Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data

Márton Mestyán[1], Taha Yasseri[1,2,3,*], János Kertész[1,3,4]

1 Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary

2 Oxford Internet Institute, University of Oxford, Oxford, UK

3 Department of Biomedical Engineering and Computational Science, Aalto University, Aalto, Finland

4 Center for Network Science, Central European University, Budapest, Hungary

∗ **E-mail: yasseri@oii.ox.ac.uk**

Abstract

Use of socially generated "big data" to access information about collective states of the minds in human societies has become a new paradigm in the emerging field of computational social science. A natural application of this would be the prediction of the society's reaction to a new product in the sense of popularity and adoption rate. However, bridging the gap between "real time monitoring" and "early predicting" remains a big challenge. Here we report on an endeavor to build a minimalistic predictive model for the financial success of movies based on collective activity data of online users. We show that the popularity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia, the well-known online encyclopedia.

Introduction

Living in the digital world of today, along with all the advantages also has its side effects and byproducts. Our daily life nowadays leaves a digital trace of all our activities in the recently developed Information and Communications Technology based environments. Our social communications through different digital channels, financial activities within e-commerce, physical locations registered by cell phone providers etc., are traced and recorded. In addition to such passive collection of data about online activity, we also actively share information about our feelings, emotional moods, opinions and views through the so called Web 2.0. or user generated content within social media. In addition to providing us with novel answers to classic questions about individual and social aspects of human life from scientific point of view, precise analysis of this huge amount of data can have practical applications to predict, monitor, and cope with many different type of events, from simple matters of daily life to massive crises in the global scale. For example, Sakaki et al. have developed an alerting system based on Tweets (posts in the Twitter microblogging service), being able to detect earthquakes almost in real time [1]. They elaborate their detection system further to detect rainbows in the sky, and traffic jams in cities [2]. The practical point of their work is that the alerting system could perform so promptly that the alert message could arrive faster than the earthquake waves to certain regions. Bollen et al. have analyzed moods of Tweets and based on their investigations they could predict daily up and down changes in Dow Jones Industrial Average values with an accuracy of 87.6% [3]. Saavedra et al. investigated the relationship between the content of traders' messages and market dynamics. They show that there is a positive correlation between the usage of "bundles" of positive and negative words with agents' overall financial performance [4]. Another example is using Twitter to predict electoral outcomes [5], however with its biases and limitations [6, 7]. Interesting studies have appeared treating the use of social media indicators to predict the scientific impact of research articles, e.g., short-term web usage (number of downloads from the pre-print sharing web site "arXiv") [8] and Twitter mentions [9]. In a recent work, it is shown that

Twitter mentions and arXiv downloads follow two distinct temporal patterns of activity, however, the volume of Twitter mentions is statistically correlated with arXiv downloads and early citations [10]. Preis et al. found a correlation between weekly transaction volumes of "S&P 500 companies" and weekly Google search volumes of corresponding company names [11]. By analyzing search queries for information about preceding and following years, a "striking" correlation between a country's GDP and the predisposition of its inhabitants to look forward is observed [12]. Based on Google search logs, Ginsberg et al. estimated the spread of influenza in the United States [13]. There are other examples of using social media streams to make predictions on news popularity in terms of the number of user-generated comments [14, 15] or the number of news visitors [16]. For a comprehensive literature review see [17].

Statistical analysis of motion picture markets has led to intriguing results, such as observing the evidence for a Pareto law for movie income [18, 19] along with a log-normal distribution of the gross income per theater and a bimodal distribution of the number of theaters in which a movie is shown [20]. By analyzing historical data about 70 years of the American movie market, Sreenivasan has argued that the movies with higher level of novelty (assigned based on keywords from the Internet Movie Database) produce larger revenue [21]. Despite much effort with different approaches, predicting the financial success of a movie remains a challenging open problem. For example, Sharda and Delen have trained a neural network to process pre-release data, such as quality and popularity variables, and classify movies into nine categories according to their anticipated income, from "flop" to "blockbuster". For test samples, the neural network classifies only 36.9% of the movies correctly, while 75.2% of the movies are at most one category away from correct [22]. Joshi et al. have built a multivariate linear regression model that joined meta-data with text features from pre-release critiques to predict the revenue with a coefficient of determination $R^2 = 0.671$ [23]. Since predictions based on classic quality factors fail to reach a level of accuracy high enough for practical application, usage of user-generated data to predict the success of a movie becomes a very tempting approach. Ishii et al. present a mathematical framework for the spread of popularity in society [24]. Their model, which takes the advertisement budget as an input parameter and generates a dynamic popularity variable, is validated against the number of blog posts on the particular movies in the Japanese Blogosphere. In other words they consider the activity level of bloggers as a representative parameter for social popularity. In an earlier work [25] a quantitative model based on "word of mouth" spreading mechanism was introduced in order to assess the quality of movies based on the "aggregated consumption data". However, by analyzing the sentiment of blog stories on movies, Mishne and Glances emphasize that the correlation between pre-release sentiment and sales is not at an adequate level to build up a predictive model [26]. In a very interesting approach Asur and Huberman set up a prediction system for the revenue of movies based on the volume of Twitter mentions [27]. They achieve an adjusted coefficient of determination of 0.97 on the night before the movie release for the first weekend revenue of a sample of 24 movies. In a later work, however, Wong et al. show that Tweets do not necessarily represent the financial success of movies [28]. They consider a sample of 34 movies and compare the Tweets about the movies to evaluations written by users of movie review web sites. They argue that predictions based on social media could have high precision but low recall. Yun and Gloor showed that the betweenness centrality of a movie in a network representation of its presence on the Web is correlated with its financial success [29]. In a rather novel approach, Oghina et al. have made use of Twitter and YouTube activity streams to predict the ratings in the Internet Movie Database (IMDb), which is among the most popular online movie databases [30].

Wikipedia, as a predominant example of user-generated media, has been intensely studied from different points of view. Its size and growth [31–33], topical coverage and notability of entries [34–36], conflict and editorial wars among users [37–41], editorial patterns [42] and linguistic features [43] are only few examples of research topics associated with Wikipedia. We are aware of two comprehensive reviews [44, 45] and a brief hands-on guide to some of the most recent Wikipedia research [46].

Although effects of external events on the activity of Wikipedia editors [47, 48] and the number of page views [49, 50] have been studied in detail, usage of Wikipedia as a source of information to detect

and predict events in real world has been limited to the work by Osborne et al. [51], in which they used Wikipedia page views to fine-filter the outcome of their algorithm for Twitter-based "first story detection" and a very recent work by Georgescu et al., in which Wikipedia edits are introduced as "entity-specific news tickers and time-lines" generators [52]. And finally in an interesting work published later than the first revision of the current manuscript, Moat et al. reported on the predictive power of Wikipedia data for financial fluctuations [53].

In this work we consider both the activity level of editors and the number of page views by readers to assess the popularity of a movie. We define different predictor variables and apply a linear regression model to forecast the first weekend box office revenue of a set of 312 movies, which were released in the United States in 2010. Our analysis not only outperforms the previous works by the much larger number of movies we have investigated, but also improves on the state of the art by providing reasonable predictions as early as one month prior to the release date of the movie. Finally, our statistical approach, free of any language based analysis, e.g., sentiment analysis, can be easily generalized to non-English speaking movie markets or even other kinds of products.

Results

According to data from Box Office Mojo, there were 535 movies that were screened in the United States in 2010 (see the Methods section). We could track the corresponding page in Wikipedia for 312 of them. A closer look at the history of these 312 articles shows that many of them are created a lot earlier than the release date of the movie (Fig. 1(A)). This enables us to follow the popularity of the movie much in advance. To estimate the popularity, we followed four activity measures; V: *Number of views* of the article page, U: *Number of users*, being the number of human editors who have contributed to the article, E: *Number of edits* made by human editors on the article, and R: *Collaborative rigor* (or simply *rigor* [54]) of the editing train of the article. To have a consistent time framework, we set the release time of the movie as $t = 0$. For more details see the Methods section. Examples of the daily increments of number of views and number of users are shown in Supplementary Fig. S1. The daily increments of both variables rise and fall around the day of release similarly to observations by Ishii et al. [24]. In addition to these, an essential parameter for predicting the movie revenue is *the number of theaters* that screen the movie T, which is included in our set of parameters. The complete dataset including the financial data as well as Wikipedia activity records is available via the Supplementary Data S1. To have an overall image of the sample, histograms of the accumulated values of the 4 activity parameters from the first edit on the article up to 7 days after release, along with the first weekend box office revenue, and the number of theaters screening the movie are depicted in Fig. 1(B–F). It is clear that revenues among the sample have a bimodal distribution (Fig. 1(B)). This is in accord with [20], where authors report that the distribution of the total revenue of a sample of 5,222 movies released over the period of 1999-2008 across theaters in the USA, exhibits bimodal nature and have been fit using a superposition of two log-normal distributions. It also shows that Wikipedia coverage is not limited to financially successful movies. The considerable amount of activity on Wikipedia articles (Fig. 1(D–G)) indicates the richness of the data. However, before building a regression model, the correlations between the activity parameters and the box office revenue should be examined first.

The Pearson correlation coefficient $r_j(t)$ between the accumulated value $x_j(t)$ of the j-th predictor variable from the inception of the article up to time t before the movie release and the box office revenue y is calculated as

$$r_j(t) = \frac{\langle x_j(t)y\rangle - \langle x_j(t)\rangle\langle y\rangle}{\sqrt{\langle x_j^2(t)\rangle - \langle x_j(t)\rangle^2}\sqrt{\langle y^2\rangle - \langle y\rangle^2}}, \quad (1)$$

with $\langle . \rangle$ indicating average over the whole sample. Temporal correlations are shown in Fig. 2. For all activity based predictors the correlation coefficient gradually increases as time approaches the day

of release and around the day of release, correlation suddenly rises. Note that V shows the highest correlation with the revenue prior to the release pf movies.

We build a multivariate linear regression model for predicting the box office revenue y. The general form of a regression model at time t before release, based on a set of predictor variables S is

$$y = \sum_{j \in S} \alpha_j(t) x_j(t) + C_S(t) + \varepsilon_S(t), \tag{2}$$

where $\alpha_j(t)$s are time varying parameters of the linear regression model, $C_S(t)$ is a constant and $\varepsilon_S(t)$ is the noise term. We feed the model with different combinations of predictor variables and characterize the goodness of different sets by calculating the coefficient of determination $R^2(t)$. The coefficient of determination is calculated using 10-fold cross-validation (See Methods section). Temporal evolution of $R^2(t)$ is shown for different predictor sets S in Fig. 3. While a model employing $\{T\}$ can be seen as a benchmark of the state of the art in real market predictions, the model solely fed by $\{V\}$ predicts roughly as well as that. Combinations of $\{V, T\}$ and $\{U, T\}$ score well above the benchmark indicating the relevance of activity measures for prediction. Among all sets considered (not shown here), $\{V, U, R, E, T\}$ yields the highest coefficient of determination, which reaches 0.77 around a month before the movie release.

Discussion

Results presented above clearly show how simple use of user generated data in a social environment like Wikipedia can enhance our ability to predict the collective reaction of society to a cultural product. While these results can be of practical application for marketing purpose, especially in combination with other source of information, our main aim is to demonstrate the extent of engagement of members of the public in the peer-production platforms. The introduced approach can be easily generalized to other fields where mining of public opinion provides valuable insights, e.g., financial decisions, policy making, and governance. We believe that Wikipedia and similar mass-collaboration platforms can serve as alternative resources for social media streams with higher level of professionalism and deeper engagement of users. Since the methods presented here are independent of the language of the medium, they can be easily generalized to other languages and local markets.

It is worth mentioning that to feed our predictive model, we have tried several other activity measures, which can potentially be predictive parameters, e.g., time span between the creation of the article and the release time and length of the article. However these quantities did not show any significant correlation with the box office revenue and consequently were excluded from the model.

We also compare the predictive model based on Wikipedia activity measures with the results of the Twitter-based model provided in the 2010 study of Asur and Huberman [27]. Asur and Huberman use a sample of 24 movies to train and test their model. In the same approach we train and test our model focusing on the same set of movies. The $R^2(t)$ of our Wikipedia model reaches 0.94 few days before release, while it is 0.98 for the Twitter model. However, the results of the Twitter study are limited to the night before release, while the analysis presented here can make predictions with reasonable accuracy ($R^2 > 0.925$) as early as one month before release (See Fig. 4). One should also bear in mind that the Wikipedia model does not require any complex content analysis and only relies on statistical measures of activity level. The predicting power of the Wikipedia-based model, despite its simplicity compared to the Twitter, can be explained by the fact that many of the Wikipedia editors are committed followers of movie industry who gather information and edit related articles significantly earlier than the release date, whereas the "mass" production of tweets only occurs very close to the release time, mostly evoked by marketing campaigns.

Fig. 5 shows the actual revenue of movies in the sample against the predicted revenue at $t = -30$ days. It is evident that the prediction is more precise for more successful movies. When less successful movies

are considered, deviations from the diagonal line denoting perfect prediction, increase. Some examples of the movies whose box office receipts were predicted accurately are *Iron Man 2*, *Alice in Wonderland*, *Toy Story 3*, *Inception*, *Clash of the Titans*, and *Shutter Island*. However, the model failed to provide accurate predictions for less successful movies, e.g., *Never Let Me Go*, *Animal Kingdom*, *The Girl on the Train*, *The Killer Inside Me*, and *The Lottery*. This systematic difference in precision can be explained by the amount of data available for each class of movies. Clearly the model works more accurately when the movie is more popular and the volume of the related data is larger. By considering the green squares which represent the movies in the sample predicted by the Twitter model, one realizes that most of the movies predicted by the Twitter method are among the successful ones, therefore applicability of the Twitter model on movies with medium and low popularity levels remains an open question.

While we tried to keep our model as simple as possible and based on only a few variables, one could possibly enhance the efficiency of prediction by applying more sophisticated statistical methods, such as neural networks on more detailed content-related parameters e.g., the controversy measure of the article [38].

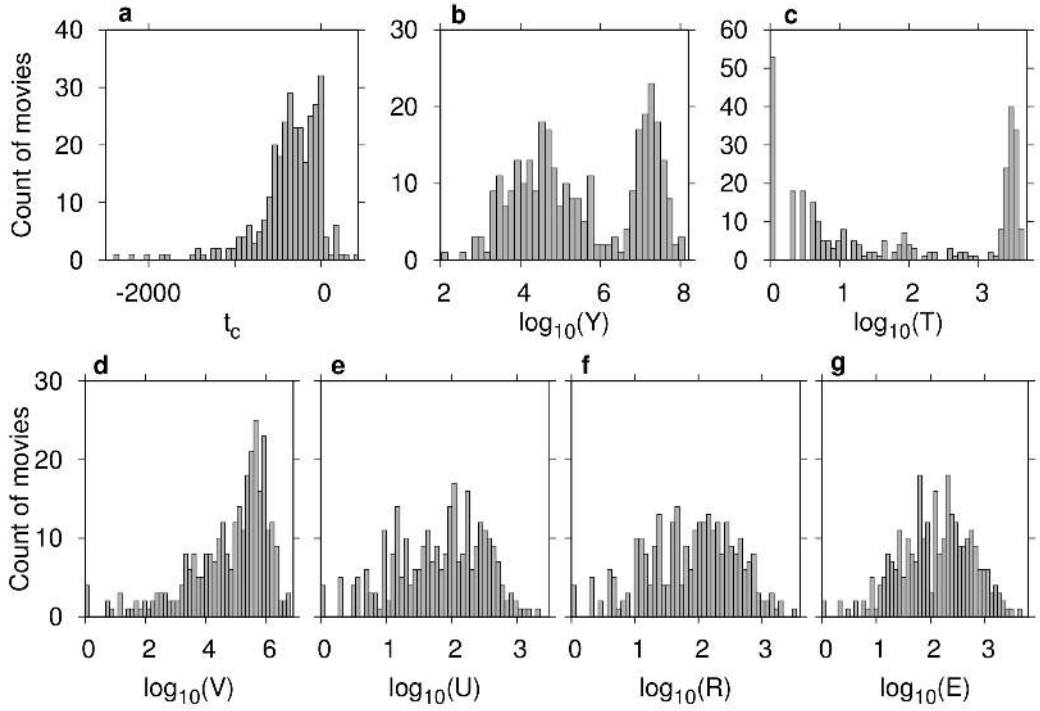

Figure 1. Histograms of different variables for our sample of $n = 312$ movies from 2010. A: Time of creation t_c of the corresponding article in Wikipedia, shown in days of *movie time* ($t = 0$ is the release time), B: Release weekend box office revenue in the U. S., in USD C: *number of theaters* that screened the movie on the first weekend, D: Accumulated *number of views*, and E: *users*, F: *edits*, G: *rigor* for the Wikipedia page up to $t = 7$ days after release.

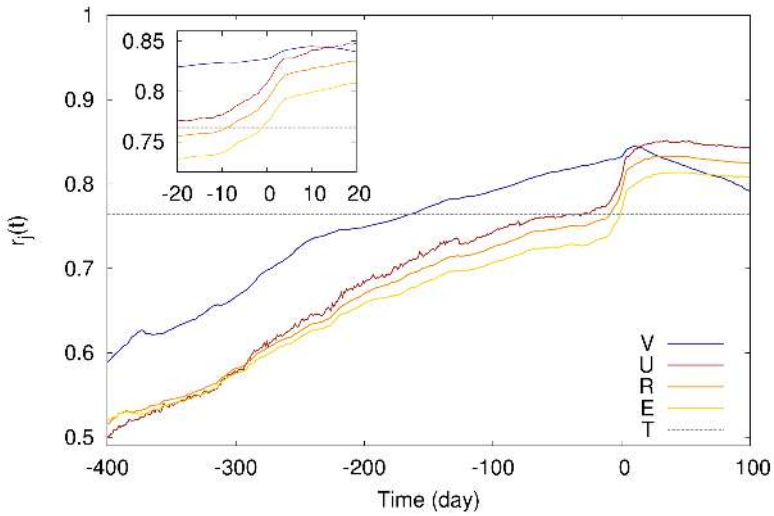

Figure 2. Temporal evolution of $r_j(t)$, the Pearson correlation of the box office revenue with different predictors. The shorthands V, U, R, E, and T denote the *number of views*, the *number of users*, the *rigor*, the *number of edits*, and the *number of theaters*, respectively. Time is measured in movie time. *Inset:* magnified detail of the main panel, showing the Pearson correlation around the day of release. Dashed horizontal line shows the correlation for *the number of theaters.*

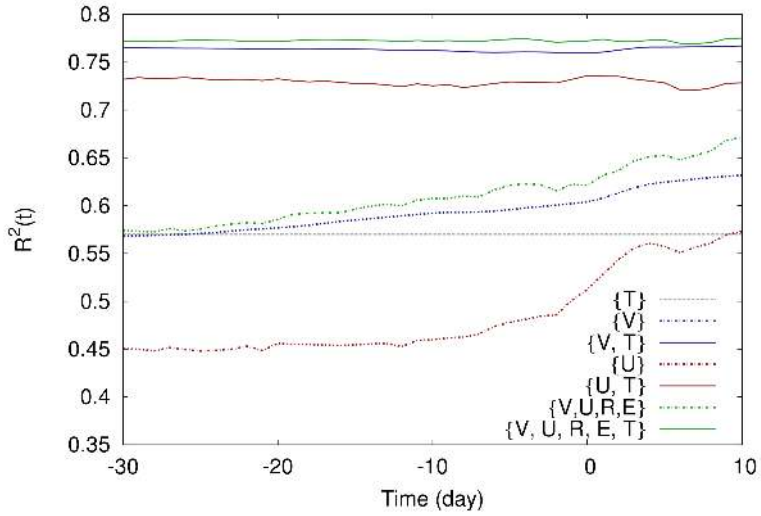

Figure 3. Coefficient of determination of the multivariate linear regression model fed by different set of input variables. The shorthands V, U, R, E, and T denote the *number of views*, the *number of users*, the *rigor*, the *number of edits*, and the *number of theaters*, respectively. The coefficient of determination was calculated using 10-fold cross-validation (see the Methods section). The dashed gray line shows the coefficient of determination for linear regression solely based on the *number of theaters.*

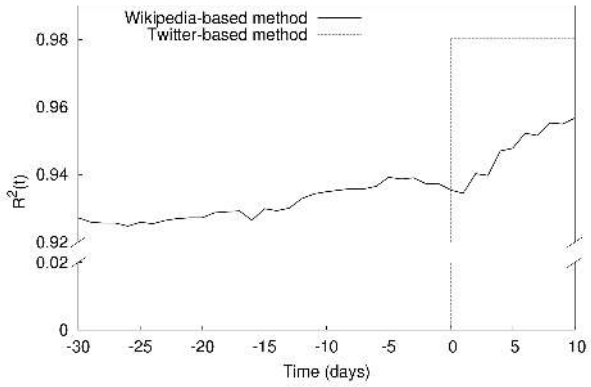

Figure 4. Comparison of the results with the Twitter-based prediction in Asur and Huberman work [27]. Same sample of 24 movies is considered as both training and test set. The coefficient of determination obtained with the Twitter-based method is 0.98 at the night of the release (day 0 in movie time).

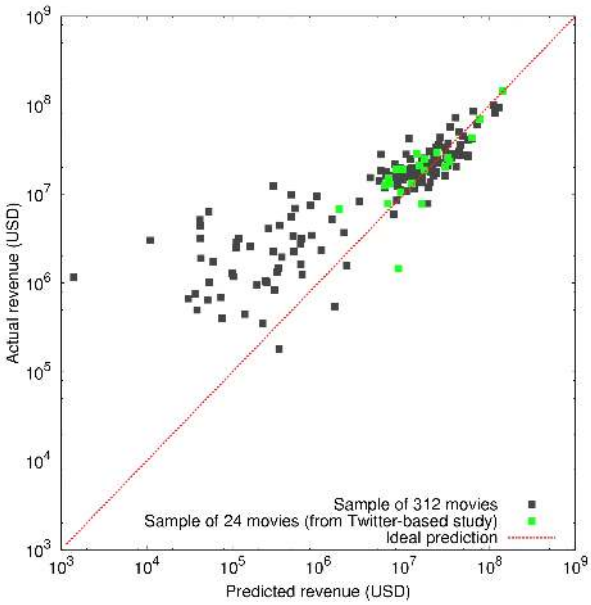

Figure 5. First weekend box office revenue in the U. S. against its predicted value by the Wikipedia model at $t = -30$ days. Green dots are representing the smaller sample of 24 movies common in Twitter and Wikipedia studies, and black dots are movies from the 2010 sample of 312 movies. Note that negative predicted revenues for some of the very unpopular movies could not be shown in the logarithmic scale.

Methods

In this study we consider a sample of 312 movies, which were released in the United States in 2010. The complete dataset including the financial data as well as Wikipedia activity records is available via the Supplementary Data S1. To obtain this dataset, first the list of 2010 movies distributed in the U. S. is acquired from Box Office Mojo (`http://boxofficemojo.com`) along with their accompanying financial data (535 movies). Financial data consist of the opening weekend box office revenue and the number of theaters screening the movie.

In order to locate the corresponding articles in Wikipedia, we use the category system of Wikipedia. Wikipedia articles are classified into one or more categories by users. We match the title of the movies in the Mojo database with the title of Wikipedia pages in categories `2009 films` and `2010 films`. Inclusion of the category `2009 films` is necessary because of movies that were released in 2010 in the U.S. but which could have already entered the international market during 2009, and hence were classified in the category `2009 films` in Wikipedia. To achieve the best possible match of the titles, they were stripped of punctuation and postfixes. Wikipedia uses the latter to maintain the uniqueness of every title, such as in the case of `Avatar (2009 film)` and `Avatar (computing)`. As a result of the matching process described above, a sample consisting of the financial data and the corresponding Wikipedia page for 312 movies was obtained.

For the sake of convenience we introduce *movie time*, a common time coordinate for the movies in the scope of our study. By definition, movie time is measured from the time of release in the U.S. All temporal variables are measured in movie time. Throughout this study, we consider accumulated values of parameters from the inception of the article to the prediction time t for each activity measure. The four activity measures are defined as the following:

Number of users, U: the number of different human users who contributed to the page.

Number of edits, E: the number of modifications made by human users on the article.

Collaborative rigor, R: similar to the number of edits; however it counts multiple subsequent edits by the same user as one edit [54]. It avoids counting multiple edits by the same user in a short period, e.g., to correct errors in their previous contribution.

A schematic illustration of these activity measures is presented in Fig. 6. These three variables are calculated using the page history databases of Wikimedia Toolserver (!http://toolserver.wikimedia.org!), which register information about every modification made to the pages of Wikipedia. To ensure that the above variables count solely human activity, contributions made by *bots* are excluded from calculations. Bots are automated scripts which facilitate automatic tasks such as spell checking. Contributions made by bots are registered in the same way as revisions by humans; however, they can be distinguished from human activity by noting a special entry in the databases of Wikimedia Toolserver, called the *bot flag*.

Number of views, V: the number of times a given page is viewed from its inception up to the time t. This data is extracted from the page view statistics section of the Wikimedia Downloads site (!http://dumps.wikimedia.org/other/pagecounts-raw!) through the web-based interface of "Wikipedia article traffic statistics" (!http://stats.grok.se!). Wikimedia Downloads counts views only since December 2007 and the view count data for July 2008 is corrupted. Therefore it is impossible to count the exact total number of views till the time of prediction for all considered pages. We have counted the page hits from $t = -500$ days before release, which according to Fig. 1(A), is sufficiently early. Another challenge is created by the renaming of the articles, which splits page hit counts into subsets according to the various titles the page possesses throughout its history. To cope with this problem, we followed the logs of "title moves" in the article history to track back and merge the whole page hits. Note that in the the dataset there are records on Wikipedia page requests for non-existing pages as well, which give us an indicator of the public interest in a movie even before its Wikipedia article is created and therefore we did not exclude such records from the data. *Number of theaters*: the count of movie theaters that screen the movie on the first weekend of its release.

To calculate the coefficient of determination, we carry out 10-fold cross-validation by randomly di-

viding our sample of 2010 movies into 10 subsets first. In the next step the model is trained for the union of the 9 subsets and tested on the remaining 10th subset. This is repeated for all 10 permutations of the subsets and the coefficient of determination for the model is obtained as the average over the permutations.

Figure 6. Illustration of different variables characterizing the activity of Wikipedia editors on an article. Each tick on the axis represents a modification of the page. Different tick styles refer to different users.

Acknowledgments

We thank Wikimedia Deutschland e.V. for providing access to its databases on the Wikimedia Toolserver and IMDb, Inc. for the access to Box Office Mojo database. We also thank the PLoS ONE anonymous reviewers for useful comments. Partial financial support from EU's 7th Framework Program's FET-Open to ICTeCollective project no. 238597 and by the Academy of Finland, the Finnish Center of Excellence program, project no. 129670, and TEKES (FiDiPro) are gratefully acknowledged.

References

1. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. New York, NY, USA: ACM, WWW '10, pp. 851–860. doi:10.1145/1772690.1772777.

2. Okazaki M, Matsuo Y (2011) Semantic twitter: Analyzing tweets for real-time event notification. In: Breslin J, Burg T, Kim HG, Raftery T, Schmidt JH, editors, Recent Trends and Developments in Social Software, Springer Berlin / Heidelberg, volume 6045 of *Lecture Notes in Computer Science*. pp. 63-74.

3. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. Journal of Computational Science 2: 1 - 8.

4. Saavedra S, Duch J, Uzzi B (2011) Tracking traders' understanding of the market using e-communication data. PLoS ONE 6: e26705.

5. Tumasjan A, Sprenger TO, Sander PG, Welpe IM (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 178-185.

6. Gayo-Avello D, Melaxas P, Mustafaraj E (2011) Limits of electoral predictions using Twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. pp. 490-493.

7. Gayo-Avello D (2012) "i wanted to predict elections with Twitter and all I got was this lousy paper" – a balanced survey on election prediction using Twitter data. priprint; arXiv:12046441 .

8. Brody T, Harnad S, Carr L (2006) Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology 57: 1060–1072.

9. Eysenbach G (2011) Can tweets predict citations? metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. J Med Internet Res .

10. Shuai X, Pepe A, Bollen J How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. PLoS ONE 7.

11. Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of The Royal Society A 368: 5707-5719.

12. Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. Sci Rep 2.

13. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012-1014.

14. Tsagkias E, de Rijke M, Weerkamp W (2009) Predicting the volume of comments on online news stories. In: ACM 18th Conference on Information and Knowledge Managment (CIKM 2009). ACM, Hong Kong: ACM, pp. 1765–1768.

15. Tsagkias E, Weerkamp W, de Rijke M (2010) News comments: Exploring, modeling, and online predicting. In: 32nd European Conference on Information Retrieval (ECIR 2010). Springer, Springer, pp. 109-203.

16. Carlos Castillo JPMS Mohammed El-Haddad (2013) Characterizing the life cycle of online news stories using social media reactions. priprint; arXiv:13043010 .

17. Tsagkias M (2012) Mining Social Media: Tracking Content and Predicting Behavior. Ph.D. thesis, University of Amsterdam.

18. Sinha S, Raghavendra S (2004) Hollywood blockbusters and long-tailed distributions: An empirical study of the popularity of movies. Eur Phys J B 42: 293-296.

19. Sinha S, Pan RK (2005) Blockbusters, bombs and sleepers: The income distribution of movies. In: Chatterjee A, Yarlagadda S, Chakrabarti BK, editors, Econophysics of Wealth Distributions, Springer Milan, New Economic Windows. pp. 43-47.

20. Pan RK, Sinha S (2010) The statistical laws of popularity: universal properties of the box-office dynamics of motion pictures. New Journal of Physics 12: 115004.

21. Sreenivasan S (2013) Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. priprint; arXiv:13040786 .

22. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications 30: 243-254.

23. Joshi M, Das D, Gimpel K, Smith N (2010) Movie reviews and revenues: An experiment in text regression. In: Proceedings of NAACL-HLT 2010, Short Papers Track.

24. Ishii A, Arakaki H, Matsuda N, Umemura S, Urushidani T, et al. (2012) The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process. New Journal of Physics 14: 063018.

25. R CAH, Castro A, Rodriguez-Sickert C (2006) The effect of social interactions in the primary consumption life cycle of motion pictures. New Journal of Physics 8: 52.

26. Mishne G, Glance N (2006) Predicting movie sales from Blogger sentiment. In: Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW).

27. Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. pp. 492-499.

28. Wong FMF, Sen S, Chiang M (2012) Why watching movie tweets won't tell the whole story? In: Proceedings of the 2012 ACM workshop on Workshop on online social networks. New York, NY, USA: ACM, WOSN '12, pp. 61–66. doi:10.1145/2342549.2342564.

29. Yun Q, Gloor PA (2012) The web mirrors value in the real world comparing a firms valuation with its Web network position. MIT Sloan Research Paper No 4973-12 Available at SSRN: http://ssrncom/abstract=2157278 .

30. Oghina A, Breuss M, Tsagkias E, de Rijke M (2012) Predicting imdb movie ratings using social media. In: ECIR 2012: 34th European Conference on Information Retrieval. Springer-Verlag, Barcelona, Spain: Springer-Verlag, pp. 503–507.

31. Voss J (2005) Measuring Wikipedia. In: International Conference of the International Society for Scientometrics and Informetrics : 10th, Stockholm (Sweden), 24-28 July 2005.

32. Almeida RB, Mozafari B, Cho J (2007) On the evolution of Wikipedia. In: Proceedings of the International Conference on Weblogs and Social Media. ICWSM'07.

33. Suh B, Convertino G, Chi EH, Pirolli P (2009) The singularity is not near: slowing growth of Wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration. New York, NY, USA: ACM, WikiSym '09, pp. 8:1–8:10.

34. Holloway T, Bozicevic M, Börner K (2007) Analyzing and visualizing the semantic coverage of Wikipedia and its authors. Complexity 12: 30–40.

35. Halavais A, Lackaff D (2008) An analysis of topical coverage of Wikipedia. Journal of Computer-Mediated Communication 13: 429–440.

36. Taraborelli D, Ciampaglia G (2010) Beyond notability. collective deliberation on content inclusion in Wikipedia. In: Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on. pp. 122 -125.

37. Sumi R, Yasseri T, Rung A, Kornai A, Kertész J (2011) Characterization and prediction of Wikipedia edit wars. In: Proceedings of the ACM WebSci'11, Koblenz, Germany. pp. 1–3.

38. Sumi R, Yasseri T, Rung A, Kornai A, Kertész J (2011) Edit wars in Wikipedia. In: Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom). pp. 724-727.

39. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J (2012) Dynamics of conflicts in Wikipedia. PloS ONE 7: e38869.

40. Yasseri T, Spoerri A, Graham M, Kertész J (2014) The most controversial topics in Wikipedia: A multilingual and geographical analysis. In: Fichman P, Hara N, editors, Global Wikipedia: International and cross-cultural issues in online collaboration. Scarecrow Press.

41. Török J, Iñiguez G, Yasseri T, San Miguel M, Kaski K, et al. (2013) Opinions, conflicts and consensus: Modeling social dynamics in a collaborative environment. Phys Rev Lett 110: 088701.

42. Yasseri T, Sumi R, Kertész J (2012) Circadian patterns of Wikipedia editorial activity: A demographic analysis. PLoS ONE 7: e30091.

43. Yasseri T, Kornai A, Kertész J (2012) A practical approach to language complexity: a Wikipedia case study. PLoS ONE 7: e48386.

44. Nielsen FA (2011). Wikipedia research and tools: Review and comments. Available at `http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6012/pdf/imm6012.pdf`.

45. Jullien N (2012). What we know about Wikipedia: A review of the literature analyzing the project(s). Available at SSRN: `http://ssrn.com/abstract=2053597`.

46. Yasseri T, Kertész J (2013) Value production in a collaborative environment. Journal of Statistical Physics 151: 414-439.

47. Keegan B, Gergle D, Contractor NS (2011) Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the tōhoku catastrophes. In: Int. Sym. Wikis. pp. 105-113.

48. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A (2010) Characterizing and modeling the dynamics of online popularity. Phys Rev Lett 105: 158701.

49. Spoerri A (2007) What is popular on Wikipedia and why? First Monday 12.

50. Spoerri A (2007) Visualizing the overlap between the 100 most visited pages on Wikipedia for September 2006 to January 2007. First Monday 12.

51. Osborne M, Petrović S, McCreadie R, Macdonald C, Ounis I (2012) Bieber no more: First story detection using Twitter and Wikipedia. In: Proceedings of the Workshop on Time-aware Information Access. TAIA'12.

52. Georgescu M, Kanhabua N, Krause D, Nejdl W, Siersdorfer S (2013) Extracting event-related information from article updates in wikipedia. In: Serdyukov P, Braslavski P, Kuznetsov O Sergei, Kamps J, Rüger S, et al., editors, Advances in Information Retrieval, Springer Berlin Heidelberg, volume 7814 of *Lecture Notes in Computer Science*. pp. 254-266.

53. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. Sci Rep 3: 1801.

54. Kimmons R (2011) Understanding collaboration in Wikipedia. First Monday 16.

Supporting Information

Dataset S1

The dataset under study, including the financial and Wikipedia activity data is also available at `http://wwm.phy.bme.hu/SupplementaryDataS1.zip`

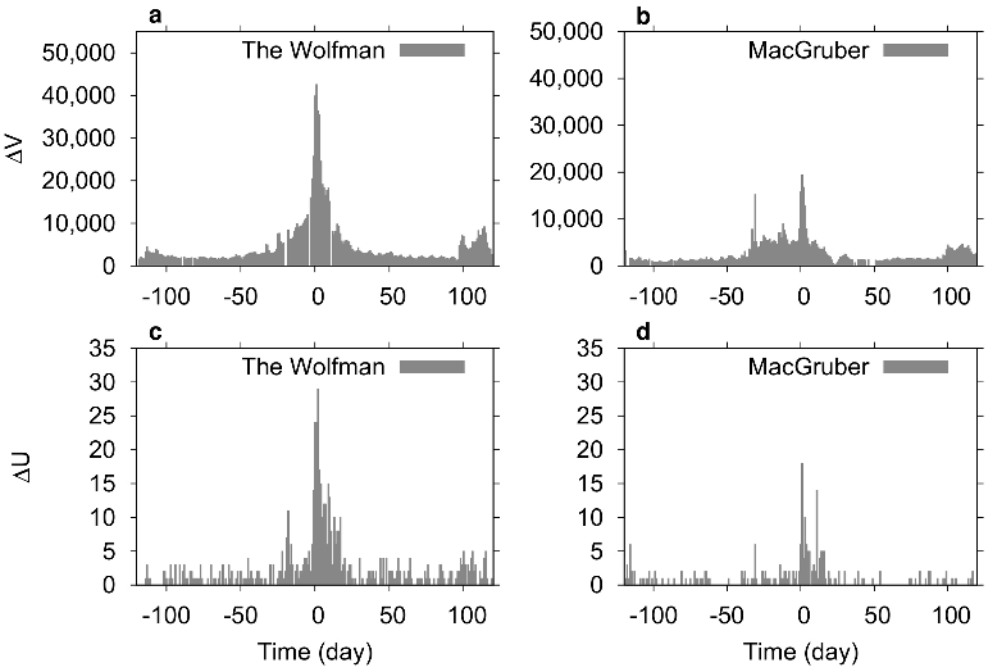

Figure S1. Temporal evolution of Wikipedia-based predictors for two individual movies: `The Wolfman (2010)` and `MacGruber`. The daily increments of *number of views* ΔV and *number of users* ΔU are shown for the articles in English Wikipedia that correspond to the two movies. The temporal axis shows movie time, i.e., a time-frame in which $t = 0$ corresponds to the release date. `The Wolfman` earned a box office revenue of $\$31,479,235$ on the release weekend while `MacGruber` gained only $\$4,043,495$. Accordingly, predictor variables take larger values in the case of `The Wolfman`.