Forecasting Audience Increase on YouTube

Matthew Rowe

Knowledge Media Institute, The Open University,
Milton Keynes, United Kingdom
m.c.rowe@open.ac.uk

Abstract. User profiles constructed on Social Web platforms are often motivated by the need to maximise user reputation within a community. Subscriber, or follower, counts are an indicator of the influence and standing that the user has, where greater values indicate a greater perception or regard for what the user has to say or share. However, at present there lacks an understanding of the factors that lead to an increase in such audience levels, and how a user's behaviour can affect their reputation. In this paper we attempt to fill this gap, by examining data collected from YouTube over regular time intervals. We explore the correlation between the subscriber counts and several behaviour features - extracted from both the user's profile and the content they have shared. Through the use of a Multiple Linear Regression model we are able to forecast the audience levels that users will yield based on observed behaviour. Combining such a model with an exhaustive feature selection process, we yield statistically significant performance over a baseline model containing all features.

Keywords: User modelling, Forecasting, Social Web, Data Mining, Behaviour

1 Introduction

A fundamental quality of Social Web platforms is the provision of profile building functionality, allowing users to construct an identity within such systems that conveys their interests and persona in a bespoke form. The motivation for constructing user profiles is often born of the need to share content and develop a reputation within a given community, or site. The Oxford English dictionary defines reputation as:

> *...the beliefs or opinions that are generally held about someone or something.*[1]

In developing a reputation within a Social Web platform, one is able to accrue a larger audience comprised of other users who are keen to listen to, read or watch the content that is being shared. Therefore the greater the reputation and standing of a user within the platform, the more subscribers and followers the user will gain. Reputation is also synonymous with influence, providing a

[1] http://oxforddictionaries.com/view/entry/m_en_gb0702720

mutually beneficial relationship where the greater the reputation of an individual the greater their influence, and vice versa.

Quantifying the standing and perception of a user can be achieved by measuring his/her in-degree on a given Social Web platform - i.e. the number of subscribers they have, or rather, the size of their audience. Therefore a key motivation behind the participation of users on Social Web platforms, and the construction of their profiles and development of reputation, is to increase this in-degree level. Exploring the relation between subscriber counts and behaviour features is particularly important on Social Web platforms that revolve around the sharing of *social objects* - i.e. videos, photos, etc. At present there is no understanding of what factors contribute to an increase in subscriber counts, and audience levels, nor is there an analysis of the affects of user behaviour on such levels. Exploring such a relationship would identify patterns of usage and isolate the correlation between certain behaviours and the advance of a user's reputation.

In this paper we seek to provide the link between reputation - quantified through audience levels - and the behaviour charecteristics of both users and the content they share. For our analysis we use the video-sharing, social networking platform YouTube,[2] exploring the question: *What factors influence audience levels?* For behavioural features we extract user profile attributes, such as the number of channels the user has subscribed to and the number of videos they have watched, and attributes of the content they have shared, such as the number of views and favourites their content has had.

To perform our analysis we employ Multiple Linear Regression (MLR) Models. First, we assess the correlation between subscriber counts and our collected behavioural features - identifying key feature correlations. Second we use observations from our analysis to forecast audience levels of individual users at various time-steps, comparing a MLR model using all features against another MLR model post-feature selection. Experiments are performed over 10 days worth of data collected from YouTube over 4 hour time intervals.

This paper is structured as follows: section 2 explains our data collection approach, the features we model with an overview of the collected datasets, and the role that semantics plays in our work. Section 3 presents our approach to forecasting subscribers using Multiple Linear Regression modelling over time-series data, detailing the correlation between the in-degree of users and other behaviour features, and our experiments with forecasting in-degree changes. Section 4 describes related work to our approach and section 5 concludes the paper.

2 Data Collection and Overview

In order to predict the number of followers - denoted as subscribers on YouTube - that a given person will develop over time we required data against which our forecasting methods could be empirically compared. To gather such a collection

[2] `http://www.youtube.com`

we utilised the YouTube Data API,[3] using the following process: we queried the API for the most recent 100 uploaded videos to the UK version of YouTube - in order to increase the likelihood that we gathered English-language videos, as our future work will look at the language features of the videos and comments. For each video collected, we wanted to analyse the relationship between audience levels of the person who shared the content and both the behaviour of the user and the reception of the content they have shared. Therefore, every 4 hours, we logged 6 individual statistical features associated with both the video uploader and the video itself. Below we describe these features and our reason for their selection:

- *User Statistics:*
  - *In-degree:* Measures the number of users who are currently subscribed to the user - i.e. the video uploader. This feature provides the value that we wish to predict for individual users: audience levels.
  - *Out-degree:* Measures the number of YouTube channels, and therefore other users, that the given user is currently following. This captures the openness of the user and the degree to which they are interested in following the content of other users.
  - *User View Count:* Measures the number of unique views of videos on YouTube. Provides a measure of the extent to which the user watches content and the time spent on the platform, thereby gauging the activity behaviour of the user.
  - *Post count:* Measures the number of unique videos uploaded by the user onto the platform. Allows the extent to which the user shares original content with the community to be gauged.
- *Post Statistics:*
  - *Post View Count:* Measures the number of times a given video has been watched. This provides a notion of popularity for the user's uploaded content. As we are only logging the statistics for individual videos, this measure is for a user's single video, not for all videos uploaded by the user.
  - *Favourite count:* Measures the number of times a given video has been *'favourited'* by users. Like the Post View Count, this feature assesses the popularity of the uploaded content, but in a more extreme sense - given that the action of *'favouriting'* a video requires the viewer to watch the video and be impressed by what he/she sees.

At 4 hour time intervals we collected the most recent 100 uploads to UK YouTube, thereby progressively building up a collection of videos, stopping once we had collected 2000. This process was continued for 10 days, logging the details of the 2000 collected videos. This method of collection is similar to the one used in [8], which collected a dataset for inspection of view counts as an indicator of popularity analysed over time.

[3] `http://code.google.com/apis/YouTube/2.0/reference.html`

For our analysis however we wanted to use a scaled down version of the dataset, that would allow graphical inspection of the data. Therefore we randomly chose 10% of this dataset for use in our analysis and experiments - producing a collection of 200 videos. For each of the videos we analyse the first 10 time steps from each video's upload. In the following section we detail our approach for forecasting subscriber counts for individual users, utilising Multiple Linear Regression Models for this process. In order to induce these models we require analysis to be performed over the data, thereby assessing the fit of the models with respect to the data. To do this *model selection* phase we divided the dataset of 200 videos, and their logged information over time, into a training/testing split using an 80/20 random split. The former set provides the data over which we perform analysis of features and their change over time.

2.1 The Role of Semantics

At present our analysis only covers a small facet of the Social Web: YouTube. The models that we explore and learn in this paper will form the basis for comparisons against other Social Web platforms - e.g. Twitter[4] and Flckr[5] - allowing patterns in behaviour to be associated with audience levels and the effects of applying such patterns observed. The heterogeneity of the Social Web hinders such explorations without a common understanding of statistical features across domains, and the annotation of such features using a common schema or ontology. To this end the information that we have logged is converted to RDF and labelled with concepts from a Behaviour Ontology[6] specifically designed for this purpose.

The ontology contains an abstract concept called *Impact* that captures the impact of either a post or a user at a given point in time. This concept is specialised for a post using the class *PostImpact*, allowing the number of views and favourites that a piece of content has incurred to be recorded as data type properties at a given point in time. Likewise *UserImpact* records statistical features of the user at a given point in time from the above list. We omit further details of the ontology in order to focus on feature analysis and forecasting, however it is suffice to say that the role of semantics at present is not fully exploited, given that we are only using information collected from a single Social Web platform. Our future work, that we describe within the conclusions of this paper, seeks to extend this work to cover forecasting using other social data.

3 Forecasting Audience Increase

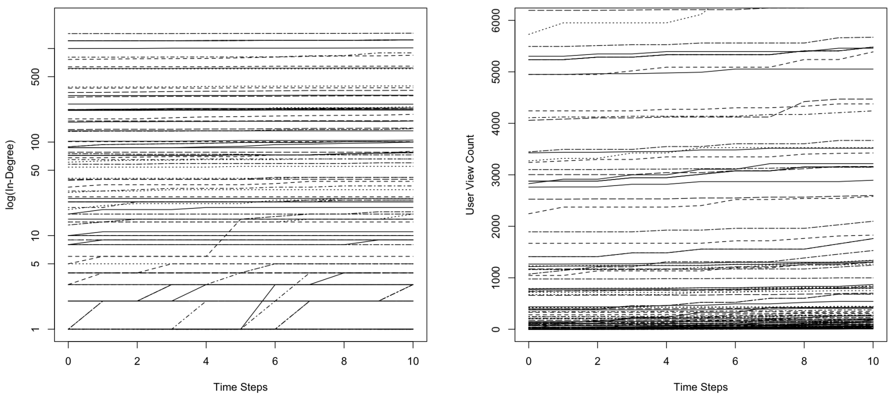
Figure 1 shows the change in in-degree level of YouTube users over time - shown as a *logged* value in Figure 1(a) - and the increase in view counts of the uploaded videos over time - shown in Figure 1(b). These graphs demonstrate that, in each

[4] http://www.twitter.com

[5] http://www.flickr.com

[6] http://people.kmi.open.ac.uk/rowe/ontologies/UserOnto_0.23.rdf

case, we observe that as time increases, so too does the in-degree of the user and the view counts of the video. This prompts the question as to which features of the user, and the content that they have uploaded, are correlated with the increase in their subscriber count, and thus the reputation and influence that they may have within the community.



(a) Time vs In-Degree

(b) Time vs View Count

Fig. 1. Analysis of 80% training split, showing an increase in the in-degree of the user over time, and the increase in view counts over time also.

To analyse this correlation we employ a Multiple Linear Regression (MLR) model, allowing multivariate data to be employed as the empirical basis for model induction. The general form MLR, the form of which is shown in Eq. (1), is to learn some coefficients - contained within β - that fit the observed data, allowing the model to be assessed for *goodness-of-fit*. Within this model α denotes the intercept of the induced plane within an n-dimensional space and ϵ denotes a normally-distributed random error vector.

$$Y = \alpha + \sum_{i=1}^{n} \beta_i X_i + \epsilon \tag{1}$$

As we described previously, our data contains various variables that we have measured over incremental time periods. We wish to incorporate these variables into a MLR model, and then assess the correlation that each feature has with the dependent variable - in our case the in-degree of the user. Table 1 details the features used, together with their vector notation labels. Our data is however time-series, in that we have collected it over successive 4-hourly increments, therefore X_i is actually a vector of length n corresponding to n time steps.

Table 1. Features used within the model and their corresponding independent variable labels within the models

Label	Time-series Feature
X1	User In Degree
X2	User Out Degree
X3	User View Count
X4	User Post Count
X5	Post View Count
X6	Post Favourite Count

3.1 Model Selection

The first analysis that we perform is to assess the fit of a model using all possible features. Given that our dependent variable is the in-degree of a user - denoted by $X1$ - we can rewrite Eq. (1) as the form shown in Eq. (2):

$$X1 = \alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon \qquad (2)$$

Using this model we can assess the correlation of the model's variables with the in-degree of the user, thereby asking: *What is the correlation between in-degree and other features?* To do this we took the training split of our dataset and derived average measures for each of the features from Table 1 from all of the users within that split. Our intuition is that we can observe a general pattern in the data that can be later used for forecasting over individual samples (users) - we denote this model, including all the features, as Ψ_{all}.

Table 2. Model results for Ψ_{all}

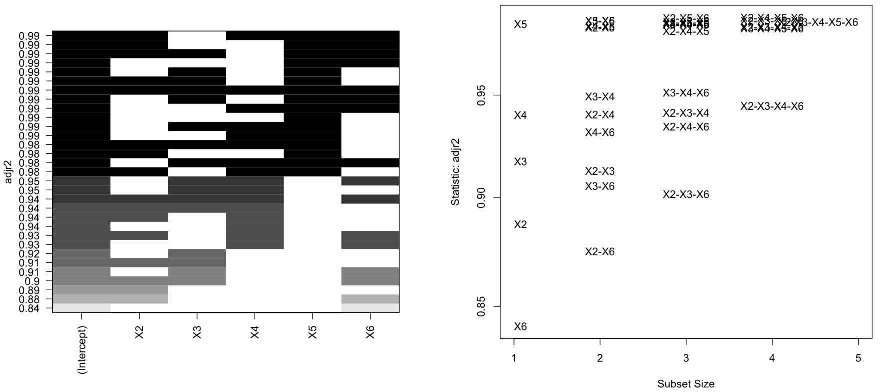
Feature	Est' Coefficient	Standard Error	t-Value	P(x >t)
X2	-4.565e+00	3.025e+00	-1.509	0.205719
X3	2.239e-04	7.193e-04	0.311	0.771099
X4	-1.165e+00	1.408e+00	-0.827	0.454550
X5	3.160e-02	7.518e-03	4.203	0.013664*
X6	2.220e+00	1.594e+00	1.393	0.236016
Summary Res. St Err: 0.2447 Adj R^2: 0.9872 $F_{5,4}$: 140.3 p-value: 0.0001399				

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1

The results shown in Table 2 are form evaluating the null hypothesis that no significant relationship exists between the dependent variable - the in-degree of the user - and the individual independent variables, using the t-test and the given t-Value for specific features. This hypothesis is rejected if a statistically significant relationship exists between the two variables and the p-value is below a given significance level. The results indicate that only X5 - post view count

- is found to have a significant correlation with the in-degree of the user. This correlation suggests that as the video that the user has uploaded incurs more views, they gain more subscribers. We also note the low p-values for the number of favourites that the user's content has incurred, indicating a good correlation between the in-degree and rating of their content, but one that is not significant. As the results show, using a model fitted with all the features does not identify any other features as having a significant correlation with the in-degree of the user. The adjusted coefficient of determination - adjusted R^2 - shows a strong fit of the model to the data, and the ability of the model to predict future in-degree values.

Up to this point our analysis has concentrated on using all possible features, and their correlation with the in-degree of YouTube users. However, some feature combinations may have a stronger correlation with the dependent variable, prompting the question: *What are the best features for in-degree prediction?* To identify such features we used an exhaustive search of possible feature combinations and evaluated the fit of each combination using adjusted R^2 values. This method trialled different subsets of the entire set of features, and ranked their fit with respect to the dependent variable.



(a) Feature combination rankings by R^2 values

(b) Feature subsets plotted by R^2 values

Fig. 2. Adjusted R^2 values achieved using different feature subsets

Figure 2 shows the ranking of different feature subsets, where the best fitting model is found to be Ψ_{all} without X3 - the number of videos that the user has viewed. Figure 2(b) shows the ranking of the different trial subsets, and the various adjusted R^2 values that the differing combinations achieve. The optimum feature combination is that which maximises the adjusted R^2 values and appears above other feature permutations in the graph. As expected, as the subset size increases, the performance of the subsets models does too. This figure also demonstrates the strong correlation between the number of views that a video receives and the in-degree of the user - given the high adjusted R^2 value yielded for the use of solely X5 within a model. At the same subset size of 1, the number of posts that the user has made on the site - i.e. videos uploaded - is found also have high predictive value, followed by the number of the videos that the user has watched. As one would expect, as a user interacts more on the platform and watches more videos and uploads more content, other members of the community's awareness is raised, and the user is noticed - resulting in an increase in subscribers and therefore audience members.

Table 3. Results for Ψ_{best}

Feature	Est' Coefficient	Standard Error	t-Value	P(x >t)
X2	-4.594196	2.736807	-1.679	0.15406
X4	-1.301767	1.211048	-1.075	0.33153
X5	0.032853	0.005746	5.718	0.00229**
X6	2.528960	1.128538	2.241	0.07513.
Summary Res. St Err: 0.2215 Adj R^2: 0.9895 $F_{4,5}$: 213.9 p-value: 8.963e-06				

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1

Feature selection identifies an optimal combination of features as being the inclusion of all original features from Ψ_{all} without X3, we therefore construct another model using this combination of features and denote this as Ψ_{best}. The fit of this new model may alter the correlation of the independent variables with the in-degree of the user, given the omission of a previous feature, therefore we explore the following question: *What is the correlation between in-degree and best features following feature selection?* Table 3 shows the results from a MLR model fitted with the best performing features. We note that unlike the previous model, the number of favourites that a video has yielded is now found to have a significant correlation with the in-degree of the user - at a significance level of $\alpha = 0.1$. The induced coefficients show an interesting pattern in the relationship of the model, given that an increase in the number of videos viewed by the user actually has a negative effect on in-degree increase, however this feature is not statistically significant within the model, and therefore conclusions drawn from such a correlation are not well supported. We formally define this new model as:

$$X1 = \alpha + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon \tag{3}$$

3.2 Forecasting

Model selection has identified two models that we can use for prediction: 1) Ψ_{all} - using all 5 features; and 2) Ψ_{best} - using 4 features identified as optimising adjusted R^2 values. Our analysis was performed over the 80% training split, allowing general patterns in the data to be observed by averaging the features. In this stage we now wish to test the predictive quality of these two models and their ability to forecast the in-degree of individual users. For our experiments we test two scenarios: the first predicts the in-degree at time t using the previous k time steps as training data for inducing the model's coefficients. The second experiment predicts the in-degree at the final time step - i.e. $t = 10$ - when trained on the previous k steps.

Unlike the previous model selection stage, at this point we perform predictions at the *micro-level* for each user in our held-out test split of 20% - therefore predicting the in-degree for 40 users. To test the performance of each of our models we train both models for a single person, using their respective features collected over time, rather than building a general model as in the previous section using averages from which the coefficients are then induced. Our predicted in-degree should match, as close as possible, to the actual observed in-degree in the data, therefore to measure the error in predictions we measure the Root Mean Square Error of the prediction as follows: let $\hat{Y}$ denote our predicted value and Y denote our actual observed value, we define the Root Mean Square Error (RMSE) as:

$$RMSE(\hat{Y}, Y) = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n}} \tag{4}$$

One-Step Forecast Our first experiment, as described above, tests the predictive quality of the two models when forecasting the in-degree of each user one-step ahead. That is to say that at time $t = k + 1$, the model is trained using the previous k time steps. Table 4 presents the RMSE achieved by both Ψ_{all} and Ψ_{best} as k is iteratively increased. Training each model using only 1 step achieves equivalent performance, however as k is increased we see differences in the RMSE produced by each model, culminating in an average performance by the latter - Ψ_{best} - which outperforms the use of all features.

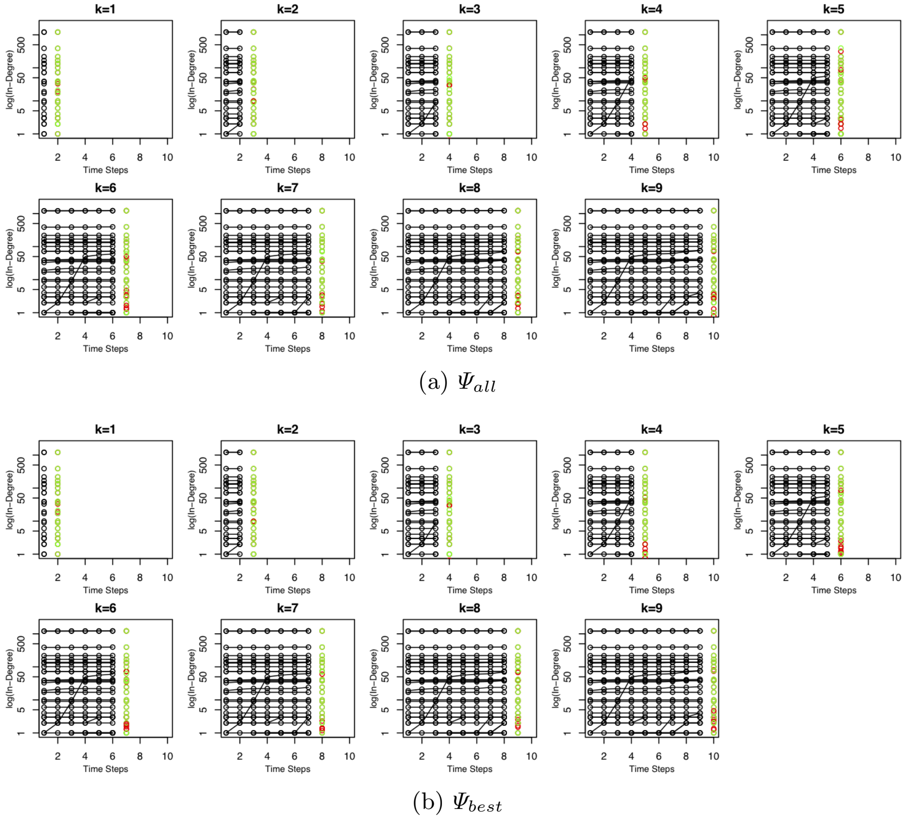
Table 4. Prediction Error Rates for Ψ_{all} and Ψ_{best} predicting the in-degree at the next time-step when trained using the past k steps.

Model	1	2	3	4	5	6	7	8	9	Average
Ψ_{all}	0.707	0.720	1.101	0.602	1.282	0.345	3.794	0.464	0.478	1.055
Ψ_{best}	0.707	0.622	1.027	0.542	0.575	0.248	0.868.	0.795	0.605	0.665

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 .

In order to assess the significance of the error values achieved at differing values of k for the tested models, we utilised the Sign Test. We tested the null hypothesis that there was no difference in performance between the models, rejecting this hypothesis should a significant improvement be found when using

Ψ_{best}. For each value of k the test was performed for all 40 test samples - denoting the 40 users in the test set whose in-degree was to be predicted. From our analysis we note that where $k = 7$, the model derived following feature selection outperforms the use of all features at a liberal significance level of $\alpha = 0.1$. The difference in performance between Ψ_{all} and Ψ_{best} over all values of k was not found to be statistically significant, indicating that the feature selection improves predictive quality but not in a significant manner.



(a) Ψ_{all}

(b) Ψ_{best}

Fig. 3. Predictions (green) and observations (red) for user in-degree at next $k + 1$ time-step, when trained using past k steps.

Figure 3 presents the overlay plots for predicted in-degree values against the observed values for each of our 40 tested users. Predictions are denoted using green circles and observations using red circles, while training time-steps are denoted by black circles and path lines. Our goal is to minimise the number of red circles that can be seen by overlaying the green circles - denoting the predicted values, thereby providing a qualitative assessment of prediction accuracy.

Differences in performance between the two models can be observed at various values of k, for instance at $k = 5$ where Ψ_{best} has a notably less red circles, and therefore errors, for larger audience levels than Ψ_{all}.

Final-Step Forecast For our second experiment we explored the prediction of a user's in-degree at the final time step in the collected data - i.e. $t = 10$. Therefore each of the MLR models was trained using the previous k steps, and then applied to the final time step. As previous, we measure the RMSE achieved as k is iteratively increased and report on statistically significant results. As Table 5 indicates Ψ_{all} outperforms Ψ_{best} on average, where the difference in performance post-feature selection was found to be statistically significant at a significance level of $\alpha = 0.05$. However when comparing the results at individual values of k no significant improvements in performance were found. Qualitative analysis of the predictions using both models is shown in Figure 4, where we observe at $k = 6$ for larger in-degree levels the reduction in prediction errors when using Ψ_{best} over Ψ_{all}.

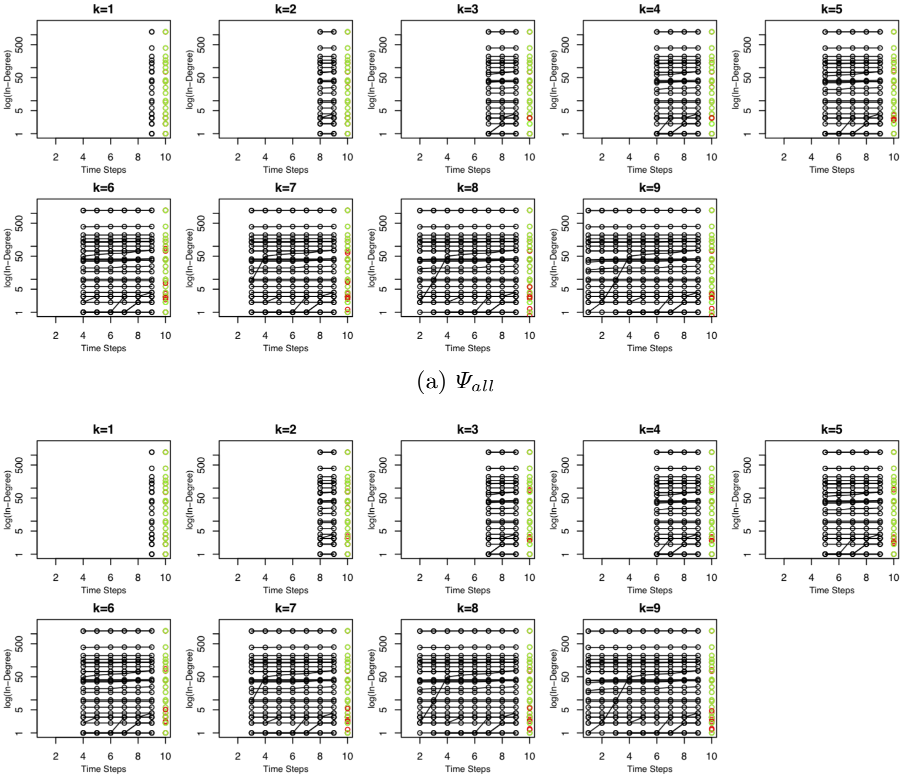
Table 5. Prediction Error Rates for Ψ_{all} and Ψ_{best} predicting the in-degree at the final step ($t = 10$) when trained using the past k steps.

Model	1	2	3	4	5	6	7	8	9	Average
Ψ_{all}	0.806	0.739	0.641	0.438	0.261	0.329	0.761	0.495	0.478	0.550
Ψ_{best}	0.806	0.546	0.352	0.161	0.149	0.259	0.571	0.490	0.605	0.438*

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 .

4 Related Work

The measurement of subscriber counts, and therefore in-degree values, as a gauge of influence has been studied in various pieces of recent work. For instance, work in [9] explored influence in citation networks, to assess the flow of topics amongst authors. Using Topical Affinity Propagation, the authors model the topic distribution of authors - the intension being to identity experts, or heavily influential nodes in citation networks. Their experiments found, naturally, that those nodes with a higher in-degree - derived from incoming citations - for a given topic, represented an influential member of that community. Citation networks were also studied by [10] - although they are defined as '*social networks*' - also, like our work, using Multiple Linear Regression models over time-series data. The motivation behind their work was to note the influence that content network measures had upon social network measures and vice-versa - explained through analysing the coefficients of the induced models. Experiments were performed over 4 yearly time steps of the collected citation networks - formed using SPARQL queries against bibliographic endpoints.



(a) Ψ_{all}

(b) Ψ_{best}

Fig. 4. Predictions (green) and observations (red) for user in-degree at final time step, when trained using previous k steps.

Assessing influence on Twitter is described in recent work by [4] where three distinct notions of influence are defined: in-degree influence, retweet influence and mention influence. The authors found that the in-degree of a user is not correlated with audience engagement - unlike retweets and mentions. Influence among social network members is also studied in [1], this time over data collected from the photo-sharing platform Flickr.[7] To test the affects of influence the authors examine the propagation of tags from nodes in the network, analysing whether a connected node, after a certain point, begin using a tag that their network members had used. The authors found no significant influence affect, however this could be explained by the nature of the data source used - given that Flickr is used to share photos of external resources, not necessarily connecting users by common topics.

[7] `http://www.flickr.com`

Prediction and forecasting using social data has been explored in [3] where the notion of user authority is compared against user affinity in the context of Yahoo! Answers. The described approach models the objectivity of the question poster: where high-objectivity is associated with the poster looking for an expert. Their results show that a combination of affinity and authority - the latter denoting influence - provides the best approach to predicting the best answer selection. The work described in [5] predicts whether a user will retweet a given URL or not. The authors claim that so-called '*powerusers*' are highly-influential users with many followers who contribute to information spread via retweets. They model a user's influence over time-series data collected from Twitter, inducing the model's parameter up to a point, and then using this model to predict propagation after the point in time. The model uses the notion of neighbourhood influence as one parameter, where the propagation probability is dependent on past influence in the surrounding network. Twitter is also used for prediction in [2] which predicts stock market levels from Twitter mood. First analysis is performed of the correlation between independent mood variables and the Dowjones Index - denoting the dependent variable. A fuzzy neural network is then used to predict the index based on the mood variables.

The analysis of in-degree distributions within social networks on Social Web platforms has been presented in work by [7] and [6]. The former, [7], assesses the statistical properties of social networks collected from several platforms, including YouTube, assessing the clustering coefficients of the collected networks, and the in-degree/out-degree distributions. Their findings showed a high level of local clustering, indicating the possibility of topical cliques forming around niche subjects, and a power-law distribution for the in-degree. In [6] the in-degree distribution on the popular technology-news web site Slashdot[8] is analysed. On Slashdot moderators post stories on to the site, functionality which is not available to commenters. However after sorting users by their in-degree, the authors identified the top-ranked individual not as a main content contributor, but as a commenter, indicating that the increase in subscribers/followers/friends is based on community interaction and discussion, and not on the creation of initial content.

Of the work studied, although several pieces deal with influence and the measurement of in-degree levels, no known work attempts to predict in-degree levels and the gain in reputation that a user could yield based on behavioural characteristics. The approach presented within this paper attempts to fill this gap by studying the correlation between in-degree levels and various features - similar to the modelling approach described in [2] - in order to identify predictive features for forecasting the evolution of a user's standing on a Social Web platform.

5 Conclusions

In this paper we have examined the relationship between behavioural features, of both users and the content they share, and user subscriber levels - attempting

[8] `http://slashdot.org/`

to quantify the reputation that a user has on a given Social Web platform. In essence we sought to explore *what factors influence audience levels?* We found that the reception of content by the platform had a strong link to an increase in audience levels. In particular we observed two key aspects from our analysis:

1. *The greater the number of views of uploaded content, the greater the audience levels of the user*
2. *The more content is 'favourited' by users, the greater the audience levels of the user*

Interestingly we observed a negative relationship between audience levels and the behaviour of the user, suggesting that as the user participates more in the community by watching more videos and uploading more content, that such behaviour can have a negative effect on the person's audience levels, and therefore his/her reputation. However it is worth noting that such relationships, although present in our analysis, are not statistically significant.

Combining the behavioural features into a single Multiple Linear Regression Model (Ψ_{all}) sought to explore the predictive power of feature combinations and the fit of the model to the empirical data. Post-feature selection, using exhaustive subset tests, found an alternative model (Ψ_{best}) using all possible behavioural features without the view count of the user. Comparison of the two models was then performed via two forecasting experiments over a held-out sample from our previous model selection stage. We found that for predicting a user's in-degree at a final time-step when trained on the previous k steps, Ψ_{best} showed statistically significant performance over Ψ_{all}.

Our future work will explore two key avenues of work: the first is to extend this study over a larger dataset, one that we are currently collecting, and expanding the feature scope to include reciprocal features to capture the interaction behaviour of the user - i.e. via commenting activity. The second avenue is to expand this approach to other Social Web platform - e.g. Flickr and Twitter - to mine patterns that explain the link between reputation on such platforms and the behaviour of its users. For this second area of work, we require semantics to encode information in a common, machine-readable form, from which our methods can then function. In section 2 of this paper we briefly discussed the behaviour ontology that is used to represent the statistical features employed without our approach in a common form. This ontology will form the basis for our future work, allowing information from disparate platforms and provided using heterogeneous data schemas, to be represented in a common, machine-readable form.

Acknowledgements

This work is funded by the EC-FP7 project Robust (grant number 257859).

References

1. A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 7–15, New York, NY, USA, 2008. ACM.
2. J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.
3. S. Budalakoti and K. S. Barber. Authority vs affinity: Modeling user intent in expert finding. *Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010 IEEE International Conference on*, 0:371–378, 2010.
4. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.
5. W. Galuba, D. Chakraborty, K. Aberer, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *3rd Workshop on Online Social Networks (WOSN 2010)*, 2010.
6. V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 645–654, New York, NY, USA, 2008. ACM.
7. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07*, 2007.
8. G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, 2010.
9. J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM.
10. S. Wang and P. Groth. Measuring the dynamic bi-directional influence between content and social networks. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference*, volume 6496 of *Lecture Notes in Computer Science*, pages 814–829. Springer Berlin / Heidelberg, 2010.