer Communications*, *33: 3*, 259-268.

Cristelli, M., Batty, M., Pietronero, L. (2012). There is More than a Power Law in Zipf. *Scientific Reports*, *2:812*, 1-7.

Crow, M.M, Dabars, W.B. (2015). *Designing the New American University*, Johns Hopkins University Press.

DeMillo, R.A. (2013). *Abelard to Apple: The Fate of American Colleges and Universities*, The MIT Press.

DeMillo, R.A. (2015). *Revolution in Higher Education: How a Small Band of Innovators Will Make College Accessible and Affordable*, MIT Press.

EF-EPI (2015). EF English proficiency index. Retrieved August 5 2016, from http://www.ef.com/ca/epi/downloads/

Internet Live Stats (2015). Internet users by country (2015). Retrieved August 13 2016, from http://www.internetlivestats.com/internet-users-by-country/2015/

Kamenetz, A. (2013). Harvard-size ambitions. *The New York Times*, November 1, 2013.

Khan, S. (2012). *The One World Schoolhouse: Education Reimagined*. Twelve.

Lewin, T. (2013i). Colleges adapt online courses to ease burden. *The New York Times*, April 29, 2013.

Lewin, T. (2013ii). Master's degree is new frontier of study online. *The New York Times*, August 17, 2013.

Perez-Pena, R. (2014). Starbucks to provide free college education to thousands of workers. *The New York Times*, June 15, 2014.

Selingo, J.J. (2013). *College Unbound: The Future of Higher Education and What It Means for Students*, Amazon Publishing.

Selingo, J. (2014). Demystifying the MOOC. *The New York Times*, October 29, 2014.

WEF (2015). The human capital report: employment, skills and human capital global challenge insight report. Retrieved August 5 2016, from http://reports.weforum.org/human-capital-report-2015/the-human-capital-index/

YouTube (2016), System Requirements. Retrieved 10 August 2016. https://support.google.com/youtube/answer/78358?hl=en