Better Options Than Harvard: Customized School Rankings and Education Requirements Based on Social Media Data Analysis

Pauline C. Ng

pauline.c.ng<at symbol>gmail.com

Abstract

College ranking lists, such as the one published annually by U.S. News & World Report (USNWR), provide guidance on which colleges are "better". However, the rankings are weighted heavily by subjective opinions and it is unknown whether these rankings have a correlation with students' future successes. To address these issues, we compute college rankings based on notable individuals in the crowd-sourced Wikipedia. We extracted college information from 42% (42,385/100,791) of recent American biographies in Wikipedia by building a decision tree with 77% accuracy. We find that Wikipedia rankings correlate with USNWR rankings (Spearman's rank correlation 0.70), thus validating that USNWR rankings are useful in predicting which schools are likely to produce successful individuals.

Due to the large and diverse dataset of individuals in Wikipedia, we are able to create tailored college rankings for different careers. We also assess whether higher education is a requirement for success for every career path. We find that the need for education in business and entertainment has decreased over time ($p=0.003$ and $p<10^{-16}$ respectively), suggesting that other alternative paths for success in these professions are now available. In conclusion, Wikipedia validates popular college ranking lists and can provide information on higher education for specific professions.

Introduction

Deciding which college to go to is a major decision for many high school students. College ranking lists attempt to provide guidance on which colleges are "better". These rankings are fairly powerful: the U.S. News & World Report's college ranking issue is one of its highest selling issues (Ehrenberg 2005) and two thirds of parents surveyed find college rankings helpful (Monks 1999). However, the validity of the rankings has come into question because schools have fabricated data to improve their rankings (Perez-Pena & Slotnik 2012, Anderson 2013). Furthermore, college rankings are partly based on school reputation, which is subjective as it is based on high school counselors' and university staff's opinions. Indeed, it may be circular – counselors and university staff may look at rankings to determine the reputation of other schools, thus reinforcing the rankings.

College rankings take into account factors such as a school's reputation, student/faculty ratio, and graduation rate. These factors can have a tenuous correlation to the common reason why students go to college: to increase future income prospects and to have a better chance at being 'successful' in life. How can one measure success and rank colleges to provide a more relevant ranking to students' primary goals? When Forbes magazine created its college rankings list, they incorporated how likely a college graduate would appear in Who's Who, a compilation of notable individuals. However, they were criticized because individuals can purportedly buy their way into Who's Who (Carlson,1999; Hamilton, 2005). Who's Who, compiled by research staff, uses arbitrary criteria on whether someone is worthy of inclusion (Hamilton, 2005). In response, Forbes decreased the weight of this measure in their rankings.

This study asks if, instead of relying upon limited curated lists compiled by financially motivated staff, whether crowd-sourced social media can improve or validate the existing ranking systems. Specifically, Wikipedia can be used as a resource to mine characteristics of notable individuals (Ng 2012). Wikipedia contains a diverse set of notable individuals from web sensations to college coaches, with notability defined as citation in independent public media (e.g. newspaper, magazine articles). The person does not necessarily have to be wealthy, but the assumption is that notable individuals

have been 'successful' in accomplishing something that the public would find noteworthy.

Another question is whether a college education is needed for success. This question arose from observational studies which show that 18% of U.S. billionaires and 7% of CEOs at the 500 largest corporations do not have college degrees (Herper 2000, Wecker 2012, Lenzner, 2012). However, these analyses are based on a small number of successful people. A larger dataset such as one offered by Wikipedia provides the opportunity to quantitatively examine this question with statistical robustness.

In this work, we create college rankings based on notable individuals in Wikipedia. We find that Wikipedia rankings correlate with the U.S. News World & Report's and Forbes' rankings, thus validating that these rankings do predict which schools are likely to produce successful individuals. Furthermore, Wikipedia's large and diverse dataset enables us to answer specific questions about college education. In this study, we analyze the impact of a college education on specific career paths, but the technique can be extended to additional analyses. The results of the career-specific college ranking analysis show that the typical prestigious schools are not as important for some professions. We analyze education levels for different professions and how these have changed over time. Enhanced analysis reveals trends, such as the need for education in business and entertainment decreasing over time. These results suggest that alternative paths for success in these professions are available.

Methods

Data Collection

In order to analyze notable Americans, we downloaded Wikipedia pages of Americans born between 1930-1984 on January 7, 2011. We ignored people born 1985 and later as they may be too young to obtain a college education. This yielded 100,791 biographies.

Text Extraction of School, Degree, and Major

For each person's entry, we want to identify which sentences contain information about education, and then parse from that sentence the school, degree, and major. We added the various degrees (e.g. B.S., B.A., M.B.A.) to the NLTK Punkt English tokenizer so that it could properly parse sentences with degrees. We then implemented a decision tree to identify sentences pertinent to education.

To build the decision tree, we retrieved sentences containing the word 'school', 'college', 'university' and any of the top 25 schools listed in the U.S. News & World Report. We manually classified these sentences to obtain 465 positive and 465 negative examples. A decision tree was trained on 75% of the dataset, and tested on the remaining 25%. The decision tree had an accuracy of 78%, precision of 82%, and recall of 72%. The decision tree was reimplemented with modifications – specific names due to overtraining were removed, and more degrees and university names were added. For example, one branch of the decision tree returns a positive result if a sentence contains a capital "B" (presumably because the sentence contains "Bachelor's degree", "B.A.", or "B.S."). We edited this branch of the decision tree to classify a sentence positively as an education sentence if the sentence contained any degrees (e.g. "Bachelor", "B.S.", "PhD", etc.) The manually edited decision tree achieved 77% accuracy, 81% precision, and 70% recall.

Once a sentence in a Wikipedia biography was classified as relevant to education, we identified schools and degrees in this sentence by looking for string matches using Simstring (Okazaki and Tsujii, 2010). In order to capture colleges containing prepositions (e.g. "University of California *at* Berkeley" and "University of California *in* Berkeley"), if there was no exact match to known universities, we took the uppercase words in a sentence and looked for university matches to these strings.

As an independent validation to the manually edited decision tree, we compared schools extracted from the decision tree with the 6,518 Wikipedia biographies that had education information in their corresponding Wikipedia Infobox. When compared to college information obtained from Wikipedia Infobox for each biography, the decision tree achieved 81% recall, and 75% precision. For completeness, we combined education information from the Infobox with the education information that was parsed from text of a Wikipedia entry. After combining, 42,385 of the Wikipedia biographies have education information. This is a substantial increase compared to obtaining education information from the Wikipedia Infobox alone, which provides education data for only 6,518 biographies. Therefore, additional text parsing with our decision tree dramatically increases sensitivity.

Some education sentences had the structure "majored in <subject>" or "graduated with a degree in <subject>." To parse out the subject that a person majored in, we took the education sentence and extracted out words that followed "in". The most frequent words and bigrams were used to construct a word cloud.

Calculating Scores and Ranks for Colleges

After identifying the schools attended by persons in Wikipedia, we calculate $Wiki\_Freq_{college}$, the fraction of Wikipedia notables from *college*:

$$Wiki\_Freq_{college} = \frac{Num\_Of\_Wiki\_Bios_{college}}{\sum_{college} Num\_Of\_Wiki\_Bios_{college}}$$

where $Num\_of\_Wiki\_Bios_{college}$ is the number of Wikipedia individuals who have attended college *college*.

The number of Wikipedia notables from a certain

college has to be adjusted for the college's enrollment size because schools with larger class sizes will have a larger pool of students who can be incorporated into Wikipedia. $Num\_Students\_Enrolled_{college}$ are college enrollments obtained from collegedata.com (2012) and includes undergraduates and postgraduates.

$$Student\_Fraction_{college} = \frac{Num\_Students\_Enrolled_{college}}{\sum_{college} Num\_Students\_Enrolled_{college}}$$

where the denominator is summed over all colleges with at least one Wikipedia notable.

Then the relative ratio is:

$$RR_{college} = \frac{Wiki\_Freq_{college}}{Student\_Fraction_{college}}$$

If $RR_{college}$ equals 1 then the frequency of college alumni in Wikipedia matches the expected frequency based on enrollment. A high $RR_{college}$ (>1) means that a disproportionately higher number of individuals from that school appears in Wikipedia than expected. Universities with at least 10 notable alumni were ranked, where universities with higher $RR_{college}$ received higher ranks.

Categories

Wikipedia has a diverse set of notable individuals. Because different professions may have different education trends, we subdivided the Wikipedia notables into five professions: entertainment, sports, academics, business, and government. To classify individuals into these professions, we extracted the "Categories" section from all Wikipedia biographies, identified the most common terms, and then manually classified the most frequent categories. For example, individuals that belong to a category containing the string "football" were classified as sports. Other examples are: "senator" classified as government, "entrepreneur" as business, "singer" or "writer" as entertainment, and "professor" or "scientist" as academics. Over 100 Wikipedia categories were reclassified into these 5 professions and the groupings can be found at **<website withheld>**. After classifying all Wikipedia individuals, our dataset is composed of 45,949 entertainment biographies, 35,602 sports biographies, 12,059 academics biographies, 5,253 business biographies, and 12,272 government biographies.

Results

Validation Against Established College Rankings

We compared the rankings obtained from Wikipedia with two other well-known ranking lists: U.S. News & World Report and Forbes. We obtained the top 200 ranked colleges on the U.S. College News & World Report website (2012) and compare it with the relative ratios derived from Wikipedia (Figure 1). The Spearman's rank correlation between the U.S. College News & World Report and the Wikipedia RR's is 0.70. Thus our scoring methodology correlates well with U.S. News & World Report rankings. The top 50 schools in U.S. News & World Report show especially high enrichment in Wikipedia: 91% have RR > 1 and these have an average RR of 7.4. This suggests that if one attends a school ranked in the top 50, the student is 7.4x more likely to appear in Wikipedia than if he/she attends another college. For the remaining schools on the list with ranks 51-200, only 43% have a RR > 1, and their mean enrichment in Wikipedia is 1.9.

We also compared our Wikipedia college rankings against Forbes college rankings. One of Forbes' criteria is how likely someone who graduates from a college will appear in Who's Who, and therefore rankings based on this criteria should be similar to our Wikipedia rankings. In order to obtain rankings derived solely from Who's Who and not other factors, we used the Forbes website which allows users to create rankings based on their own criteria (www.forbes.com/2009/08/05/best-colleges-ranking-screener-opinions). To obtain rankings based on Who's Who only, we set "graduates achieve success" (the Who's Who criteria) as "very important" and set all other criteria as "not important". Twenty colleges are returned, and we compared ranks for these twenty colleges on the Forbes list with the corresponding Wikipedia ranks. The Spearman's rank correlation coefficient is 0.69.

The high correlations of college ranks between Wikipedia with both Forbes and U.S. News & World Report validate the Wikipedia methodology.

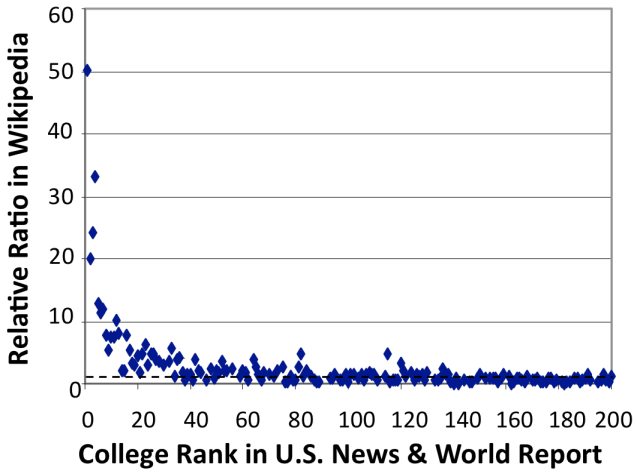

Figure 1. Comparison of a college's rank from U.S. News & World Report (x-axis) with its enrichment of alumni in Wikipedia (y-axis).

Wikipedia Rankings

In the previous section, we showed that Wikipedia

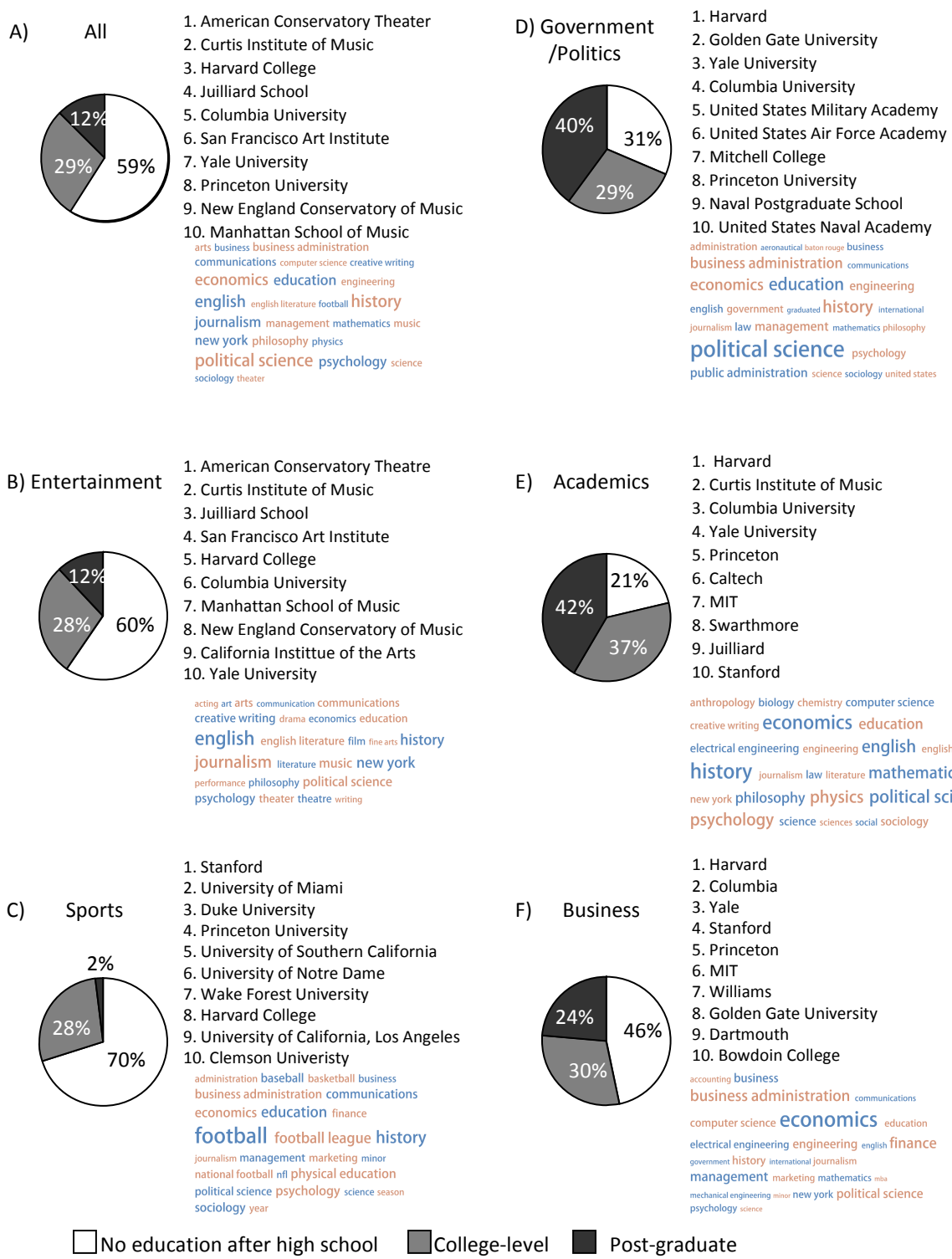

Figure 2. Top ten universities and institutes according to profession. The pie graph shows the level of education for a given profession; shaded areas represent proportion of Wikipedia individuals that have education beyond high school. The tag cloud shows the 25 most frequent majors.

rankings correlate with well-known published college rankings from U.S. News & World Report and Forbes. However, the standard college rankings do not include nontraditional and highly specialized schools. When we examined the entire list of Wikipedia rankings, we see that some specialized schools rank as highly as other renowned Ivy League schools such as Harvard and Princeton (Figure 2A). For example, the American Conservatory Theatre was ranked first among all of the institutes ranked by Wikipedia. This was unexpected, but further investigation shows that Wikipedia is dominated by entertainers (41% of biographies). Therefore, we separated the Wikipedia notables into five professions and created college rankings for each category: Academics (e.g. teachers, engineers), Entertainment (e.g. authors, theatre), Business (e.g. managers, billionaires), Government (e.g. military, senators), and Athletics (e.g. Olympians, football players). After subcategorization, only academic rankings are affected by the dominance of entertainers (Figure 2E): Curtis Institute of Music and Juillard are ranked highly because these institutes' students later become faculty, which is classified as academics.

Different career categories have varying levels of education and school rankings differ for each job category (Figure 2). Not surprisingly, a large proportion of academics in Wikipedia are educated (79%), and many with degrees beyond college level (42%). The second most educated category is government (69%), and then business (54%). A smaller proportion of artists and athletes are educated according to Wikipedia (40% and 30%, respectively). This suggests that successful entertainers and athletes have a lower need for education compared to academics, business, and government officials. Thus, different levels of education are needed for different careers. For comparison, 30% of the general population have Bachelor's degrees or higher (United States Census Bureau, 2010). With the exception of athletes, all professions (academics, government/politics, business, and entertainment) have higher levels of education compared to the general population. This indicates that notable individuals tend to have higher levels of education compared to the general population. The education level of athletes is similar to the general population.

Education Over Time

There is some controversy on whether a college education is needed for success based on the observation that Bill Gates, Steve Jobs, and Mark Zuckenberg were college dropouts and 20% of billionaires do not have college degrees (Williams, 2012, Herper 2000). However, these observations are based on a small number of successful people. Simultaneously, the U.S. population has become more educated over time; the college graduation rate has almost doubled over the past 5 decades (Stoops, 2004). As a result, one would expect that the fraction of people in Wikipedia with education would also increase over time, to match the general population. Therefore, we examined education levels over time.

In the previous section, we show that different professions have different levels of education, supporting the concept that certain professions have less emphasis on higher education. We continue to explore the question of the value of education by looking at time trends for education.

We examined education levels among Wikipedia notables over time. We performed a linear regression on the fraction of educated notables versus time for the different professions. If education levels increase over time, then the slope of the line should be positive; if it decreases, the slope will be negative. If educational attainment does not change over time, the slope is equal to zero.

People in entertainment and business show a small but significant decline in education over time (Entertainment $p< 10^{-16}$; Business $p=0.003$) (Figure 3). This suggests that in recent years there is less reliance on education in order to excel in these fields and there may be alternative career paths to become successful. The decline in education for businesspeople starts for those born in the 1960's. This would correspond to being educated in the mid-1980's which coincides with a recession where college tuition may have been impractical. It also coincides with the development of the World Wide Web, which enabled non-traditional career paths to success.

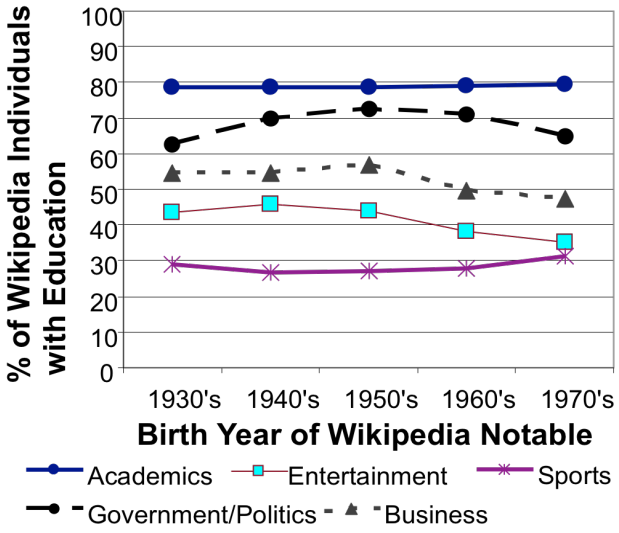

Figure 3. Education rates over time. The x-axis is the birth year of the Wikipedia notable, so people born in the 1960's would be ready to enter college around the 1980's.

Successful athletes show increased education rates in recent years (Figure 3, $p= 3 *10^{-6}$). However, college

attendance may not mean graduation because athletes may become professional before graduation. This result may also be biased by the popularity of college athletics, and the tendency for professional sports leagues to recruit players from college athletics programs. For people who are classified in the categories of Academics and Government, educational levels appear to have stayed constant over the past 40 years (we are unable to reject the null hypothesis that the slope = 0; Academics p=0.455; Government p=0.820).

Conclusions

In this study, we analyze universities and education levels for notable individuals in Wikipedia. The purpose of this study was two-fold. First, college ranking lists are under scrutiny because they use subjective criteria such as reputation and most of the data comes from colleges themselves (Lynch 2012, Anderson 2013). Crowd-sourced datasets such as Wikipedia do not have the bias of self-reporting, and is used to corroborate the rankings. Second, Wikipedia's large dataset enables creation of career-specific rankings and analysis of education patterns over time.

Using Wikipedia's database of notable individuals, we confirm that the popular U.S. News & World Report rankings are useful in predicting if a college produces 'successful' people despite not using success as an explicit criteria. The enrichment for notable individuals is especially high for schools ranked in the top 50 of U.S. News & World Report. Furthermore, our analysis shows that the role of highly specialized schools should not be underestimated.

U.S. News & World Report determines school reputation by surveying high school counselors and university staff. Using a crowd-sourced dataset to imply reputation has the advantage of being less biased because the compilation of notable individuals and the schools they attend is an amalgamation of community effort. However, it is possible that Wikipedia contributors may be more likely to record attendance to prestigious universities than less well known schools. Another weakness is that anyone can create entries in Wikipedia. As there are reports of college staff falsifying data to improve rankings (Anderson 2013, Perez-Pena & Slotnik 2012), one could impose a requirement of a minimum number of citations and contributors in order to mitigate possible inflation of notability by certain universities.

An advantage of using the Wikipedia dataset is we are able to recognize that many notable individuals do not have standard professions. As a result, there is value to creating career-specific rankings to tease out the importance of a college education versus other forms of experience. This leads to the interesting observation that art institutes such as the American Conservatory Theater and Curtis Institute of Music are highly ranked – as high as the most reputable colleges, such as Harvard and Yale. In retrospect, this is not surprising because admission to these specialized institutes is based on competitive auditions, where candidates have already performed or created a portfolio. Furthermore, these artists are likely to be written up in newspapers that can be cited because their profession lends them to appear in mainstream media, such as in movies or in theater.

One of the major weaknesses of this analysis is the difficulty discerning whether someone has graduated from an institute. Many sentences in Wikipedia describing people's educations used ambiguous phrases such as "He attended <university>" or "She studied at <university" and our decision tree designated sentences with these words ("studied" or "attended") as college-educated. However, the text does not explicitly say whether the person graduated, and it is possible the person did not complete their education. For example, many successful athletes are drafted into a professional team before graduation. The problem with dropouts is that this information may be be missing or ambiguous. Wikipedia's entry on Mark Zuckerberg, Facebook founder, states that he attended Harvard but never explicitly says that he dropped out of college (as of Jan 2,2013). This would result in overestimating the fraction of college graduates in Wikipedia notables. Under-reporting can occur if the Wikipedia entry never mentions a college history because contributors to the entry do not feel it is relevant or important to the person's biography. This may be true for entertainers and athletes. One possible way to address this issue is to use degree verification services such as National Student Clearinghouse.

Some have questioned if a college degree is needed for success (Williams, 2012). Past studies have been anecdotal or based on only a few hundred individuals. To the best of our knowledge, this is the first large-scale analytical study exploring this question. While education levels of notable individuals in Wikipedia exceed that of the general population (with athletes being the exception), we find a declining need for education for entertainers and businessmen since the 1980's. This slight but significant downward trend for the education levels of entertainers and businessmen occurs despite the fact that the general population has become more educated. This suggests that alternative paths to notability and success may be available for these professions. For example, singers and actors can now become successful in their teen years through channels such as YouTube and Disney and may not need formal education. Young entrepreneurs can bootstrap their own companies, and increasingly investors do not require a college degree (Williams 2012). In contrast, athletes have increased education rates over the recent years. NFL and NBA require players to be out of high school for three years and one year, respectively. Therefore athletes that wish to become professional go to college so they can

demonstrate their skills and be later recruited. Thus, different social expectations for education and the availability of alternative paths to a profession can affect education levels for different careers.

References

Anderson, N. 2013, February 7. Five colleges misreported data to U.S. News, raising concerns about rankings, reputation. The Washington Post.

Carlson, T. March 8, 1999. The hall of lame. Forbes.

Ehrenberg, R. G. (2005). Method or madness? Inside the U.S. News & World Report College Rankings [Electronic version]. Journal of College Admission, 189, 29-35.

Hamilton, W. L. 2005, November 13. Who are you? Why are you here? New York Times.

Herper, M. 2000, June 29. Some billionaires choose school of hard knocks. Forbes.

Lenzner, R. 2012, October 29. You can be CEO of a top 500 company without a Harvard degree. Forbes.

Monks, J. and Ehrenberg, R. G. (1999). The impact of U.S. News & World Report college rankings on admissions outcomes and pricing policies at selective private institutions (CHERI Working Paper #1). Retrieved Jan. 8, 2013, from Cornell University, ILR School site: http://digitalcommons.ilr.cornell.edu/cheri/1

Ng, P.C. (2012) What Kobe Bryant and Britney Spears have in common: Mining Wikipedia for characteristics of notable individuals. Proceedings of the Sixth International Conference on Weblogs and Social Media.

Okazaki, N. and Tsujii, J. Simple and Efficient Algorithm for Approximate Dictionary Matchin},Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)},August 2010,Beijing, China, pp. 851—859, http://www.aclweb.org/anthology/C10-1096

Perez-Pena, R. & Slotnik D.E. Gaming the College Rankings. January 31, 2012 The New York Times. http://www.nytimes.com/2012/02/01/education/gaming-the-college-rankings.html?pagewanted=all&_r=0

Stoops, N.S. Educational Attainment in the United States: 2003. U.S. Census Bureau. June 2004. Retrieved Jan. 8, 2013, from http://www.census.gov/prod/2004pubs/p20-550.pdf

U.S. Census Bureau, Educational Attainment in the United States: 2010. Table 2, Educational Attainment of the Population 25 Years and Over, by Selected Characteristics: 2010. Retrieved Feb. 10,2013 from http://www.census.gov/hhes/socdemo/education/data/cps/2010/tables.html

Wecker, M. 2012, May 14. Where the Fortune 500 CEOs went to school. U.S. News.

Williams, A. December 2, 2012. Saying No to College. New York Times. Retrieved Jan 8, 2013.