Towards Quantifying Behaviour in Social Crowdsourcing Communities

Khobaib Zaamout, Ken Barker
{kzaamout,kbarker}@ucalgary.ca
Department of Computer Science - 602 ICT Building - University of Calgary
2500 University Dr NW. Calgary, AB. Canada T2N 1N4

Abstract

We analyze crowdsourcing communities by creating a detailed process for quantifying individual behaviour in online environments. The key feature of our communities is their social interactions so we call these *social crowdsourcing communities* (SCC). First, we derive factors based on actions captured about textual contributions. We interpret and name these factors. Then we demonstrate their utility in predicting the quality of team contributions. We capture the actions of members using measurable variables and perform factor analysis on these to produce factors of behaviour in SCCs (i.e. dimensions of behaviour). We derive factor scores for each member. An abstract notion of teams is used that is based on the social interactions. Team scores are then determined by the aggregation of the individual factor scores. The relationship between the team-level factor scores and the quality of contributions made by each team are then used as a proxy for team effectiveness. We found that member behaviour has three dimensions/factors: *Impact*, *Activity*, *Policing/Rowdiness* and there is a linear relationship between a team's contribution quality and their *Impact* scores. We also found a moderate negative linear relationship between the smallest *Activity* scores in each team with the quality of their individual contributions. This shows that teams that produce higher quality contributions tend to have higher total and maximum *Impact* score with lower levels of *Activity*. Thus, we demonstrate that properly aggregated behavioural factors can predict the quality of team-level contributions.

The way we behave says a lot about many aspects of our lives, such as our personalities (McCrae and John 1998; McCrae 2010), personal values and religiousness (Roccas et al. 2002; Saroglou 2002), performance at our jobs (Salgado 1997), and performance working in teams (Mount, Barrick, and Stewart 1998; Lim and Ployhart 2004). There is a clear relationship between the way we behave and other aspects of our lives. Anthropologists, psychologists, sociologists, and many other social scientists have been investigating this relationship for centuries. The Internet gave rise to an opportunity for computer and data scientists to contribute to this research. They analyze the behaviour of individuals in online environments by finding patterns that are indicative of aspects of individuals' life, such as personality (Kalish and Robins 2006; Krause, James, and Croft 2010; Wehrli 2008; Quercia et al. 2011; Gosling et al. 2011; Golbeck, Robles, and Turner 2011; Kosinski et al. 2014; Kern et al. 2014), ability to work in teams (Yang and Tang 2004; Janhonen and Johanson 2011; Aldrich and Kim 2007), interest and expertise in specific areas (Zhang, Ackerman, and Adamic 2007), strength of relationships (Gilbert and Karahalios 2009), and surfing behaviour and buying habits (Awad, Khan, and Thuraisingham 2008; Hansen, Jensen, and Solgaard 2004). Most researchers in this field focus on quantifying and relating aspects of online behaviour to online (or offline) aspects of individuals. This popular approach provides interesting insights that tend to be specific in nature. Our approach draws upon more detailed data in an attempt to acquire more holistic views of member behaviours.

We perform our analysis on a Social Crowdsourcing Community (SCC). SCCs are crowdsourcing communities where members have access to social features that allows them to interact with each other. Interactions among members of these communities produce digital footprints like that of social networking platform[1].

This paper answers three fundamental questions: How can we quantify behaviour in an online environment? What are the dimensions, *i.e.* factors, of the quantified behaviour one such online environment, namely SCC? How do these factors relate to other aspects of individuals' behaviour? We establish a standard process that helps in answering these questions for our and other online environments. Specifically, we present a process for quantifying behaviour in SCC by first defining what we mean by behaviour in this specific environment, present the process to arrive at its dimensions, *i.e.* factors, derive the factors, interpret their meanings, and show their utility in predicting a team's quality of contribution.

We achieve three outcomes. First, we present a standard process to quantifying behaviour in online environments. This process is sufficiently generic that it applies to future studies, which will help develop an understanding of behaviour in various online environments in a systemic and reliable way (see Section 2.2). Second, we derive, name, and interpret the factors of behaviour in an SCC (see Section 2.3). Finally, we show the presence of relationship be-

[1] (Zaamout and Barker 2017) provide detailed descriptions of SCCs consistent with our definition.

tween these factors and member actions (see Section 3) and show the utility of these factors in predicting teams' quality of contributions (see Section 4).

These outcomes are important because they expand our understanding of human behaviour, help improve product or service recommendations, and help find potential threats or problematic individuals in social networks. It is also important to SCCs where individuals collaborate to achieve a common goal (Zaamout and Barker 2017). This is because quantifying behaviour in SCCs and understanding its effect on aspects, such as team work, team cohesion, quality of contributions, *etc*., is critical to the success of social crowdsourcing communities.

This paper is organized as follows. Section 1 describes the data and shows demographic information of members of our SCC. Section 2 describes the process to arrive at the factors of behaviours concerning textual contributions. Section 3 demonstrates the relationship between our factors and other aspects of behaviour. Section 4 demonstrates the utility of our factors in predicting the quality of team-level contributions. Section 5 summarizes our contributions and highlights future research.

1 Data

We use an anonymized back-end database of an active and diverse SCC from Chaordix[2]. This SCC is a private community of brand loyalists of a major mobile phone company[3]. Members of this community take part in various activities related to mobile devices in return for exclusive information from within the company and frequent giveaways of desired company branded merchandise. This database has member demographic information, and their actions. From this database, we produce several datasets that we use to answer our core questions.

Member demographic data consists of identification number, age, sex, country of citizenship, number of spoken languages, occupation, and level of education for 7,524 members. 89% of members are male and 10% are female. These members are from 119 countries on six continents (see Table 1). The average age in this community is 31 years with 9 years standard deviation and over 73% of members within one standard deviation.

Table 1: Distribution of members on continents

Continent	Percentage
Africa	0.9%
Asia	18.9%
Australia (Oceania)	2.0%
Europe	33.0%
North America	44.4%
South America	0.3%
Unspecified	0.6%

Table 2: Education level of members

Level of education	Percentage
Unspecified	1.5%
Less than High School	1.4%
High School or equivalent	16.0%
Some college (but no degree)	26.0%
Associate degree	8.7%
Bachelor degree	32.9%
Graduate degree Masters	12.2%
Graduate degree Doctorate	1.5%

68.1% of members speak English as a primary language and the majority can speak more languages. 83% of members specified their occupation. The largest portion of members is students and the smallest portion of members is stay-at-home parents. Table 2 shows the percentages of the highest level of education achieved by members.

2 What are the Factors of Behaviour?

To answer this question, we produce a dataset (see Section 2.1) that quantifies all relevant actions that members perform on this SCC. We do factor analysis to derive factors of behaviour. We then analyze these factors, name them, and interpret their meaning.

2.1 Dataset

We produce a dataset that consists of members and base variables. Base variables hold simple counts of actions for each member. We do a Factor Analysis (FA) on this dataset to derive *factors of behaviour* (see Section 2.2).

We now consider why we require factor analysis instead of using this dataset directly. We define behaviour as the set of all measurable actions that members can perform on our SCC through features offered by the SCC website/platform, such as following other members, creating posts, commenting, voting, *etc*. Thus, this definition implies the dataset quantifies behaviour. However, it consists of elaborate and platform-dependent variables, which prevent us from producing any generalized conclusions. FA yields factors, *i.e.* dimensions, that are latent, *i.e.* unmeasured, and generalized, *i.e.* applicable to more than a single SCC platform.

Why are base variables adequate? Why do we not use derived variables? For two reasons: First, derived variables, such as rates, inter-arrival times, and regularities are more specific and complex than base variables, which make them harder to understand and make interpreting their relationships with other variables difficult. They also pose a challenge in FA because they tend to cross-load[4], and to be redundant[5](Hinkin, Tracey, and Enz 1997). Second, the number of derived variables is extremely large so we only consider base variables.

We focus on artifacts and actions concerning one type of textual contributions, namely submissions. Submissions are

[2] www.chaordix.com

[3] The community's name must be kept confidential to be consistent with our access agreement.

[4] Variables that have a loading value ≤ 0.3 on multiple factors.

[5] Variables derived from multiple base variables tend to highly correlate with the base variables, which is problematic for FA.

the highest level of written content that members produce as a direct response to a moderator's newly initiated discussion topic or question. Members can remove their own submissions and can tag their submissions with descriptive tags. These tags help members and data analysts to find and organize submissions. Moderators can remove and tag any member's submissions. Any member can vote, follow, and comment on a submission and any other member can reply to any of the comments. A reply to a comment creates a comment thread. Comment threads are at most two levels deep. Members can also flag submissions as "inappropriate", which notify moderators to review and potentially censor the flagged submission. Moderators can censor, *i.e.* delete, submissions and can flag them as significant contributions. They also can pin - and unpin - submissions to display them in a highly visible area. Pins indicate that the submission is significant.

We capture actions that a member performs, such as creating submissions and voting on others' comments, and actions that a member receives on written contributions, such as comments received from others on their own submissions or comments and votes received on submissions or comments. From these actions, we calculate simple quantitative variables that have clear meaning and mapping to the SCC actions and avoid any complex variables that address several aspects of behaviour. All base variables have a minimum value of zero and an unbound maximum value.

Let A denote actions that our crowdsourcing community offers (see Table 3 for details). Let M be the set of members, $M = \{m_0, m_1, \ \ldots, m_n\}$. We create a dataset[6], MA, as a matrix that holds values for each member action combination. That is, $MA = M \times A$ where each cell, ma_{ij}, contain the total number of times member m_i performed action a_j since joining the SCC. Since members joined the SCC at various times, they have been on the SCC for different periods, and they login and participate at different rates, we must normalize the dataset. Therefore, we divide each member's variables by the number of days this member logged in to the SCC, which gives us the average number of actions this member performed per day. In other words, we normalize our data by dividing each ma_{ij} by the number of days member m_i logged into the SCC, which will produce an average a_j per days logged in. We calculated these variables for all members who logged in three or more times ($5,593$ members).

2.2 Factor Analysis

Factor Analysis (FA) is a term that describes a suite of methods - mathematical procedures and algorithms - that factorize and confirm factorization of a given dataset. FA derives factors - called factor solution - as linear combinations of input variables and offers statistical methods to assess the reliability of the factorization. These factors give a summary of the original dataset, which makes understanding and interpreting relationships easy (Yong and Pearce 2013). In other words, FA simplifies complex measured variables by deriving unmeasured "latent" variables, *i.e.* factors, that affect the measured variables. The output of FA is a variable-factor loadings matrix, which has the loading value each variable has on each factor. Variable loading is the amount that a variable contributes to a factor. Since a variable loads on all factors - *i.e.* factors are linear combinations of variables - these factors can be thought of as groups that variables belong to as defined by the variable loadings matrix. To simplify, we say that a variable belongs to a factor on which it has the highest loading value.

In mathematical terms, we put the MA dataset through FA to yield a new dataset VF[7]. $VF = A \times F$, where A is the set of actions and $F = \{f_0, f_1, \ldots\}$ is the set of factors we arrive at through FA. Thus, vf_{ij} is the loading value variable v_i has on factor f_j.

Subjectivity plays a key role in obtaining and interpreting FA results. This subjectivity is necessary because it allows analysts to use their professional judgment and domain knowledge to tune the analysis process and interpret its results within the context of the research question. FA results are credible if the analyst reports all relevant information about the dataset, the analysis process, and the results to allow others to critique and reproduce them.

There are two steps to performing FA (Froman 2001; Hinkin, Tracey, and Enz 1997): Exploratory FA (EFA) and Confirmatory FA (CFA). EFA is an algorithmic procedure that summarizes a dataset by merging interrelated variables into factors, where interrelated variables load higher on the same factor (Suhr 2006; Froman 2001; Yong and Pearce 2013). CFA is a FA procedure that verifies whether a given dataset supports a specific variable grouping, *i.e.* factorization or factor solution, through various statistical tests (Suhr 2006; Froman 2001; Yong and Pearce 2013). We aim to quantify behaviour in SCC. To the best of our knowledge, no rigorous theoretical treatment of SCCs appears in the literature. Thus, we will not suggest a factorization based on preconceived understandings. Instead, we will let the dataset speak for itself by performing EFA first. We will guide, and document, our EFA procedure based on our understandings and then use CFA to test the reliability of our factorization.

Exploratory Factor Analysis Exploratory Factor Analysis (EFA) consists of six steps: gathering a dataset, choosing an EFA method, preparing the dataset, performing EFA, obtaining and interpreting factorization, and obtaining factor scores.

Gathering a dataset is the accumulation of records, *i.e.* samples or observations, in sufficient quantities and in adequate diversity so analysis is possible and the results are stable, credible, and generalizable. The dataset is sufficient if the ratio of observations to variables surpasses some threshold (Froman 2001; Osborne and Costello 2009). There are many possible threshold values; the most conservative of which is an observations to variables ratio of 30 to 1 (Yong and Pearce 2013). Section 2.1 shows our dataset has 5,593 observations and 19 variables, which exceeds the most conservative ratio. The dataset is an adequate sample if it contains diverse subjects, and if the variables measure various

[6] We use the term create with respect to the dataset to indicate that we are extracting, projecting, or aggregating the real data.

[7] VF stands for variable-factor loading matrix, which contains variable loadings on factors.

Table 3: Variables and their descriptions. None of the variables have univariate or multivariate normality (tested using Kolmogorov-Smirnov goodness of fit test (Henderson 2006) and Henze-Zirkler's Multivariate Normality Test (Korkmaz, Goksuluk, and Zararsiz 2014). * indicates that a variable's distribution does not fit power-law distribution ($p < 0.05$).

Variable Name	**Description**
comment made by me on me submissions	replies a member made to comments received on submissions
*comment made by me on others submissions**	comments a member made on others' submissions
comment others made on me submissions	comments a member received on submission
*submission**	submissions made by a member
submission censors made	censors a moderator made
submission censors received	censors a member received
submission flags made	flags a member made on others' inappropriate submissions
submission flags received	flags a member received on inappropriate submissions
*submission follows made**	follows a member made on others' submissions
submission follows received	follows a member's submissions received
submission pins made	pins a moderator made on others' submissions
submission pins received	pins a member received on submissions
submission significant sets made	times a moderator set others' submissions to significant
submission significant sets received	times a member's submissions were set to significant
*submission tags by me on me**	tags a member made on submissions
submission tags made	tags a moderator made on others' submissions
submission tags received	tags a member received on submissions
submission votes made	total number votes a member made on others' submissions
*submission votes received**	votes a member received on submissions

aspects of the problem. Lack of diversity in the data, *i.e.* homogeneity of the subjects, prevents factors from emerging (Yong and Pearce 2013). Section 1 shows our dataset contains diverse subjects, which makes our findings applicable to a diverse population.

Choosing an EFA method is the process of analyzing the dataset for specific characteristics that help determine a suitable method. The choice between the various EFA methods is complex (Kline 1987; Osborne and Costello 2009). The argument for the preferred approach of EFA by analysts is contentious from various fields including statistics (Kline 1987; Osborne and Costello 2009). While some analysts suggest that all methods produce the same results when presented with "sufficiently" large dataset (Harman 1976), others believe that the differences between the various approaches are significant and suggest different approaches to differently distributed dataset (Osborne and Costello 2009; Fabrigar et al. 1999; Yong and Pearce 2013). We used Maximum Likelihood due to their popularity in the specific *R* packages we used to perform FA.

Preparing the dataset is cleaning and transforming the dataset in the ways needed by the chosen EFA method. We check and remove variables with multicollinearity and singularity (Froman 2001; Yong and Pearce 2013) by using Squared Multiple Correlation (SMC). Otherwise, FA will not be able to run. Variables that have issues with singularity will have a SMC value close to zero and variables that have issues with multicollinearity will have SMC value close to one. Removing a variable from the dataset will affect the SMC scores of the others. Since we aim to keep as many of our variables as possible, we created an iterative process to help us in deciding which variable to remove and when to stop removing variables. This process helps us avoid subjectivity in deciding what SMC scores are "close enough" to singularity or multicollinearity and allow us to maximize the variables that are considered while ensuring the completion of the analysis for a given computing constraint.

The iterative process stops when FA can be run to completion on our dataset. If FA does not run, we run SMC on the dataset. We sort the variables by their SMC scores. Variables that suffer from multicollinearity or singularity will be at either end of the list. We start by removing a variable with multicollinearity problem. Multicollinearity affects groups of variables. Thus, if our dataset has this problem, then two or more variables have it. Multicollinear variables have similar SMC scores. We inspect these variables together as group and decide to remove the variable that is least meaningful to our analysis. If we do not have a multicollinearity problem, we remove the variable with the least SMC score.

After running this process, we removed: *submission pins made*, *submission significant sets made*, *submission significant sets received*, and *submission tags made* because of low SMC scores. The dataset contains no multicollinearity within the variables.

Performing EFA is the process of setting the chosen EFA method's required options, feeding it with the cleaned and transformed dataset, and obtaining results, *i.e.* factors, and factor loadings. Typically, this is an iterative process where the analyst checks the results for any anomalies that may require changing the method's options and running it again. Typically, analysts change two options: the number of factors the method will derive and the rotation method.

Deciding on the number of factors to derive is the most crucial decision the analyst must make. Deriving too many factors or too few has a direct effect on the correctness and interpretability of these factors. There are many methods that help analysts arrive at a number and there are many heuristics that help analysts decide on the quality of a factorization, *i.e.* factor solution. Kaiser's criterion (Kaiser 1960; Kline 1987; Osborne and Costello 2009), scree plot test (Cattell 2012), and parallel analysis (Horn 1965) are popular methods for deciding the number of factors to derive. Kaiser's criterion suggests keeping all factors with an Eigen value greater than one, Scree plot test helps analysts decide the number of factors based on visual cues in the plot, and parallel analysis suggests keeping all factors that have an Eigen value larger than the average Eigen value obtained from a series of random data matrices.

Several heuristics assess the quality and feasibility of a factor solution (Osborne and Costello 2009; Kline 1987; Yong and Pearce 2013). These heuristics concern four areas: weak variable loadings, cross-loading, weak factors, and factor interpretability. A factor has weak variable loadings if the highest variable loading is less than 0.3 (some suggested 0.32 (Tabachnick, Fidell, and Osterlind 2001) for sample sizes larger than or equal to 300). Cross-loading refers to a variable that has a high loading value (> 0.3) on two or more factors. This may be acceptable if both factors have a high correlation with each other. A weak factor is a factor that has fewer than three variables that load highest on it. Some (Froman 2001) suggest that a factor that has two variables that loads highest on it is acceptable if the loading values are high and the two variables are both highly correlated with each other and have low correlation with others. That is, if we change the number of factors and rerun the FA and this factor remains the same, then this is a stable and acceptable factor. Factors interpretability refers of the ease of interpreting the meaning of the factors and the degree of which they make sense. Since the aim of EFA is to produce simple and interpretable factors some argued (Kline 1987) that the emergence of simple structures (factors), *i.e.* factors with few high loadings that are easy to interpret, is a key sign of the success of the FA.

We use the three methods and the factor quality heuristics in an iterative process to arrive at the best factorization. We run all the methods to get an estimate of the number of factors each method suggests. We use the largest estimate as a safe starting point. We run FA and obtain the factors. We inspect the factors and variable loadings for specific characteristics. If these properties exist, we reduce the number of factors and run FA again. If a certain variable proved to be consistently problematic, we remove it from our dataset, and run FA without reducing the number of factors. We avoid removing multiple variables at once. We found that at the end of this process, most methods for estimating the number of factors to derive converge to the same number (see Figure 1).

After running this iterative process, we removed three variables: *submission flags made*, *submission follows made*, and *submission tags received* because they consistently loaded at less than 0.3 on any factor.

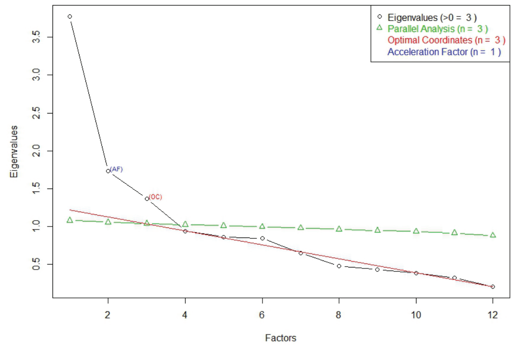

Figure 1: Scree plot test, parallel analysis, and Kaiser's criterion methods all converged to the same number of factors.

The aim of rotation is to simplify the factors by minimizing the number of variables that factors load on while maximizing the loading value of each factor (Osborne and Costello 2009; Yong and Pearce 2013). This improves the factorization fit to the dataset and therefore improves interpretability of the factors. There are two types of rotations, orthogonal and oblique. Orthogonal rotation methods produce factors that are uncorrelated and oblique rotation methods produce factors that are correlated (Osborne and Costello 2009; Yong and Pearce 2013). Some (Guilford 1959) argue that orthogonal rotation methods produce simpler factorizations and are therefore easier to interpret. We believe that human behaviour is rarely partitioned into "neatly packaged units that function independently of one another" and therefore using oblique rotation would produce a more accurate, interpretable, and reproducible outcome (Osborne and Costello 2009; Yong and Pearce 2013; Kline 1987; Cattell 1973). We used a variety of oblique rotation methods and found no significant difference in our results. Therefore, we use oblimin factor rotation.

Obtaining and interpreting factorization refers to analyzing variable loading values in the context of the research topic to give meaning and justification to the factors and to name them. A variable that loads highest on a certain factor belongs to that factor. By reflecting on the variables that belong to each factor, we can arrive at the proper naming and meaning behind it. This is ultimately the goal behind EFA and a key to deciding its success.

Obtaining factor scores refers to the process of deriving factor scores for each subject, *i.e.* SCC member in our dataset, based on a specific factorization. There are several methods to do this (Odum 2011; DiStefano, Zhu, and Mindrila 2009). We tested two of the common methods, *i.e.* Bartlett and Anderson-Rubbin methods (Yong and Pearce 2013), and found that they had no significant impact on outcomes of this paper.

Confirmatory Factor Analysis Confirmatory Factor Analysis (CFA) measures the reliability of our factorization. Given the factorization, and the dataset, CFA quantifies reliability of the factorization by calculating the degree of which this factorization "fit" the dataset. There are many measures that quantify goodness of fit that we can choose (Pedrosa et al. 2016; Hinkin, Tracey, and Enz 1997). We ultimately chose Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) due to their popularity in the specific *R* packages we used to perform FA.

2.3 Factors of Behaviour

We ran the Maximum Likelihood Estimation (MLE) FA method on the dataset MA to obtain factors and obtain variable loadings on these factors, VF (see Section 2.2).

Variable Loadings Factor analysis on our initial dataset yields three factors with noteworthy consistency (see Section 2.3). We name these factors: *Impact*, *Activity*, and *Policing/Rowdiness*. Table 4 shows the three factors along with the variable loadings. *Impact* is a factor that represents the attention a member's submissions receive through comments, votes, follows, and pins from other members of the SCC. Thus, variables concerning received reactions from other members about a member's submissions load highest on the *Impact* factor. This factor is a positive factor because its variables are positive in value and in sentiment. Meaning, the smallest value a member can have in any of these variables is a zero and these variables have no negative sentiments (no vote down, no negative comments, *etc.*). *Activity* is a factor that present the amount of work a member performs in the SCC. Thus, variables concerning number of submissions, tags, commenting, and voting on others' submissions load highest on this factor. *Activity* is a positive factor. *Policing/Rowdiness* is a factor that quantifies actions of flagging and receiving of flags for inappropriate submissions or comments as well as censoring inappropriate submissions or comments. This is a complex factor where higher scores imply higher level of policing activities or rowdiness. Two of the three variables that loads highest on this factor concern rowdiness: *submission censors received* and *submission flags received*. These two variables represent the number of inappropriate submissions and comments a member made, which other members flagged and moderators censored. These variables are positive in values but negative in sentiment. Meaning, the smallest value a member can have is zero and having higher values in this factor means a higher number of flags or flagged and censored content.

Factor Correlations We believe that factors *Impact* and *Activity* should correlate with each other because a member with a high score on *Activity* factor reads and reacts to others' textual contributions (as the interpretation of *Activity* in Section 2.3 factor states). We believe that this would give the member an advantage for creating new original content. However, our results shows that *Impact* and *Activity* are not correlated (-0.125[8]). This means that a member's level of impact is not relevant to the amount of content (s)he goes through, reacts to, or produces. This also implies that reactions in SCCs are not reciprocal, which explains the low level of reciprocity in SCC interaction networks (Zaamout and Barker 2017). We expect that *Activity* correlates with *Policing*, since active members are more likely to see and report bad behaviour and given their tendency to produce a large number of content, are more likely to produce content that others find negative. Our results supports this hypothesis (0.471). Factors *Impact* and *Policing/Rowdiness* are not correlated (-0.0756), which suggests that they are independent of one another. This means that impact scores are irrelevant to policing activities or rowdy behaviour.

Table 4: Factors and loadings. We bolded highest loading values for each variable and underlined cross-loadings.

Variables	Impact	Activity	Policing
comment others made on me submissions	**0.86**	0.14	0.02
submission follows received	**0.81**	-0.19	-0.02
submission votes received	**0.50**	0.37	0.03
submission pins received	**0.31**	-0.16	0.031
comment made by me on others submissions	-0.02	**0.76**	-0.01
submission votes made	-0.073	**0.67**	-0.02
submission	0.20	**0.50**	0.08
comment made by me on me submissions	0.40	**0.48**	-0.01
submission tags by me on me	0.01	**0.47**	0.043
submission censors received	-0.01	-0.02	**1.00**
submission censors made	0.04	0.02	**0.52**
submission flags received	-0.01	0.05	**0.49**

Factoring Reliability Our factorization has a 0.9 Tucker-Lewis Index (TLI) value, which is within the acceptable range ($>= 0.9$) (Pedrosa et al. 2016). It also has a Root Mean Square Error of Approximation (RMSEA) value of 0.086, which is slightly higher than the reported acceptable range for this measure, *i.e.* $<= 0.08$ (Pedrosa et al. 2016). Overall, this factorization shows sufficient reliability.

3 Members' Factor Scores Relate to Other Behaviour in SCC

What do the factors of behaviour concerning textual contributions tell us about the behaviour of members in general? To answer this question, we calculate members' factor scores and then perform correlation analysis between these factor scores and base variables concerning other actions in our SCC that were not included in the FA.

3.1 Dataset

There are many methods to calculate member factor scores. We found that these methods have no significant impact on our results. We choose to use Anderson (DiStefano, Zhu, and Mindrila 2009) to calculate factor scores. This

[8] All factor correlations are significant at the 0.01 level two-tailed.

method requires VF and MA datasets to produce these scores. We create a Members Factors Scores (MFS) dataset $MFS = Anderson(MA, VF)$, where MA is the member actions dataset and VF is the variable-factor loadings dataset. MFS is a matrix that contains the score each member has on each of the factors. Thus, mfs_{ij} is the factor score that member m_i has on factor f_j.

3.2 Analysis and Results

Table 5 shows the correlation coefficients between the three factors and other base variables. Overall, we found statistically significant high correlations between our factors and other base variables. *Impact* factor correlates positively with all variables concerning quality of work, such as receiving comments, votes, follows, pins, and points and badges concerning conversations. *Activity* correlates with all variables concerning quantity of work, such as the total and unique number of badges, number of blogpost submissions, comments made on others and in response to others' comments, following others, voting, tagging, and points concerning community activities. It is interesting that *Policing/Rowdiness* factor did not correlate significantly with any of the variables. These correlation results tell us that our factors, although derived from a small set of variables concerning one type of textual contributions, are strong indicators of overall behaviour in this SCC.

4 Factors Predict Teams' Quality of Contributions

What impact do the factors of behaviour have on teams' quality of contributions? To answer this question, we create a dataset that allow us to analyze this relationship. This dataset has the factors of behaviour scores, aggregated to team-level, and the quality score of every project a team worked on. We found the presence of linear and non-linear relationships between the team-level aggregates of factors of behaviour with teams' quality of contributions. To show the presence of a linear relationship, we perform Pearson's correlation between the quality of contributions and the aggregated factor scores. To show the presence of a non-linear relationship, we use a neural network, a popular supervised machine learning function estimator algorithm (Haykin and Network 2004).

We must define the concepts of teams and projects. We consider discussion threads to be projects. A discussion thread consists of a submission and comments, which are replies to the submission or other comments under this submission. We use the term *textual contributions* to refer to the submission and comments collectively. A team consists of members who contribute - textually - to this project, *i.e.* discussion thread. Specifically, we define a team to be two or more members who participate in a discussion thread. A member is part of a team if the number of textual contributions of this member exceeds a certain threshold value, which we determine empirically. The quality of a project is the measurable positive effect it has on members who are not part of the team. We calculate the quality of each project as the sum of endorsements it receives from others. We exclude projects with fewer contributions or smaller team sizes than a specific threshold value, which we determine empirically.

4.1 Dataset

We create a dataset, D, which has 1,235 projects, as follows. First, we retrieve a list of all projects, *i.e.* discussion threads in our database. For each project, we retrieve the list of members who made textual contributions to it in a quantity that equals or exceeds a threshold value. These members are the team for this project. We include this project if its team size is larger than or equal to a threshold value. If the project's team size is large enough, we calculate the quality score of the project by summing up the number of followers each textual contribution in this project attracted from non-team members. This creates a dataset that has records of projects, team members, and project quality scores. We join this dataset with the MFS dataset to produce a new dataset that has records of projects, team members, factor scores, and the projects' quality scores. We aggregate this dataset to team-level to create a matrix D, which contains records of projects, aggregate factor scores for teach team through summations, averaging, standard deviation, *etc.*, and the project quality scores.

Thus, D is a matrix, $D = P \times F'$, where P is a set of projects and corresponding quality scores, and F' is the set of aggregated factors scores using all the aggregation methods listed in (Judge, Heller, and Mount 2002), including sum, average, and standard deviation. Thus, F'={sum(Impact_score), avg(Impact_score), std(Impact_score), …, *etc.*}. In D dataset, d_{ij} holds the j^{th} aggregation of a factor score for i^{th} project.

4.2 Showing the Presence of a Linear Relationship Through Correlation

We perform Pearson's Product-Moment Correlation using SPSS Statistics on D, the results of which are in Table 6. The correlation coefficients are statistically significant but do not show strong correlation. However, there is enough correlation to show the presence of a linear relationship between some of the factors of behaviour with a team's quality of contributions. We arrive at four findings.

First, IMPACT_TTL is the sum of the *Impact* factor score for all members in a team. This aggregated factor score (factor aggregate) moderately and positively correlates with project quality (see Table 6). This shows that teams that accumulate a high impact total are more prone to producing higher quality contributions. Second, IMPACT_MAX is the greatest *Impact* factor score in each team. This factor aggregate moderately and positively correlates with project quality (see Table 6), which shows that the highest *Impact* score in a team is indicative of its propensity to produce higher quality contributions. Third, ACTIVITY_MIN is the smallest *Activity* factor score in each team. This factor aggregate moderately and negatively correlates with project quality (see Table 6). A team having a large smallest *Activity* factor score is a team with the least active member being quite active. The negative correlation shows that highly active teams (as demonstrated by the high level of activity of

Table 5: Factors' correlations with other base variables using Pearson's Correlation. ** correlation is significant at the 0.01 level and * at the 0.05 level (two-tailed). We kept and bolded correlation records with absolute values ≥ 0.3.

Variable	Impact	Activity	Policing	Description
album flags made	**.398****	-.157**	0.011	flags made on others' inappropriate albums
album pins received	**.519****	-.211**	0.018	pins received from others
badge conversation	**.430****	**.569****	.044*	conversation badges received
badge conversation unique	.244**	**.326****	0.028	unique conversation badges received
blogpost submissions	.136**	**.341****	0.015	blogposts made by a member.
blogpost follows received	**.477****	-.175**	0.014	follows a member's blogposts received
blogpost pins received	**.409****	-.160**	0.015	pins a member's blogposts received
blogpost tags by me on me	**.470****	-.196**	0.014	tags a member made on own blogposts
blogpost votes received	**.518****	-.208**	0.017	votes a member's blogposts received
comment by me on me	**.669****	**.456****	.070**	replies a member made to comments received
comment by me on others	.238**	**.760****	.035*	comments a member made on others' submissions
comment by others on me	**.718****	-.098**	.053**	comments a member received
comment follows made	**.438****	**.523****	.037*	follows a member made on others' comments
comment follows received	.236**	**.537****	.035*	follows a member received on comments
comment votes made	.290**	**.496****	.053**	votes a member made on others' comments
comment votes received	**.501****	**.463****	.049**	votes a member received on own comments
points	**.347****	.202**	.073**	points earned by a member
point community	.234**	**.313****	.084**	points for community related activities
point conversation	**.568****	.224**	.047**	points for conversation related activities
tag made	.224**	**.415****	.067**	tags a member made
tag received	.235**	**.444****	.075**	tags a member received

Table 6: Factor aggregates correlation coefficients with projects' quality.** correlation is significant at the 0.01 level and * at the 0.05 level (two-tailed). We bold correlation coefficients that have an absolute value ≥ 0.3.

Factor Aggregates	Project Quality
IMPACT_TTL	**.307****
IMPACT_MAX	**.319****
IMPACT_MIN	-.220**
IMPACT_STD	.247**
ACTIVITY_TTL	.144**
ACTIVITY_AVG	-.196**
ACTIVITY_MDN	-.161**
ACTIVITY_MIN	**-.363****
POLICING_TTL	.143**
POLICING_MAX	.137**

even the least active member), produce lower quality of contributions. In other words, a team made exclusively of highly active members is counterproductive. Although such members may read or produce a lot of content; the quality of their contributions may be low and will not attract endorsement. This is further solidified by the lack of correlation between *Impact* and *Activity* factors reported in Section 2.3. Fourth, *Policing/Rowdiness* factor has no high or moderate correlations with project quality. Its highest correlations are POLICING_TTL and POLICING_MAX, which may show that the higher the total and maximum rowdiness and policing activities of members, the more follows the project will receive. That is, higher policing or rowdiness score in teams generates more attention in the form of following actions. However, because of the low correlation coefficients, this should be interpreted with caution.

4.3 Showing the Presence of a Non-linear Relationship Through Machine Learning

Using WEKA[9], we submitted D to a function approximation machine learning algorithm: Multilayer Perceptron, *i.e.* neural network. A neural network is a parametric machine learning method, which means it requires a set of parameters to be set before running, including learning rate, momentum, learning iterations, *etc*. There are ways to approximate these parameters to arrive at "better" results (Leung et al. 2003). However, we are not concerned with achieving higher prediction scores. We are concerned with demonstrating the presence of a non-linear relationship. Therefore, we would like to mitigate the effect of these parameters on our results by using the same parameter values across all our experiments. We set the number of hidden layers, number of nodes in hidden layers, learning rate, momentum, learning iterations, and cross-validation, to standard values set by WEKA: one hidden layer with nodes equal to the number of variables and the number of distinct project quality values (classes), 0.3, 0.2, 500, and 10. The neural network was able to predict the project quality from the aggregated factor scores within MAE of 2.4 follows, which establishes the presence of a non-linear relationship.

To find the aggregate factors of behaviour that are most important for prediction, we used three feature selection

[9] https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Table 7: MAE achieved by a neural network trained on the corresponding subset of factor aggregates.

Factor Aggregates	**WSE**	**CFS**	**ReliefF**
IMPACT_TTL	Y		Y
IMPACT_AVG	Y		
IMPACT_MIN	Y	Y	
IMPACT_STD			Y
ACTIVITY_TTL			Y
ACTIVITY_MIN	Y	Y	
ACTIVITY_MAX			Y
POLICING_TTL			Y
POLICING_AVG	Y		
POLICING_MAX	Y	Y	
POLICING_MDN	Y		
MAE	**2.3982**	**2.4853**	**2.603**

methods available in WEKA that can handle continuous data: Wrapper Subset Evaluation (WSE) (Kohavi and John 1997) for neural networks[10], Correlation-based Feature Subset Selection (CFS) (Hall 1999), and ReliefF Attribute Evaluation (ReliefF) (Kira and Rendell 1992; Robnik-Šikonja and Kononenko 1997). We ran each of these methods to obtain its suggested subset of features, *i.e.* factor aggregates, that are "best" for the prediction task. Finally, we ran each of the subsets using a neural network described above and reported the Mean Absolute Error (MAE) in Table 7.

4.4 Results

We see from Table 7 that a neural network trained on the WSE subset achieves the lowest MAE, which contains factors that were found in the previous section to have moderate but statistically significant relationship with the quality of contributions. This should not be interpreted to mean that only this subset is indicative of the quality of contributions. Given that *Activity* and *Policing/Rowdiness* factors are highly correlated, we would expect that they are redundant to one another for prediction. We learn from this result that through these factors, we can predict the quality of contributions of teams with remarkably low error.

Impact factor can be thought of as a consequence of conscientiousness factor in the Five Factor Model (FFM) of personality (McCrae and John 1992). Conscientiousness is the personality dimension that indicants a person's tendency to do what is "right" and perform these duties well (McCrae and John 1992; Buss 1991). A meta-analysis of personality's role in team's quality of contributions shows that the elevation of conscientiousness, *i.e.* sum, average, *etc.* is positively related to team performance (Judge, Heller, and Mount 2002). This is in tune with our finding.

Activity factor can be thought of as a consequence of extroverted personalities. Extroverts tend to be highly social individuals, with a great deal of energy (McCrae and John 1992). The elevation of extroversion in teams has no baring on team performance. Variability, *i.e.* standard deviation, variance, *etc.* is positively related to team performance (Judge, Heller, and Mount 2002). This is consistent with our results, which show that having a high minimum level of activity (lower variability in levels of activity) is negatively correlated with the team's quality of contributions.

5 Conclusion and Future work

We presented a process to arrive at factors of behaviour in online environments. We applied this process to a SCC to arrive at factors of behaviour regarding textual contributions. We named and interpreted these factors in the context of SCCs. We showed the presence of a correlative and predictive relationship between our factors and teams' quality of contributions. We showed that teams composed of higher levels of Impact and lower levels of Activity tend to produce higher quality of content. We will expand our work by applying it to other SCCs to assess the generality of our findings. We will also test various project quality measures. Moreover, we will repeat this process but with more variables that involves all aspects of behaviour in SCC to get a more holistic factors. Then assess their utility in predicting various aspects of individuals and teams, such as leadership ranking, and quality of contributions.

References

Aldrich, H. E., and Kim, P. H. 2007. Small worlds, infinite possibilities? how social networks affect entrepreneurial team formation and search. *Strategic Entrepreneurship Journal* 1(1-2):147–165.

Awad, M.; Khan, L.; and Thuraisingham, B. 2008. Predicting www surfing using multiple evidence combination. *The VLDB JournalThe International Journal on Very Large Data Bases* 17(3):401–417.

Buss, D. M. 1991. Evolutionary personality psychology. *Annual review of psychology* 42(1):459–491.

Cattell, R. B. 1973. *Personality and mood by questionnaire.* Jossey-Bass.

Cattell, R. 2012. *The scientific use of factor analysis in behavioral and life sciences.* Springer Science & Business Media.

DiStefano, C.; Zhu, M.; and Mindrila, D. 2009. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation* 14(20):1–11.

Fabrigar, L. R.; Wegener, D. T.; MacCallum, R. C.; and Strahan, E. J. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* 4(3):272.

Froman, R. D. 2001. Elements to consider in planning the use of factor analysis. *Southern Online Journal of Nursing Research* 2(5).

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 211–220. ACM.

Golbeck, J.; Robles, C.; and Turner, K. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, 253–262. ACM.

Gosling, S. D.; Augustine, A. A.; Vazire, S.; Holtzman, N.; and Gaddis, S. 2011. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking* 14(9):483–488.

Guilford, J. 1959. Personality mcgraw hill and company. *New York.*

[10] We use the same parameter values for this neural network.

Hall, M. A. 1999. Correlation-based feature selection for machine learning.

Hansen, T.; Jensen, J. M.; and Solgaard, H. S. 2004. Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior. *International Journal of Information Management* 24(6):539–550.

Harman, H. H. 1976. *Modern factor analysis*. University of Chicago Press.

Haykin, S., and Network, N. 2004. A comprehensive foundation. *Neural networks* 2(2004):41.

Henderson, A. R. 2006. Testing experimental data for univariate normality. *Clinica Chimica Acta* 366(1):112–129.

Hinkin, T. R.; Tracey, J. B.; and Enz, C. A. 1997. Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research* 21(1):100–120.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2):179–185.

Janhonen, M., and Johanson, J.-E. 2011. Role of knowledge conversion and social networks in team performance. *International Journal of Information Management* 31(3):217–225.

Judge, T. A.; Heller, D.; and Mount, M. K. 2002. Five-factor model of personality and job satisfaction: a meta-analysis. *Journal of applied psychology* 87(3):530.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and psychological measurement* 20(1):141–151.

Kalish, Y., and Robins, G. 2006. Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks* 28(1):56–84.

Kern, M. L.; Eichstaedt, J. C.; Schwartz, H. A.; Dziurzynski, L.; Ungar, L. H.; Stillwell, D. J.; Kosinski, M.; Ramones, S. M.; and Seligman, M. E. 2014. The online social self an open vocabulary approach to personality. *Assessment* 21(2):158–169.

Kira, K., and Rendell, L. A. 1992. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, 249–256.

Kline, P. 1987. Factor analysis and personality theory. *European Journal of Personality* 1(1):21–36.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97(1-2):273–324.

Korkmaz, S.; Goksuluk, D.; and Zararsiz, G. 2014. Mvn: an r package for assessing multivariate normality. *The R Journal* 6(2):151–162.

Kosinski, M.; Bachrach, Y.; Kohli, P.; Stillwell, D.; and Graepel, T. 2014. Manifestations of user personality in website choice and behaviour on online social networks. *Machine learning* 95(3):357–380.

Krause, J.; James, R.; and Croft, D. 2010. Personality in the context of social networks. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1560):4099–4106.

Leung, F. H.-F.; Lam, H.-K.; Ling, S.-H.; and Tam, P. K.-S. 2003. Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Transactions on Neural networks* 14(1):79–88.

Lim, B.-C., and Ployhart, R. E. 2004. Transformational leadership: relations to the five-factor model and team performance in typical and maximum contexts. *Journal of Applied Psychology* 89(4):610.

McCrae, R. R., and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60(2):175–215.

McCrae, R. R., and John, O. P. 1998. An introduction to the five-factor model and its applications. *Personality: critical concepts in psychology* 60:295.

McCrae, R. R. 2010. The place of the ffm in personality psychology. *Psychological Inquiry* 21(1):57–64.

Mount, M. K.; Barrick, M. R.; and Stewart, G. L. 1998. Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human performance* 11(2-3):145–165.

Odum, M. 2011. Factor scores, structure and communality coefficients: A primer. *Online Submission*.

Osborne, J. W., and Costello, A. B. 2009. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review* 12(2):131–146.

Pedrosa, R. B. D. S.; Rodrigues, R. C. M.; Padilha, K. M.; Gallani, M. C. B. J.; and Alexandre, N. M. C. 2016. Factor analysis of an instrument to measure the impact of disease on daily life. *Revista brasileira de enfermagem* 69(4):697704.

Quercia, D.; Kosinski, M.; Stillwell, D.; and Crowcroft, J. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 180–185. IEEE.

Robnik-Šikonja, M., and Kononenko, I. 1997. An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML97)*, 296–304.

Roccas, S.; Sagiv, L.; Schwartz, S. H.; and Knafo, A. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin* 28(6):789–801.

Salgado, J. F. 1997. The five factor model of personality and job performance in the european community. *Journal of Applied psychology* 82(1):30.

Saroglou, V. 2002. Religion and the five factors of personality: A meta-analytic review. *Personality and individual differences* 32(1):15–25.

Suhr, D. D. 2006. *Exploratory or confirmatory factor analysis?* SAS Institute Cary.

Tabachnick, B. G.; Fidell, L. S.; and Osterlind, S. J. 2001. Using multivariate statistics.

Wehrli, S. 2008. Personality on social network sites: An application of the five factor model. *Zurich: ETH Sociology (Working Paper No. 7)*.

Yang, H.-L., and Tang, J.-H. 2004. Team structure and team performance in is development: a social network perspective. *Information & Management* 41(3):335–349.

Yong, A. G., and Pearce, S. 2013. A beginners guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology* 9(2):79–94.

Zaamout, K., and Barker, K. 2017. Structure of crowdsourcing community networks. *IEEE Transactions on Computational Social Systems*.

Zhang, J.; Ackerman, M. S.; and Adamic, L. 2007. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, 221–230. ACM.