Modelling and Analysis of User Behaviour in Online Communities

Sofia Angeletou, Matthew Rowe, Harith Alani

Knowledge Media institute, Open University, UK
{s.angeletou, m.c.rowe, h.alani}@open.ac.uk

Abstract. Understanding and forecasting the health of an online community is of great value to its owners and managers who have vested interests in its longevity and success. Nevertheless, the association between community evolution and the behavioural patterns and trends of its members is not clearly understood, which hinders our ability of making accurate predictions of whether a community is flourishing or diminishing. In this paper we use statistical analysis, combined with a semantic model and rules for representing and computing behaviour in online communities. We apply this model on a number of forum communities from Boards.ie to categorise behaviour of community members over time, and report on how different behaviour compositions correlate with positive and negative community growth in these forums.

1 Introduction

Online communities form a fundamental part of the web today where a large portion of the Internet's traffic is driven by and through them [**?**]. These communities are where the majority of web users share content, seek support, and socialise. On the one hand, for companies and businesses, such online communities tend to yield much value in terms of idea generation, customer support, problem solving, etc. [**?**]. On the other hand, managing and hosting these communities can be very costly and time consuming, and hence their owners and managers have a great vested interest in ensuring that these communities continue to flourish, and that their members remain active and productive.

One of the main metrics often used by community managers to measure community health is the number of members and posts. These numbers give a good indication of community popularity. However, for deeper assessment and forecasting, other more complex qualitative and behavioural parameters need to be considered [**?**,**?**]. For example, behavioural analytics complement other community assessment tools and increase the value of the data [**?**].

Health of online communities is a relatively new and complex concept that is codependent on the emergence and evolution of user behaviour in those communities. Domination of any type of behaviour, whether positive or negative, could encourage others to change their behaviour or even abandon the community [**?**]. Therefore, monitoring and analysis of behaviour and its evolution over time, in addition to straightforward metrics such as post and user counts, can provide valuable information on how healthy an online community currently is or will

be in the near future. Behaviour in online communities is usually associated with various social and technical parameters which influence the roles users hold in different settings [**?**]. Associating users with behavioural categories involves identifying and applying constraints, expectations and frameworks to categorise and follow user behaviour in the community [**?**].

To support community owners and managers in observing and maintaining the good health of their communities, we first need to (a) model, capture and monitor the activities of community members, (b) analyse emergent behaviours and their change over time, (c) understand the correlation of certain types of behaviour with community evolution, and (d) learn how and when to intervene to influence the interactions and behaviour of community members. In this work we focus on the first three tasks; the first task is concerned with producing a semantic model for representing user activities in online communities and the attention they generate in those communities. The second task focuses on comparing the emergence and change in patterns of behaviour with the evolution of those communities. And the third task explores how community composition, i.e. a macro-analysis of the community, is correlated with the health of the community. By monitoring activities in online communities we will be able to better understand and predict their evolution directions; i.e. whether they are flourishing (positive evolution) or diminishing (negative evolution).The main contributions in this paper are as follows:

1. *Method to infer user roles in online communities:* We employ semantic rules to label community users with their role and utilise dynamic feature binning to account for the dynamic nature of communities and their propensity to evolve.
2. *Ontology to model behavioural features and support community role inference:* Allowing user features to be captured in a common machine-readable format across communities and platforms.
3. *Analysis of community health through role composition:* We demonstrate the utility of our approach by analysing three communities over a 3 year period, showing the effects of behaviour composition changes on community health and compositions that are key signifiers of healthy or unhealthy communities.

In the following section we report on various related works in the area of behaviour and community analysis. In section 3 we present our methodology for user behaviour analysis and how we utilise Semantic Web technologies to infer the role that a user has within an online community. Section 4 describes our analysis of community health in the three sample forums, followed by discussions and future plans in section 5, and finishes with conclusions in section 6.

2 Related Work

In this section we report on existing works on analyses of behaviour patterns and roles in online communities. The identification of behaviour is often based on features which reflect the intensity, persistence, focus, reciprocity and polarity of user activities.

For instance, users who contribute with high intensity, reciprocity and persistence, positive polarity and are focused on supporting and contributing to the community are characterised as *moderators*, *mediators* [**?**], *captains* and *pillars* [**?**]. When such users are able to set the standard for community interactions, they get labelled as *celebrities* [**?**]. *Popular initiator*, *popular participant* and *joining conversationalist* [**?**] are three roles very similar to the celebrity type since their intensity, persistence and reciprocity are also quite high. Another type of prolific, but not as widely popular, user is the *elitist*, who demonstrates high values for the above dimensions but communicates with a smaller group of users. On the lower end of the activity scale the *lurker* is the most frequently observed role and is defined as a participant who consumes but does not contribute and usually has a strong personal focus [**?**,**?**,**?**]. Similarly described roles are those of *content consumers* [**?**], *grunts* and *taciturns* [**?**] who do contribute but with low intensity. The polarity of the user contribution has also been used to distinguish the negative roles of *troll* and *flamer* who exhibit disruptive behaviour similar to the *ranter*. Like celebrities, ranters also demonstrate high intensity and persistence yet their primary goal is to raise discussions on the topic of their interest for some personal goals, same as *over-riders* and *generators* [**?**].

Although there is no commonly agreed set of behaviour patterns and labels, the social and technical features considered by the above works when categorising behaviour do share some characteristics, albeit sometimes tailored to suit the online communities under investigation. The approaches followed in the above references are normally based on correlating a set of features taken from a specified snapshot of a community, then labelling users with behaviour roles that fits the results from that snapshot.

Our analysis extends these approaches by introducing a framework for representing, computing, and monitoring users' behaviour over time. We extend the state of the art by demonstrating how various features can be modelled and articulated into semantic rules to automate the detection and categorisation of users with specific types of behaviour.

Furthermore, is it often the case that fixed ranges of feature-values are calculated when associating them with behaviour types, so if the feature value for a given user falls within that range, then that user will be labelled with the corresponding behaviour. However, it is often the case that such value ranges seize to apply if the time window or community changes. Here we present a framework that enables a more dynamic association of features to roles (Section 3.2) and allows for on-the-fly value threshold assignment that takes context into account.

Community health indicators are normally dependent on the goals and characteristics of the community [**?**]. Straightforward measures such as number and frequency of posting are often used as an index of community health. For example, it has been shown that the activity of a group can be maintained in high levels by long term members, who help keep the group together [**?**,**?**]. On the other hand, it has been found that having *lurkers* in a community does not necessarily have a negative influence [**?**]. Our framework allows us to investigate the influence that various user behaviours and interactions have on the overall

health of communities over time. This helps in understanding what the optimum compositions of behaviour should be for a given community, and in forecasting community evolution. We analyse the influence and predictability of a wide set of behaviours on community health.

3 Methodology for Behaviour Analysis

In this section we describe our approach for labelling users in online communities with the roles they hold in the context of these communities and in a specified timeframe. To perform such labelling we first need to capture the activity of users in online communities, define what sort of behaviour is associated with particular roles and compare it to a user's activity. For capturing users' activities we define an *ontology* (Section 3.1) that represents all involved entities and their interactions. We also define "*rule skeletons*" (Section 3.2) which provides high-level descriptions of how certain *features* are associated with various behaviour roles. Our community analysis then *fleshes out* these rules with *dynamic and automatically computed value-ranges* that will eventually determine which users will be categorised with which behaviours. Finally we apply these rules on all community members to infer their behaviour types (Section 3.3).

3.1 Ontology

Capturing a user's activity in online communities is a primary step to analysing his behaviour. Fig. 1 presents a portion of our Behaviour Ontology[1] which represents online community users and their interactions. The ontology extends SIOC [**?**] to refine the representation of low level user activities and interactions. It also extends the Social Reality [**?**] ontology which provides an abstract representation of social roles and their contexts. The main concepts and properties of the ontology are:

- **sioc:UserAccount** is a SIOC class to represent online community users.
- **oubo:Post** represents users' main activities; writing and replying to posts.
- **oubo:PostImpact** summarises a post's replies, comments, forwards, etc.
- **oubo:UserImpact** encodes the user impact (behaviour)
- **oubo:TimeFrame** is the temporal context during which the analysis is carried out and the association of a user to a role holds.
- **social-reality:C** represents context, such as time period (oubo:TimeFrame) and a forum (sioc:Forum).
- **oubo:Role** represents the roles we derive for users based on their activities in the community.
- **oubo:belongsToContext** links context-related concepts, such as oubo:TimeFrame, sioc:Forum, and social-reality:C.
- **social-reality:counts_as** associates a user with a oubo:Role.
- **social-reality:context** associates a user role with its context.

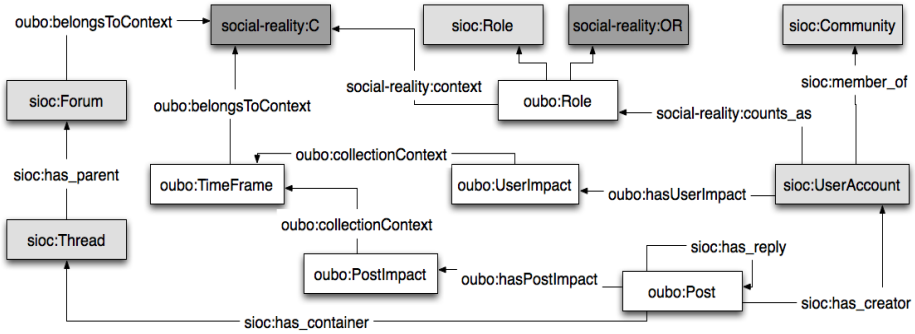


Fig. 1. Behaviour Ontology

The ontology also defines the rules for the population of features and classification of the users in different roles using dynamically populated feature weights (Fig. 3). This is discussed in Section 3.3.

3.2 Behaviour Roles

To derive the behaviour roles of community users, their activity patterns need to be compared against the behavioural characteristics of each role type. In the literature, the features (e.g. number of posts and replies, in/out degrees) associated with behaviour are often given static value ranges (i.e. min and max values for corresponding behaviours) which are calculated based on the community snapshot under analysis (e.g. [**?**,**?**,**?**]). However, a common characteristic of online communities is their propensity to evolve and develop as new users participate and the dynamic of the community changes. An effect of such dynamism is that should we learn a static value for the maximum and minimum values for each role's features then applying such values at a later point in time will lead to users being omitted from the labelling process.

To counteract such effects we use the notion of *skeleton rules*, where each rule contains a mapping between a given feature and the *level* that the value of the feature should take to indicate a certain type of behaviour: *low*, *medium* or *high*. In using this method we can shift the bounds that constitute such levels as the dynamics of the community changes, thereby allowing more users to be labelled with behaviour roles.

Many different behaviours and associated features have been proposed in the literature (section 2). Our framework for modelling and computing behaviour is not tied to any specific community or behaviour types or feature compositions. To demonstrate our framework, we selected the behaviour roles defined in [**?**], which covers a range of common activity and participation roles. In [**?**], Chan and Hayes performed clustering over users within Boards.ie community forums and then carried out manual analysis to derive the behaviour labels for each cluster. They clustered the users using a list of key features that covered (a) the structural network properties of a user within the community, (b) the user's

[1] `http://purl.org/net/oubo/0.3`

popularity amongst the community, (c) their propensity to initialise discussions, and (d) their persistence in discussions. These features are:

- **In-degree Ratio:** The proportion of users U that reply to user v_i, thus indicating the concentration of users that reply to v_i.
- **Posts Replied Ratio:** Proportion of posts by user v_i that yield a reply, used to gauge the popularity of the user's content based on replies.
- **Thread Initiation Ratio:** Proportion of threads that have been started by v_i. This feature captures the propensity of a user to instigate discussions and generate fresh content for the community.
- **Bi-directional Threads Ratio:** Proportion of threads where user v_i replies to a user and receives a reply, thus forming a *reciprocal* communication.
- **Bi-directional Neighbours Ratio:** The proportion of neighbours where a *reciprocal* interaction has taken place - e.g. v_i replied to v_j and v_j replied to v_i. This can be thought of as the intersection between the set of *repliers* and *recipients*. This measure allows the *reciprocal* characteristics of the user to be captured and their participation with users in the community, where higher values demonstrate a tendency to interact.
- **Average Posts per Thread:** The average number of posts made in every thread that user v_i has participated in. Allows the level of discussion that the user participates in to be gauged.
- **Standard Deviation of Posts per Thread:** The standard deviation of the number of posts in every thread that user v_i has participated in. This gauges the distribution of the discussion lengths, for example, one would expect that a user who often discusses at length with other users would have a high *Average Posts per Thread* and a low *Standard Deviation of Posts per Thread*, while someone who varies their participation will have a higher *Standard Deviation of Posts per Thread*.

Based on the feature-behaviour compositions in [**?**] and in other literature, we deduce a mapping of these common feature to value ranges for each behaviour role (Table 1).

3.3 Constructing and Applying Behaviour Rules

Our approach for constructing and applying rules is shown in Fig. 2 and is composed of four stages that function in a cyclical manner: *First*, we construct features for all users who participated in the given community at a specific point in time. *Second*, we derive bins for features in the community, thus providing the bounds for the *low*, *medium* or *high* levels of each feature. *Third*, the rule base is constructed using the skeleton rule base and the levels from the binning. *Fourth* we apply the rules to each member of the community and derive a role label, this provides the role composition of the community at a given time snapshot. As Fig. 2 shows, the process is repeatable over time, thereby allowing the composition of a given community to be monitored by inferring the role of each community user at a given point in time. We now explain the four steps in greater detail.

Table 1. Roles and the feature-to-level mappings

Role	Feature	Level
Elitist	In-Degree Ratio	low
	Bi-directional Threads Ratio	high
	Bi-directional Neighbours Ratio	low
Grunt	Bi-directional Threads Ratio	med
	Bi-directional Neighbours Ratio	med
	Average Posts per Thread	low
	STD of Posts per Thread	low
Joining Conversationalist	Thread Initiation Ratio	low
	Average Posts per Thread	high
	STD of Posts per Thread	high
Popular Initiator	In-Degree Ratio	high
	Thread Initiation Ratio	high
Popular Participants	In-Degree Ratio	high
	Thread Initiation Ratio	low
	Average Posts per Thread	med
	STD of Posts per Thread	med
Supporter	In-Degree Ratio	med
	Bi-directional Threads Ratio	med
	Bi-directional Neighbours Ratio	med
Taciturn	Bi-directional Threads Ratio	low
	Bi-directional Neighbours Ratio	low
	Average Posts per Thread	low
	STD of Posts per Thread	low
Ignored	Posts Replied Ratio	low

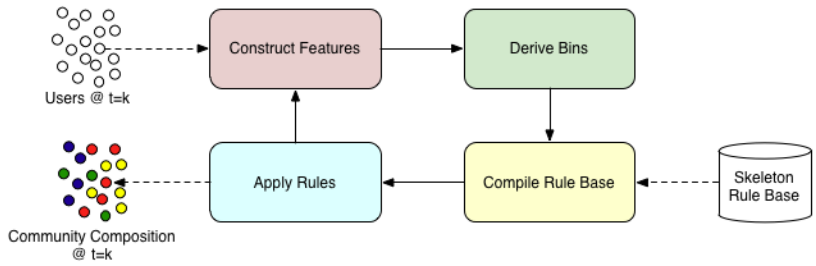


Fig. 2. Overview of the approach to analyse user behaviour, label users with behaviour roles and derive the community composition

Step 1: Constructing Features The previously defined statistical features are constructed for each user at a given point in time. For our experiments, as we describe in the following section, we use a window function to extract all posts made within a given community during that time period. Using the reply-to structure of the posts, we then compile the above features and create an instance of oubo:UserImpact that contains the features for the given user measured at a given point in time.

Step 2: Deriving Bins Our skeleton rule base contains mappings of features-to-levels, allowing the bounds of the levels to be altered depending on the dynamics of the community. We set the bounds through *binning*, a process that

discretizes continuous feature values into three bins (*low*, *medium* or *high*) using Weka's[2] discretization filter together with equal frequency binning. Using a naive binning approach - e.g. splitting a feature range into thirds - can result in a large frequency skew within a single bin, equal frequency binning avoids this and provides a distribution-dependent notion of levels. The process of deriving the bins is performed when we analyse the community at a different point in time, in doing so our intuition is that we will reduce the number of unclassified users and account for changes in the community's dynamics - we show this empirically in the following section.

Step 3: Compiling the Rule Base The association of users to roles is inferred by analysing the captured data for each user against features that each role embodies. To perform such inferences our approach employs the SPARQL Inference Notation (SPIN)[3] framework, allowing the encoding of rules as SPARQL queries. The benefits of such an approach is that the rules are embedded in an ontological model and can, therefore, be shared and executed across platforms that support SPARQL Extensions and Jena.

To compile the rule base we create a rule for each behaviour role within the community. For each role a new instance of the oubo:RoleClassifier Class is created and associated with a set of features as shown in Fig. 3. Each feature has a minimum and maximum value which specify the range of feature values a user should have for this feature in order to be assigned to this role. We use the skeleton rule of the role to provide the rule's syntax and then replace the levels with the necessary bounds produced by our binning procedure. In the majority of cases a combination of features is required for the association of a user to a role. In these cases, all the feature values of the user should belong to the ranges specified by each feature belonging to the relative RoleClassifier instance.

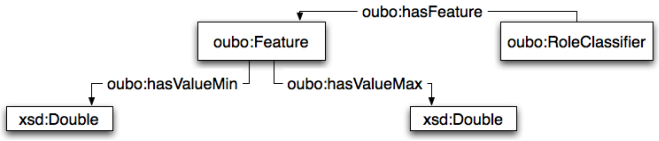


Fig. 3. Association of Roles with Features

Step 4: Applying Rules The produced rules are SPARQL Construct queries that exploit the power of SPIN Functions by testing each instance of sioc:UserAccount against each instance of oubo:RoleClassifier, and then assigning the user to the role whose associated classifier is matched. Fig. 4 presents an example of one such query that is encoded as a spin:rule for the class sioc:UserAccount. For each instance of sioc:UserAccount (represented by the variable `?this`) a set of

[2] `http://www.cs.waikato.ac.nz/ml/weka/`
[3] `http://spinrdf.org`

triples are inferred representing the association of the user with a new instance of a specific role `?role`. The `?role` is an instance of a subclass of oubo:Role, `?t` depending on the classification of the user as described in Fig. 4. The spin function oubo:fn_getRoleType executes the process described above and returns the appropriate classifier. Then using the smf:buildURI SPARQL Motion function `?t` is built so that it is associated to the correct subclass of oubo:Role. Finally, the `?context` is created and connected to the particular `?role` via the relation `social-reality:hasContext`, and is also associated to the temporal and forum relevant contexts in which it makes sense that the user holds this particular role - i.e. a given user can have multiple roles within different communities and time periods.

```
CONSTRUCT {
    ?role a ?t .
    ?this social-reality:count_as ?role .
    ?context a social-reality:C .
    ?role social-reality:content ?context .
    ?temp a oubo:TemporalContext .
    ?forum a sioc:Forum .
    ?forum oubo:belongsToContext ?context .
    ?temp oubo:belongsToContent ?context
} WHERE {
    BIND (oubo:fn_getRoleType(?this) AS ?type) .
    BIND(smf:buildURI("oubo:Role{?type}") AS ?t) .
    .....
}
```

Fig. 4. SPARQL CONSTRUCT encoded as a spin:rule in the class sioc:UserAccount

4 Analysis of Community Health

Healthy communities provide users with the resources from which information can be sought, interactions made and discussions participated in. In this section we explore the relation between the composition of a community, i.e. the various roles that users have within a community and the proportion to which such roles make up the community (e.g. 20% elitists, 10% taciturns, etc.), and the activity in the community. Through experiments and analysis of the subsequent results, we seek to identify key community compositions that are associated with both an increase and decrease in community activity. In doing so, we are provided with an understanding of how certain behaviour types are correlated with community evolution and what compositions are signifiers of healthy and unhealthy communities. We here consider the level of activity as a proxy of community health, but other parameters (e.g. reciprocity, sentiment) could also be considered.

4.1 Experimental Setup

For our experiments we used a dataset collected from the Irish community discussion forums, Boards.ie. We extracted all posts from the beginning of 2004 through to the end of 2006 for our analysis - thereby capturing a 3 year period

over which we could perform our analyses. Rather than analysing the entire site, we selected 3 forums that showed a variance in activity throughout the analysed period - the plots of post activity are shown in Fig. 6.

- *Forum 246 (Commuting and Transport):* Demonstrates a clear increase in activity over time.
- *Forum 388 (Rugby):* Exhibits periodic increase and decrease in activity and hence it provides good examples of healthy/unhealthy evolutions.
- *Forum 411 (Mobile Phones and PDAs):* Increase in activity over time with some fluctuation - i.e. reduction and increase over various time windows.

In order to compile a dataset for each forum we used the following process: beginning on 1st January 2004 we used a window from 13 weeks prior to this date up until the date as our feature window. Within this window we extracted all the posts made within the forum, and used the posts to compile the statistics for each unique user who had made a post within that window. Once we had finished building the statistics for each user at that collection date, we then rolled the date forward 84 days, leaving a 12 week gap between our last collection date. The window was compiled once again: going 13 weeks back, returning all posts within the window, and then building the user features for each unique user within the window. We repeated this process until the end of 2006. To provide a coarse measurement of the community's *health* we also counted the number of posts made in the forum during that window - allowing the activity at one point in time to be contrasted against earlier activity.

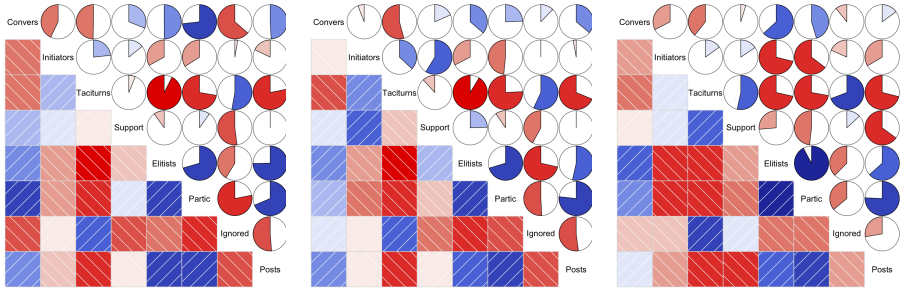
Following the compilation of our user statistics at the incremental time steps (13 time steps in total) and the instantiation of oubo:UserImpact, we categorised each user using our previously described rules. In doing so we were able to measure the composition of the community over time as differing percentages of users that have taken on such roles within the community. We then correlated this composition with the health of the community at that point in time, seeking patterns that describe a healthy and unhealthy community in terms of either an increase or decrease in activity, e.g. having many users of a certain role type reduces community activity. We also report on how our approach greatly reduces the percentage of users that could not be classified by the current behaviour rules.

4.2 Results

Fig. 5 shows the correlograms from the individual forums. The upper panel shows the extent to which a correlation exists between two features and the polarity of the correlation (i.e. positive or negative). The greater the portion of the circle that is filled then the greater the correlation. The colour indicates the polarity: blue indicates a positive correlation and red indicates a negative correlation.

For forum 246 (Commuting and Transport), shown in Fig 5(a), a positive correlation exists between the post count and both the number of *elitists* and *popular participants* ('*partic*' in the chart), indicating that as more users assume such roles within the community then activity increases. This is due to the *popular participants* driving discussions and joining in with the community to make

it more vibrant. Meanwhile, *elitists* communicate a lot with their own group and thus drive its activity. In forum 246 we also observe a negative correlation between the post count and the proportion of *taciturns* within the community, indicating that users who communicate very little with others can reduce the overall interactions and dynamic of the community. Fig 5(a) also shows that an increased number of *ignored* users has a negative effect on community activity, which follows intuition. As forum 246 concerns transport discussion, many users post questions regarding travel situations and modes of transport, and hence if a large portion of those users are ignored then activity in this community diminishes as questions remain unanswered.



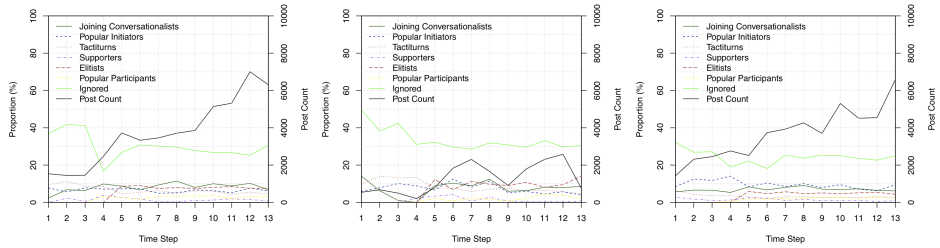
(a) Forum 246 (b) Forum 388 (c) Forum 411

Fig. 5. Correlation between the various features within each forum

Similarly to forum 246, in forum 388 (Rugby) we also find a positive correlation between post counts and number of *elitists* and *popular participants* (Fig. 5(b)). There is also a slight correlation between post count and the proportion of *conversationalists* within the community, which demonstrates the value of conversations and debates in driving this community. There is also a negative correlation between the post count and the number of *taciturns* and *ignored* within the community.

For forum 411, (Mobile Phones and PDAs), the correlogram in Fig. 5(c) demonstrates similar patterns to the previous two forums in terms of positive correlations. Once again, we find that the post count has a positive correlation with the proportion of *elitists* and *popular participants*. We also find that the post count has a negative correlation with the proportion of *taciturns* and, in this individual forum, with the proportion of *supporters*. Supporters have a mid-level range of 'Bi-directional Threads Ratio', indicating that conversation is one of the drivers behind this role. However in forum 411, debate between users appears to be limited, as users require information regarding support and are less inclined to chat with other users repeatedly. This is also supported by the lack of positive correlation between the post count and the proportion of *conversationalists* in this forum.

Within Fig. 6 we show the composition changes in each of the analysed forums over time, plotted together with the post count within each forum. For all forums we find that activity increases where the proportion of *ignored* users decreases.



(a) Forum 246 (b) Forum 388 (c) Forum 411

Fig. 6. Changes in composition over time plotted with forum post counts

For forum 246 we see that as the proportion of *joining conversationalists* and *popular initiators* increases so does activity. The same applies to forums 388 and 411, although for the latter the number of *popular initiators* was more important, confirming our earlier correlation analysis for that forum where conversations are not driving activity in this forum.

An interesting factor in each forum is the effect of composition stability on community activity. In the case of forum 246 and forum 411 we find that the composition converges on relatively stable proportions (i.e. with no extreme fluctuations in role types) and leads to a large rise in activity over time. Conversely for forum 388, Fig. 6(b) shows that the lack of stability in the community's composition leads to fluctuations in activity. This suggests that although limiting the number of ignored users and taciturns in the community would be beneficial, a stable mix of user types actually improves community health.

Unclassified Users Our rule-based approach for inferring the role of a given user at a given point in time utilises dynamic binning to update the low, mid and high bounds for each rule's features. In utilising dynamic binning our intuition was that our approach could adapt to the changing dynamics of the community as it evolves over time, i.e. the *low* bound for a feature during one year will differ from a year later. Therefore to demonstrate the utility of our approach we measured the proportion of users that are unclassified in each forum. We contrast this against the proportion of users who are unclassified when the feature bounds are not updated at each time step (40.05% unclassified users for forum 246, 37.92% for forum 388 and 39.84% for forum 411). Results show that our method of dynamically updating the bins for feature bounds enables a greater coverage of the users (29.06% unclassified users for 246, 28.06% for 388 and 28.68% for 411), and therefore enables, on average, a greater proportion of users to be labelled with a given role. Additional behaviour rules can be added to increase the percentage of classified users even more.

Predicting Community Health Thus far we have concentrated on identifying correlations between the post count within single forums and the proportion of roles within such communities. An important aspect of undertaking such analysis is the ability to forecast community health should the composition of the community change. To demonstrate the utility of such an approach we performed

a binary classification task to identify, based on the composition of the community, whether the activity had either *increased* or *decreased* since the previous time window. We built a dataset for each forum and constructed an instance for each of the 13 time windows. Each instance contained the features describing the 7 behaviour roles in the community together with the percentage of users allocated with such roles and a class label denoting the activity in the forum as having either increased (*pos*) or decreased (*neg*) since the previous time window. For our classification task we used the J48 decision tree classifier in a 10-fold cross validation setting (due to the limited size of the datasets) by: *first*, identifying increases and decreases in each of the forums, and *secondly*, identifying activity changes across communities, by combining forum datasets together into a single dataset. To report on the performance of our approach we used precision, recall, f-measure (setting $\beta = 1$) and the area under the Receiver Operator Characteristic Curve (ROC).

Table 2. Results from detecting changes in activity using community composition

Forum	P	R	F_1	ROC
246	0.799	0.769	0.780	0.800
388	0.603	0.615	0.605	0.775
411	0.765	0.692	0.714	0.617
All	0.583	0.667	0.607	0.466

Table 2 presents the results from our classification experiments. For forum 246 we achieve the highest F_1 value due to the activity in the forum steadily increasing over time and the precision value indicating that in this forum the composition patterns account for fluctuations in activity. For forum 388 we return the lowest F_1 value, indicating that the variance in activity renders the prediction of activity increase difficult within this forum, this could possibly be due to the seasonal fluctuations in interest surrounding the rugby season. For forum 411 we achieve high precision, indicating that activity can be precisely detected based on the composition in this forum. When performing cross-community health predictions we achieve lower F_1 values than those for forums 246 and 411 and the lowest ROC value. This indicates that cross-community patterns are not as reliable as individual community analysis, where patterns in compositions for single forums account for the idiosyncratic behaviour.

For our next task we induced Linear Regression models by regressing the post count on the community composition, using each of the role proportions as our predictor variables, seeking a relationship between the change in the overall composition of the community and the health of the forum. We now report on the model learnt for forum 388 (Commuting and Transport), given that this model achieved the highest coefficient of determination while forums 246 and 411 achieved R^2 values of 0.649 and 0.793 respectively.

Table 3 shows the results from the induced model.[4] The model indicates that should a community increase in its proportion of *popular initiators* and *popular*

[4] We found no multicolinearity between variables in the model when testing using the Variance Inflation Factor, suggesting that the roles are distinct in this forum and there are no clear dependencies between them.

participants while decreasing in the proportion of *supporters* and *ignored* then the community's activity will increase. However, an increase in *ignored* and *supporters* will yield a reduction in the post count and therefore a reduction in the "health" of the community. Such patterns can be used to alert a community manager of the current state of their community and its projected evolution. Managers could then use this information to decide what action to take to influence the evolution of their community in a positive way.

Our analysis of community forums has explored the correlation between community composition and health, and how predictions can be performed. Through this analysis we have identified four key *take-home* messages:

1. Healthy communities contain more elitists and popular participants.
2. Unhealthy communities contain many taciturns and ignored users.
3. Communities exhibit idiosyncratic compositions, thus reflecting the differing dynamics that are required/exhibited by individual communities.
4. A stable composition, with a mix of roles, increases community health.

Table 3. Linear regression model induced from the forum composition of f388

Role	Est' Coefficient	Standard Error	t-Value	P($x >$t)
Joining Conversationalist	69.20	43.82	1.579	0.1751
Popular Initiators	173.41	54.72	3.169	0.0248 **
Taciturns	-135.97	101.91	-1.334	0.2397
Supporters	-266.53	109.60	-2.432	0.0592 *
Elitists	-105.19	55.88	-1.882	0.1185
Popular Participants	372.44	103.24	3.608	0.0154 **
Ignored	-75.69	33.39	-2.267	0.0727 *
Summary: Res. St Err: 311.5, Adj R^2: 0.8514, $F_{7,5}$: 10.82, p-value: 0.0092				
Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1				

5 Discussion and Future Work

The communities we chose to analyse in this paper were forums from Boards.ie. It is possible of course that different behavioural patterns could emerge when analysing different communities. However, there is no reason to assume that our current behaviour types would not apply, since the basic statistics that underpin them are not specific to any community. As for the features we chose to measure users' value, we have already started comparing them with results from Twitter and highlighting variations in their influence from Boards.ie.

Churn - i.e. the loss of community members - is a risk posed to online communities and one that community operators wish to avoid. Churn is normally affected by various community features [**?**]. By analysing the community composition that is correlated with a healthy community that evolves into an unhealthy community, we will be able to learn patterns that could then be used to preempt such changes, and thus warn community managers of the possibility of such decline. Our future work will also seek to identify key users within online communities and monitor their behaviour to predict their churn which would have a detrimental effect on the community. The combination of such *micro* and *macro*-level analysis would enable community managers to identify which users to pay more attention to in order to maintain a healthy community.

The emergence and evolution of certain types of behaviour could be dependent on the rise and fall of other behavioural types. Such possible correlations need further investigation and can support prediction of community evolution. On the other hand, negative user behaviour could badly influence health of the community, and an unhealthy community could foster negative behaviour. Many studies have shown that certain features (e.g. sentiment, time, user popularity) influence the spread of some types of behaviour, or increase response to posts. Some of these findings differ from one online community to another. The ideal mix and spread of behavioural types that boosts health in communities is still unknown, and it is likely to be dependent on the characteristics and goals of the communities in question.

Our approach for inferring user roles accounts for the dynamic nature of communities by utilising the repeated binning of feature values and using skeleton rules that map features to value levels. In doing so we have shown the ability of this approach to reduce the proportion of unclassified users when compared with an approach that does not utilise such updating. However, on average our approach still misses $\sim$29% of users and is unable to associate those users with behaviour types. Our future work will explore methods to reduce this percentage by exploring the use of clustering and outlier detection techniques to account for new emerging roles within the community.

6 Conclusions

In this paper we have presented an approach to label the users of online communities with their role based on the behaviour they exhibit. We presented an ontology to capture the behavioural characteristics of users as numeric attributes and explained how semantic rules can be employed to infer the role that a given user has. There is currently no standard or agreed list of behaviour types for describing activities of users in online communities. Behaviour categories suggested in the literature are sometimes based on different observations and conceptions. In this paper our aim was not to identify the ultimate list of behavioural types, but rather to demonstrate a semantic model for representing and inferring behaviour of online community members.

A key contribution of this paper is the analysis of community composition over time and the correlation of such compositions with the health of communities, characterised by the number of posts made within a given community. Our empirical analysis of such correlations identified patterns in community composition that lead to both healthy and unhealthy communities, where a greater proportion of *elitists* and *popular participants* lead to an increase in activity, while a greater proportion of *taciturns* and *ignored* users lead to a decrease in activity. We also found that a stable community composition of role proportions lead to an increase in activity within the community, suggesting that wide fluctuations in role types could reduce community health.

Acknowledgment

This work was supported by the EU-FP7 projects WeGov (grant 248512) and Robust (grant 257859). Also many thanks to Boards.ie for providing data.

References

1. Lars Backstrom, Ravi Kumar, Cameron Marlow, Jasmine Novak, and Andrew Tomkins. Preferential behavior in online groups. In *Proc. Int. Conf. on Web Search and Web Data Mining (WSDM)*, New York, NY, USA, 2008.
2. John G. Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards semantically-interlinked online communities. In *Proc. 2nd European Semantic Web Conf. (ESWC)*. Springer, 2005.
3. Jeffrey Chan, Conor Hayes, and Elizabeth Daly. Decomposing discussion forums using common user roles. In *Proc. Web Science Conf. (WebSci10)*, Raleigh, NC: US, 2010.
4. Scott A. Golder and Judith Donath. Social roles in electronic communities. In *in Association of Internet Researchers (AoIR) 5.0*, 2004.
5. Rinke Hoekstra. Representing social reality in OWL 2. In *Proc. of OWLED*, 2010.
6. Lithium Technologies Inc. Community health index for online communities, 2009, http://pages.lithium.com/community-health-index.html.
7. Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The effect of user features on churn in social networks. In *Proc. ACM Web Science Conf. (WebSci'11)*, Koblenz, Germany, 2011.
8. Jennifer LeClaire and Jason Rushin. *Behavioral Analytics for Dummies*. Wiley, 2010.
9. Marcelo Maia, Jussara Almeida, and Virgílio Almeida. Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems*, SocialNets '08, pages 1–6, New York, NY, USA, 2008. ACM.
10. Jenny Preece. *Online Communities - Designing Usability, Supporting Sociability*. John Wiley & Sons, Ltd, 2000.
11. Jenny Preece. Sociability and usability in online communities: Determining and measuring success. *Behavior and Information Technology Journal*, 20(5):347–356, 2001.
12. Vladimir Soroka. Invisible participants: how cultural capital relates to lurking behavior. In *Proceedings of the 15th International Conference on World Wide Web*, pages 163–172, 2006.
13. Jim Sterne. *Social Media Metrics: How to Measure and Optimize Your Marketing Investment*. John Wiley & Sons, 2010.
14. Jan-Willem Strijbos and Maarten F. De Laat. Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior*, 26(4):495 – 505, 2010. Emerging and Scripted Roles in Computer-supported Collaborative Learning.
15. Don Tapscott and Anthony Williams. *Wikinomics*. Atlantic Books, 2007.
16. Nielson Wire. Led by facebook, twitter, global time spent on social media sites up 82% year over year, 2010, http://blog.nielsen.com/nielsenwire/global/led-by-facebook-twitter-global-time-spent-on-social-media-sites-up-82-year-over-year/.
17. Tian Zhu, Bai Wang, Bin Wu, and Chuanxi Zhu. Role defining using behavior-based clustering in telecommunication network. *Expert Syst. Appl.*, 38(4):3902–3908, 2011.