Analyzing a User's Contributive Social Capital Based on Acitivities in Online Social Networks and Media

Sebastian Schams
Technical University of Munich
Munich, Germany
sebastian.schams@in.tum.de

Jan Hauffa
Technical University of Munich
Munich, Germany
hauffa@in.tum.de

Georg Groh
Technical University of Munich
Munich, Germany
grohg@tum.de

ABSTRACT

To improve the quality of communication in Online Social Networks and Media (OSNEM), we envision a system that models a person's contributive social capital (CSC), which encompasses their competence, trustworthiness, and social responsibility. Having the CSC score available may inspire social behavior and mutual support. The system is based on three pillars: the analysis of OSNEM activity, interactions in virtual social capital market systems, and personal endorsements. In this paper we present our investigations regarding the first pillar. To obtain a dataset, we ran an experiment where 165 participants interacted on a custom social networking platform and assessed each other. Ground truth data was derived from these assessments. The dataset shows characteristics that are similar to larger OSNs. With different machine learning algorithms we investigated the hypothesis that contributive social capital can be extracted from network properties and networking activity, which were assessed with features such as the number of contributions of each participant. The prediction of contributive social capital showed an improvement over the baseline. A ranking of the participants following their predicted CSC scores showed a moderate correlation with the ranking according to the ground truth assessment. We also investigated the relative importance of the features for the analysis, and the effect of excluding inactive users to better understand network dynamics on a micro level. The selected features are also available in most other OSNEM platforms, like Facebook and Twitter. This allows a large-scale application of our investigations.

KEYWORDS

Network Analysis; Social Media Analysis; Contributive Social Capital; OSNEM Platforms; Information Extraction

ACM Reference Format:
Sebastian Schams, Jan Hauffa, and Georg Groh. 2018. Analyzing a User's Contributive Social Capital Based on Acitivities in Online Social Networks and Media. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191593

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191593

1 INTRODUCTION

When people converse, a considerable amount of information is transferred non-verbally. As more and more interactions take place online, some of the non-verbal cues go missing. This makes it more difficult to assess interaction partners, especially when interacting with anonymous users, whose motivations are unknown. It is a goal of current research to extract and assess user characteristics from different OSNEM platforms and thereby contribute to closing this information deficit.

A user property, whose extraction from online data sources is still relatively little studied, is social capital. There is a variety of definitions for social capital. Robison et al. [21] attribute this to the highly context-dependent nature of social capital and argue that social capital has often been defined with a specific application in mind. In general, one can differentiate between two types of social capital. The first describes the properties of social networks on a macro level. An exemplary definition in this context was given by Putnam, who describes social capital as "features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit." [18] Alternatively, one can look at social capital from the perspective of an individual (ego level) and describe their surrounding micro network. Lin, for example, describes social capital as "a broad concept, usually focused on the values obtained by being part of a social network and thus, referred to as the sum of social resources." [16]

In this paper we focus on individual social capital. But rather than looking at the social capital a person has access to by being part of a network, we look at the social capital each user adds to their social network. This has been described by Schams and Groh [22] as contributive social capital (CSC), which comprises of a person's value-add due to their competence, trustworthiness, and social responsibility. Most interactions on OSNEM can be characterized in terms of these three attributes. Knowledge and expertise are part of the competence assessment, which happens implicitly in the evaluation of many fact-based discussions and contributions. The trustworthiness aspect includes trust and reputation which guide the decision whether or not to trust the information provider. And finally, the aspect of social responsibility takes into account the willingness of a person to act socially towards others by helping or sharing information. The inclusion of trust and social responsibility are what separates the contributive social capital research from pure expert identification.

As a framework for the assessment of CSC, we envision a system that is based on three pillars: social network analysis, social capital market systems, and personal as well as institutional endorsements. The focus of this paper lies on the first pillar: social network and social content analysis.

In section 2, previous work on social capital extraction from online data sources is reviewed. The complete CSC assessment system is described in section 3. Section 4 describes the experiment we conducted to create a social network dataset with a ground truth assessment. This dataset is analyzed in section 5. Section 6 provides a summary and outlook on future work.

2 RELATED WORK

To the best of our knowledge, there are no publications encompassing the direct extraction of social capital or contributive social capital from online data sources. However, it can be argued that CSC is related to other properties, like expertise, trust, reputation, or influence. In this section we briefly review publications that investigate the extraction of these characteristics from different online data sources. A detailed overview can be found in prior work [22].

2.1 Analysis of social networking platforms

Hassan [11] investigated network features that might correlate with influence. He lists the number of likes and friends as features that belong to the class "recognition". The class of "activity generation" comprises features like the number of posts, number of received comments on written posts, number of shares of the user's posts by others, and the number of in-links (number of times the user or their posts are referenced). The number of times a user includes outlinks (references to sources given by URLs) is listed by Hassan as the class of "novelty".

An approach to infer and continuously update a user's influence was presented by Rao et al. [19] Aggregating data from different social media platforms and other sources, they trained a machine learning algorithm to calculate the Klout score, which is claimed to correlate with the real influence of a user.

2.2 Analysis of micro-blogging

Anger et al. [1] showed that ratios of different user statistics found on the micro-blogging service Twitter can be interpreted in terms of influence. On a set of Austrian Twitter users, they demonstrated that, e.g., a high ratio of Retweets and Mentions can identify an influential user.

With an algorithm similar to Google's PageRank [17], Weng et al. [25] identified influential Twitterers in different categories. In addition to their TwitterRank algorithm they investigated in-degree centrality, PageRank, and topic-sensitive PageRank.

Hadgu and Jäschke [10] used classification with support vector machines, classification and regression trees and random forests, as well as logistic regression to identify experts on Twitter. They used several features, like the total number of tweets, followers, and friends. Profile information and user statistics were also used. As ground truth for expertise they identified the profiles of scientists, as they can be regarded as experts of their fields. The precision of the classification was between 0.88 and 0.96. The most useful feature was the number of tweets.

2.3 Analysis of threaded discussion boards

Publications about threaded discussion boards are sparse. However, we want to point the reader to the publications by Richterich and Gilbert [7], who discuss Reddit's ranking algorithm, Golbeck's investigations about trust on Slashdot [9], as well as the methods discussed by Bouguessa et al. [4] to identify authoritative users in online communities.

2.4 Analysis of scientometrics

In scientometrics some of the most often used measures are indices that measure a scientist's importance. The Hirsch index is a popular example. It is defined as follows: "[…] the index h [is] defined as the number of papers with citation number ≥h, as a useful index to characterize the scientific output of a researcher." [12] A scientist who published many papers that are only cited once may have the same h-index as a scientist who only published one paper that was cited often. Variations of the h-index are the g-index by Egghe [6], or the i10 index provided by Google Scholar [3].

Other investigations were conducted by Kas et al. [14], who applied centrality measures to a scientific database, and Li et al., who identified influential scientists on an academic social media platform [15].

To summarize, most of the related work uses classical data-driven approaches that consist of extracting relevant features from collected social media data, identifying a ground truth value that is assumed to correlate with the investigated characteristic or manually labeling it directly and employing supervised machine learning algorithms for predicting the characteristic.

3 CONTRIBUTIVE SOCIAL CAPITAL SYSTEM

As explained in the introduction, in every interaction we judge our counterpart based on verbal and non-verbal communication. This assessment usually takes place automatically in every-day interactions and may be informed by past experiences and the opinion of others [23]. As a step towards increased CSC transparency, we propose a system imitating this process. As explained before, it consists of three pillars / three main sources of information: observed interactions in OSNEM, accumulation of CSC in social capital markets, and real-life expertise attested by certifications and endorsements (see figure 1). The focus of this paper lies on the first pillar, the

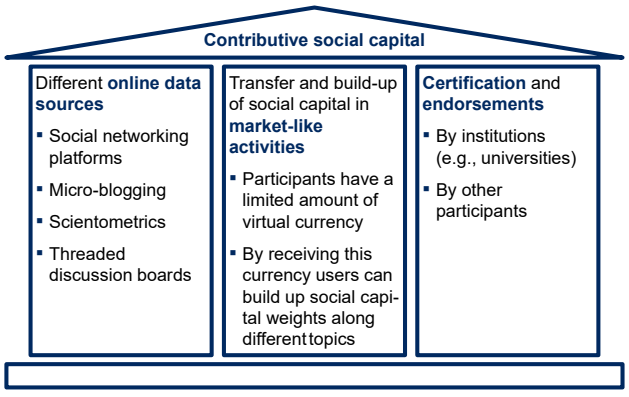

Figure 1: The three pillars of the CSC system. This article focuses on the first pillar.

extraction of CSC from online data sources. Therefore, we give an overview of the whole system and describe the social network and content analysis and our experiment in more detail in the following section.

As a measure of a person's contributive social capital we introduce the term **Contributive Social Capital Weight (**$CSCW_i^t$**)** of a person i in topic t. The first pillar of the system leverages the vast amount of data that is available from online data sources. The first step is to identify relevant data sources. Contributive social capital focuses on the value a person adds to their social network in the form of knowledge shared and help given via interactions with other network participants. Therefore, the most relevant data sources focus on interactions, rather than personal information. Five data sources that fulfill these requirements are:

- social networking platforms (like Facebook),
- micro-blogging (like Twitter),
- threaded discussion boards (like Quora or Reddit),
- scientometrics,
- and direct communication (like Email).

Direct communication in the form of Email or Whatsapp messages is usually private, therefore we do not consider it in this publication. The other four sources can be, to some extent, publicly accessed and investigated for CSC assessment. The interactions of each user in a platform associated with a data source can be described in terms of features. With these features and a ground truth value that reflects the user's CSC, one can investigate whether CSC can be extracted with machine learning algorithms. Similar systems have been used for the assessment of influence as shown by Rao et al. [19] The CSC value can be divided along the topics of a user's interest to determine topic-sensitive $CSCW_i^t$ values. Identifying these topics could, e.g., be achieved with topic modeling [2]. When assigning a CSC score to a topic one has to regard that users may talk authoritatively about matters outside their area of expertise.

Another way to infer CSC that directly assigns topics to the $CSCW_i^t$, is the use of a market system in which every participant has a certain amount of virtual currency that we call social capital currency (SCC). The SCC is either distributed when registering to the system or as a monthly payment similar to basic income. Market participants can transfer their currency freely. One can pay for information or services, thank others for good contributions on social media platforms, or acknowledge beneficial social behavior in general. When making a transaction, the sender specifies the topic of the recipient's social behavior that motivated the transaction. The recipient's $CSCW_{recipient}^{topic}$ increases in proportion to the received currency SCC. After person A transfers ΔSCC to person B in topic i, A's updated SCC'_A and $CSCW'_A$ are:

$$SCC'_A = SCC_A - \Delta SCC \tag{1}$$

$$CSCW_A^{i'} = CSCW_A^i \tag{2}$$

For the recipient B, SCC'_B and $CSCW'_B$ change as follows:

$$SCC'_B = SCC_B + \Delta SCC \tag{3}$$

$$CSCW_B^{i'} = CSCW_B^i + \alpha \cdot CSCW_A^i \cdot \Delta SCC \tag{4}$$

Two terms contribute to B's new $CSCW'_B$:

- $CSCW_B^i$ is B's CSCW before the transaction.
- $\alpha \cdot CSCW_A^i \cdot \Delta SCC$ is the increase that is influenced by a factor α, A's $CSCW_A^i$, and the amount of transferred capital ΔSCC.

B's weight was included to achieve a PageRank-like effect [17] that takes the social capital of the sender into account. The term ΔSCC assigns more weight to larger, i.e. more important, transactions. The nature of the factor α needs to be determined in future research as a compromise between avoiding inflation of CSCW and preserving the effect of small transfers.

The third pillar includes real-world knowledge as reflected by endorsements and certifications. Certifications are issued by institutions, companies, or governments and allow participants to replicate real-world contributive social capital inside the system. Examples for such certifications are degrees by universities or the completion of online courses. The amount by which the CSCW is increased should depend on three factors:

- amount of time required to obtain the degree,
- skill required to obtain the degree,
- CSCW of users with comparable endorsements (to provide a frame of reference).

The endorsement by others is a similar process that gives the chance to replicate real-life CSC. In this case the CSCW increase should depend on the CSCW of the endorser.

The extend to which this system can bring transparency in online interactions needs to be investigated in several practical experiments. In this paper we analyze the extraction from online social networking. It is important to keep in mind that such research may present a potential thread to a user's privacy as well as to keep in mind the ethical implications of such a system.

There are no publicly available datasets that have been annotated with a ground truth suitable for the encompassing analysis of CSC. Therefore, we conducted an experiment with the goal of collecting network interaction data and obtaining social capital ground truth from the participants.

4 BUILDING A SOCIAL NETWORK DATASET

The experiment was conducted within the practical part of a lecture on social computing in the summer term 2017 at Technical University of Munich. Participation was voluntary but students of the course were encouraged to do so, as they would be using anonymized excerpts of the data in the exercises that accompanied the course. This was generally perceived as more interesting than analyzing artificial social networks. Of over 400 students who took the course, 242 registered to the system and for 165 we collected at least one ground truth assessment by others. The networking platform was based on Elgg[1], an open source framework for creating custom social networking platforms. Users were provided with a functionality similar to Facebook and Twitter. They could create profiles with pictures, follow one another, write posts, or comment on own or other people's posts. They could also "like" posts and comments and send private messages.

Students could contribute to the social networking platform during a timespan of nine weeks in the middle of the semester. During this time 244 posts, 2868 comments, 1930 following relationships,

[1] https://elgg.org/

and 3651 likes were created by the 165 participants. Users were free to write about what they wanted. To encourage discussions, different conversation starters in controversial topics were given in the lecture: populism in politics, living in Munich, and healthy food and sustainability. All three topics were actively discussed by the users. The authors of this paper did not take part in the experiment and did not review or comment on any discussions in order not to influence the dynamic of the network.

The observed behavior was similar to what one might see on platforms like Facebook or Twitter. Some people discussed current affairs, others posted funny content, memes, and sometimes spam-like messages or advertisements (e.g., for university events that they organized).

4.1 Ground truth assessment

For the ground truth assessment, the students were asked to answer a questionnaire about other students in the course. They were presented with a list of all students who registered and could select as many as they felt confident to assess. There was a total number of 539 assessments for 165 students, an average of 3.3 assessments per person. Only these 165 people were considered for further analysis.

The questionnaire consisted of 8 questions, each associated with one of the three hypothesized contributing factors of CSC: competence, trustworthiness, and social responsibility. All assessments were made on a scale with 100 unmarked steps.

- The competence assessment should be related to the knowledge and expertise a person demonstrated in the network. Therefore, we asked for direct assessments in the three topics that were given as discussion starters during the experiment. For these questions the left side (0) of the scale was labeled "no experience at all", the right end (100) "extremely knowledgeable".
- The trust assessment was supposed to assess to what degree the individual was trusted by others. Three questions were used that were inspired by the research by Jones et al. on diagnosing trust [13]. They elicited with an overall assessment of trust, the belief that the other person was concerned with the other's welfare, and finally the feeling to what extent the person is fair and honest.
- The third part of CSC, the social responsibility, was assessed with two questions. The first asked about the environmental friendliness of the person, the second about their level of social support and engagement.

The eight questions provided the participants with a multi-faceted way to assess their counterpart, with the individual characteristics being easier to assess than contributive social capital directly. The full questionnaire is given in appendix A. A single CSC value per person was calculated by averaging over all values. This was used as ground truth for the following analysis. The mean CSC value was 64.0 with a standard deviation of 11.5, minimum value was 29.8, maximum value 94.8. The distribution is visualized in figure 2.

4.2 Demographics of the participants

76.4% of the 165 students are male, 23.6% female. The average age is 23.2 years. 35.2% were between 18–21, 43.0% between 22–25, 15.8% between 26–29, and 3.6% between 30–35. 2.4% decided not to disclose their age. The nationality is mainly German (64.2%), 8.5% are from India, 2.4% from Turkey, and the remaining 24.8% from 27 other countries.

Feature	Mean	σ	Min	Max
Posts	1.5	3.0	0	24
Comments	17.4	39.6	0	47
Liked posts (active)	6.2	8.1	0	47
Liked comments (active)	15.9	49.1	0	581
Liked posts (passive)	4.3	7.3	0	34
Liked comments (passive)	12.7	35.1	0	353
Comment responses to posts	16.4	30.5	0	176
Messages sent	4.2	30.3	0	384
Messages received	3.1	6.6	0	68
Followers	11.7	12.0	0	104
Friends	13.6	31.4	0	347
Characters posts	350.5	734.8	0	5331
Characters comments	1790.5	3578.8	0	34696
Characters messages	348.5	2458.9	0	30128

Table 1: Collected features from the social networking platform. The mean count per person is given, as well as the standard deviation and minimum and maximum values.

4.3 Contributions to the Network

The average number of contributions to the OSN by these 165 people is shown in table 1. It lists the respective features, their mean value (e.g., number of posts per person), the standard deviation, as well as the minimum and maximum values.

All collected features roughly follow a power-law distribution, i.e. a small number of people is responsible for most of the contributions. This is visualized in figures 3, 4, and 5 for the number of written comments, followers, and the number of likes received on comments. This is in line with what we see in larger networks [8, 20]. The ground truth roughly follows a normal distribution, as can be seen in figure 2. This is to be expected from averaging over a large quantity of evaluations.

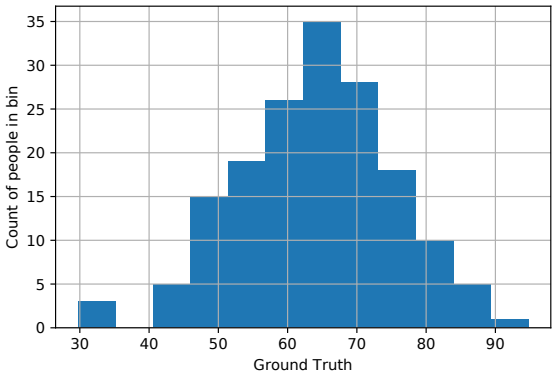

Figure 2: Distribution of ground truth values

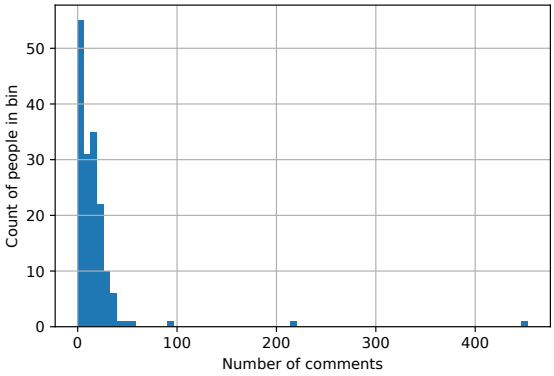

Figure 3: Histogram of comment contributions to the network

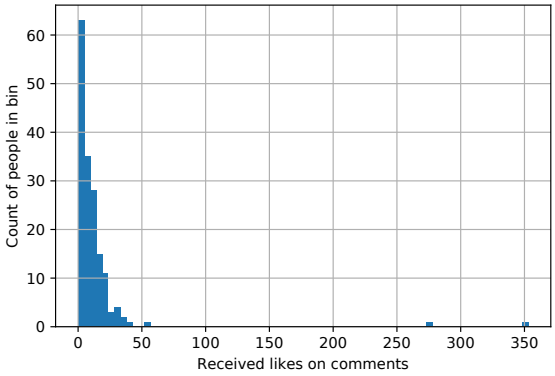

Figure 5: Histogram of number of likes participants in the network received on their comments

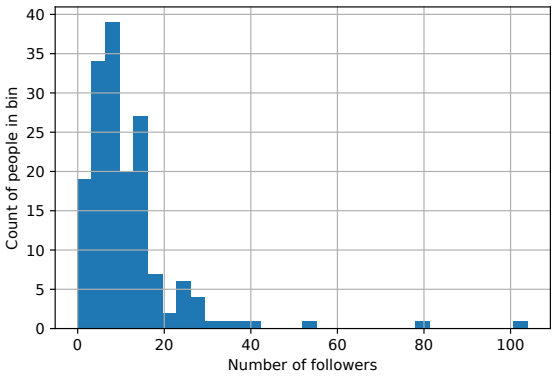

Figure 4: Histogram of number of followers

4.4 Potential shortcomings of the study

There are several potential shortcomings of the dataset that need to be mentioned.

- The majority of contributors are university educated, male students between 20 and 30 years of age. This is not representative of the total population, even though, the users of social networking platforms are predominantly below 35 and more often male [24].
- The sample size of 165 is relatively small.
- The cross-over of data collection and university lecture might have led to a bias. The participation in the network was voluntary and did not influence the grade in any way. However, we can not exclude that some students only participated or contributed in a certain way because they hoped to give a good impression. We tried to counter this bias with open communication during the experiment.
- A similar bias is possible regarding the ground truth assessment. The assessments were confidential and were never shown to the users. However, there might be a positive bias because students might not have wanted to assess others negatively. This kind of bias can be expected in all experiments of this type.
- The time frame of nine weeks is short compared to other OSNEM that are running for years.

All these shortcomings need to be kept in mind when interpreting the results of the study.

5 ANALYSIS OF CSC IN THE DATASET

The main purpose of our analysis was to investigate the research question "can a person's contributive social capital be approximated based on their interactions in a social networking platform?". For this purpose we ran two different analyses on the whole dataset and an active user subset.

5.1 Prediction and correlation with the whole dataset

Two different methods were used to test the hypothesis. The first evaluation was to predict contributive social capital scores based on network activity related features (see table 1) in comparison to a baseline estimator. The group of users was ranked according to their predicted CSC score and then compared to a ranking based on the ground truth.

5.1.1 Prediction of CSC scores based on network features. Several different algorithms were used for the evaluation: linear regression (with and without regularization directly using the features listed in table 1), as well as regression with a decision tree, a random forest, and a neural network. We used 10 fold cross validation for all algorithms. The neural network had 200 neurons in one hidden layer and a logistic sigmoid function as activation function for the hidden layer. The random forest regressors had ten trees and no restrictions on the maximum depth of the tree. To evaluate the result, the mean average error of each model's predictions (mean difference between predicted social capital score and the ground truth value) was compared to a baseline predictor that always predicts the mean ground truth of the training data. The results

Algorithm	Mean absolute error	Improvement
Baseline	9.11	–
Linear regression	9.03	0.8%
Linear regression with regularization	8.73	4.2%
Decision tree	8.45	7.2%
Random forest	7.57	16.9%
Neural network	8.87	2.6%

Table 2: Performance of the different algorithms compared to a baseline predictor for all 165 users. The improvement indicates by how much the algorithm outperforms the baseline.

Algorithm	Pearson	Spearman
Linear regression	0.24 (0.0019)	0.44 (< 0.0001)
Linear regression with regularization	0.29 (0.0001)	0.46 (<0.0001)
Decision tree	0.44 (< 0.0001)	0.41 (< 0.0001)
Random forest	0.42 (< 0.0001)	0.41 (< 0.0001)
Neural network	0.29 (0.0002)	0.29 (0.0002)

Table 3: Pearson and Spearman correlation of the respective algorithm between the predicted ranking and the ground truth ranking. The first value is the correlation, the value in brackets the p-value.

are summarized in table 2. The best result is achieved with random forest regression, which performs almost 17 percent better than the baseline predictor. This is followed by a decision tree with depth four. Linear regression with Lasso regularization, the neural network, and linear regression are only marginally better than the baseline predictor.

5.1.2 Ranking of people based on their CSC. For the ranking task we used the same algorithms to predict a CSC score for each user. All participants were then ranked according to the predicted value. The correlation between this ranking and a ranking with the ground truth was used to evaluate the goodness of the prediction. The results are summarized in table 3. For all algorithms we can observe a weak to moderate positive correlation. The p-value indicates a statistical significance at the 0.01 level for all algorithms. The largest Pearson correlation was achieved with the algorithms decision tree ($r = 0.44$) and random forest ($r = 0.42$) that both achieved the second best and best improvements in the previous analysis. The highest Spearman correlation was achieved with regularized linear regression ($\rho = 0.46$). Decision tree and random forest regression also showed a moderate positive correlation ($\rho = 0.41$ and $\rho = 0.41$). The case of linear regression and neural network regression is particularly interesting. Both algorithms only showed marginal improvements in the first analysis. In the ranking they demonstrated a weak positive correlation of $r = 0.24$ for linear regression and $r = 0.29$ for the neural network. This might indicate that it is easier for algorithms to rank people according to their CSC than to predict a concrete value.

Algorithm	Mean absolute error	Improvement
Baseline	9.06	–
Linear regression	9.86	-8.8%
Linear regression with regularization	8.98	0.9%
Decision tree	7.94	12.4%
Random forest	7.20	20.6%
Neural network	7.27	19.7%

Table 4: Performance of the different algorithms compared to a baseline predictor for the subset of 139 active users. The improvement indicates by how much the algorithm outperforms the baseline.

Algorithm	Pearson	Spearman
Linear regression	0.22 (0.0079)	0.43 (<0.0001)
Linear regression with regularization	0.34 (<0.0001)	0.46 (<0.0001)
Decision tree	0.53 (<0.0001)	0.46 (<0.0001)
Random forest	0.59 <0.0001)	0.49 (<0.0001)
Neural network	0.63 (<0.0001)	0.47 (<0.0001)

Table 5: Pearson and Spearman correlation of the respective algorithm between the predicted ranking and the ground truth ranking. The first value is the correlation, the value in brackets the p-value.

5.2 Prediction and correlation with active user subset

Some of the students in the dataset of 165 participants contributed very little to the social network. Therefore, we ran a second analysis with only active members to investigate potential differences. We performed the same two evaluations as in the previous subsection, this time only with users who wrote at least one post or comment and who did befriend at least one other user. This led to a dataset of 139 active participants.

5.2.1 Prediction of CSC scores based on network features. For the group of active participants the average ground truth CSC value was 65.0 and therefore marginally higher than the whole group's value of 64.0. All parameter settings of the employed algorithms were the same. The results of the prediction are summarized in table 4. The result is considerably better than for the whole dataset. Both, random forest regression (20.6%) and neural network regression (19.7%) led to an improvement of about 20 percent. The decision tree also demonstrated slight improvements (12.4%). The simple linear regression algorithm performed worse than on the full dataset and has a larger mean error than the baseline predictor. With Lasso regularization it performed only marginally better. This indicates that the relation between the network features and the ground truth CSC values can less well be described by a linear function when inactive users are excluded.

5.2.2 Ranking of people based on their CSC. The ranking of students according to their predicted CSC scores shows a similar result, as can be seen in table 5. All algorithms ranked the active

users in a way that their CSC rank correlates positively with the ground truth value. The highest Pearson correlation ($r = 0.63$) was achieved with the neural network. Random forest and decision tree ranking also achieve values larger than 0.5 ($r = 0.59$ and $r = 0.53$). The relatively weak correlation of $r = 0.22$ that was achieved with the linear regression indicates once more that the relation between ground truth and features is most likely not purely linear. When using Spearman correlation the best results are also achieved with random forest ($\rho = 0.49$) and the neural network ($\rho = 0.47$). However, all correlation values lie much closer together.

5.3 Discussion of Results

The experiment with the active user subset yielded an improvement of about 20% over the baseline predictor when trying to predict CSC values, and a Pearson correlation value on the ranked lists of up to 0.6. These values indicate that it might be possible to predict contributive social capital from features present in social networking platforms. However, the values are merely small to moderate and might additionally be biased due to the shortcomings of the experiment, as we discussed in section 4.4. Therefore, it is important to use caution until the findings are supported by large-scale experiments with data from existing social networking platforms. We are not aware of any similar experiments for the analysis of contributive social capital in social networks, therefore it is hard to compare the values of our results. Nevertheless, one can make several other observations:

- The best algorithms for the prediction of contributive social capital are random forest (best results for both datasets) and the neural network that performed also well on the active user network.
- For the ranking of users according to their CSC score on the whole dataset we get the best result with decision tree and random forest regression (Pearson correlation), or linear regression with and without regularization (Spearman correlation). On the active user dataset, the best results are given by the neural network and random forest (Pearson correlation), respectively random forest and neural network (Spearman correlation).
- It seems that a better result can be achieved by ranking the users than by predicting concrete CSC values.
- The quality of the analysis was increased by excluding inactive users. This led to an improvement of 20.6% compared to 16.9% for the prediction task, and a Pearson correlation of 0.63 as opposed to 0.44.

The importance of the different features for the prediction can also be investigated. As random forest regression generally led to the best results, we chose this algorithm to discuss their relative importance. As shown in table 6, the five most important features are the number of likes a user received on their comments, the number of comments written by a user, the number of characters used in written posts, the number of comments that a post inspired, and the number of followers a user has. These five features account for over 70% of the importance for the model (increase in prediction error when leaving out the feature [5]). Three of the features are indicators for the support a user gets from their surrounding network, namely the received likes, the inspired responses, and the number of followers. The number of comments as well as the length of the posts are signs for the involvement of a user. Other ways of participating, like following others or liking the posts or comments of other users are less important for the prediction of CSC. The number of posts and the likes received on them are on the very bottom of the list. This might be due to their relatively low number (on average 1.5 for posts and 4.3 for likes on posts per person).

Feature	Importance	Cumulative
Liked comments (passive)	24.9%	24.9%
Comments	15.0%	39.9%
Characters posts	14.2%	54.1%
Comment responses to posts	11.0%	65.1%
Followers	7.5%	72.6%
Characters comments	5.2%	77.7%
Liked comments (active)	4.9%	82.7%
Friends	4.4%	87.0%
Characters messages	2.9%	89.9%
Liked posts (active)	2.5%	92.4%
Messages received	2.3%	94.7%
Messages sent	2.1%	96.8%
Posts	2.1%	98.9%
Liked posts (passive)	1.1%	100.0%

Table 6: Relative and cumulative importance of the different features for random forest regression on the active user dataset

6 SUMMARY AND OUTLOOK

In this paper we described contributive social capital and presented our vision for a system to determine it based on social networking activities, market interactions, and endorsements. Furthermore, we described an experiment to investigate whether CSC is present in social networking platforms and can be detected with machine learning.

The experiment led to a dataset of 165 participants with network activity and ground truth values that were assessed through questionnaires. We ran two types of analyses on the whole network and an active user subset. The first investigation was to predict CSC scores based on the network activity and compare them to a simple baseline predictor. The second was to rank people according to their predicted CSC values and correlate the result to the true ranking. There was a small improvement regarding the prediction and a moderate correlation between both lists. However, this is just a piece of evidence for the predictability of CSC in OSNEM and not a definite proof due to the limitations of the experiment.

To address the shortcomings of our experiment, we suggest to carry out further research in larger social networking platforms. In future work we investigate whether there is a connection between CSC and market interactions with a virtual currency.

A GROUND TRUTH ASSESSMENT QUESTIONS

The participants could assess the competence, trustworthiness, and social responsibility of other known users with the help of eight questions. All assessments were made on a scale from 0 to 100.

A.1 Competence assessment

Please evaluate this person's competence (mixture of knowledge and expertise) in the following three fields:

- Populism in politics (e.g., Trump's wall to Mexico, refugee crisis in Europe, etc.)
- Living in Munich (e.g., sports and leisure activities, finding affordable living, lectures at TUM, etc.)
- Healthy food and sustainability (e.g., calorie counts, genetically altered nutrition, sustainability, etc.)

A.2 Trust assessment

Please evaluate how much you trust this student.

- What is your general level of trust towards this student?
- To what extent is this person concerned for your welfare – someone who is looking out for you, who would go out of their way to help you, and who would not knowingly do anything to hurt you?
- To what extent is this person fair and honest – do they stick to their word and use sound principles to guide themselves?

A.3 Social responsibility

Please help us understand how environmentally friendly and socially engaged the selected person is.

- Environmental friendliness (e.g., support of environmental protection institutions, sustainable food, waste separation, etc.)
- Social support/engagement (e.g., support of friendly societies, help to other students/friends/strangers, support for elderly family members, etc.)

ACKNOWLEDGMENTS

Thanks a lot to Christian Höfer, who helped with the design of the experiment, to Dennis Assmann who supported the supervision of the experiment, and Valeria Chernenko, who supported the analysis. Additional thanks to all the students who participated.

REFERENCES

[1] I. Anger and C. Kittl. 2011. Measuring influence on Twitter. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11* (2011), 1. https://doi.org/10.1145/2024288.2024326

[2] D. M. Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[3] Google Scholar Blog. 2011. Google Scholar Citations Open To All. (2011). Retrieved January 31, 2018 from https://scholar.googleblog.com/2011/11/google-scholar-citations-open-to-all.html

[4] M. Bouguessa and L. Ben Romdhane. 2015. Identifying Authorities in Online Communities. *Acm Transactions on Intelligent Systems and Technology* 6, 3 (2015), 23. https://doi.org/10.1145/2700481

[5] L. Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] L. Egghe. 2006. Theory and practise of the g-index. *Scientometrics* 69, 1 (2006), 131–152. https://doi.org/10.1007/s11192-006-0144-7

[7] E. Gilbert. 2013. Widespread Underprovision on Reddit. *Proceedings of the 2013 conference on Computer-supported Cooperative Work* (2013), 803–808. https://doi.org/10.1145/2441776.2441866

[8] M. Gjoka, M. Kurant, C. T. Butts, and Athina Markopoulou. 2009. A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. *CoRR* abs/0906.0060 (2009). arXiv:0906.0060 http://arxiv.org/abs/0906.0060

[9] J. Golbeck. 2009. *Computing with Social Trust.* 287–311 pages. https://doi.org/10.1007/978-1-84800-356-9_11

[10] A. T. Hadgu and R. Jäschke. 2014. Identifying and analyzing researchers on twitter. In *CEUR Workshop Proceedings*, Vol. 1226. 164–165. https://doi.org/10.1145/2615569.2615676

[11] S. Hassan. 2013. Identifying criteria for measuring influence of social media. 10, 1 (2013), 86–91.

[12] J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A* 102, 46 (2005), 16569–16572. https://doi.org/10.1073/pnas.0507655102 arXiv:physics/0509048

[13] S. L. Jones and P. Pradhan Shah. 2015. Diagnosing the Locus of Trust: A Temporal Perspective for Trustor, Trustee, and Dyadic Influences on Perceived Trustworthiness. *Journal of Applied Psychology* (9 2015). https://doi.org/10.1037/apl0000041

[14] M. Kas, K. M. Carley, and L. R. Carley. 2012. Trends in science networks: understanding structures and statistics of scientific networks. *Social Network Analysis and Mining* 2, 2 (2012), 169–187. https://doi.org/10.1007/s13278-011-0044-6

[15] N. Li and D. Gillet. 2013. Identifying influential scholars in academic social media platforms. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13* (2013), 608–614. https://doi.org/10.1145/2492517.2492614

[16] N. Lin. 2002. *Social Capital: A Theory of Social Structure and Action, 2001. 278 pp.* Cambridge University Press.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/ Previous number = SIDL-WP-1999-0120.

[18] R. D. Putnam. 1995. Bowling alone: America's declining social capital. *Journal of democracy* 6 (1995), 65–65.

[19] A. Rao, N. Spasojevic, Z. Li, and T. DSouza. 2015. Klout Score: Measuring Influence Across Multiple Social Networks. (2015), 8. arXiv:1510.08487 http://arxiv.org/abs/1510.08487

[20] T. Rastogi. 2016. A Power Law Approach to Estimating Fake Social Network Accounts. *CoRR* abs/1605.07984 (2016). arXiv:1605.07984 http://arxiv.org/abs/1605.07984

[21] L.J. Robison, A.A. Schmid, and M.E. Siles. 2002. Is Social Capital Really Capital? *Review of Social Economy* 60, 1 (2002), 1–21. https://doi.org/10.1080/00346760110127074 arXiv:https://doi.org/10.1080/00346760110127074

[22] S. Schams and G. Groh. 2018. Social Capital Extraction from Different Types of Online Data Sources. *submitted* (2018).

[23] B. Tracy. 2004. *The Psychology of Selling.* Thomas Nelson. https://books.google.de/books?id=8np-oAEACAAJ

[24] TUG 2017. Distribution of Facebook users worldwide as of January 2017, by age and gender. (2017). Retrieved January 28, 2018 from https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/

[25] J. Weng, E. Lim, J. Jiang, and Q. He. 2010. Twitterrank: Finding topic-sensitive influential Twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (2010), 261–270. https://doi.org/10.1145/1718487.1718520