What makes communities tick? Community health analysis using role compositions

How to cite:

For guidance on citations see FAQs.
What makes communities tick?
Community health analysis using role compositions

Matthew Rowe and Harith Alani
Knowledge Media Institute
The Open University
Milton Keynes, UK
Email: \{m.c.rowe, h.alani\}@open.ac.uk

Abstract—Today’s Web is social and largely driven by a wide variety of online communities. Many such communities are owned and managed by businesses that draw much value from these communities, in the form of efficient and cheaper customer support, generation of new ideas, fast spreading of information, etc. Understanding how to measure the health of online communities and how to predict its change over time, whether to better or to worse health, is key to developing methods and policies for supporting these communities and managing them more efficiently. In this paper we investigate the prediction of community health based on the social behaviour exhibited by their members. We apply our analysis over 25 SAP online communities, and demonstrate the feasibility of using behaviour analysis to predict change in their health metrics. We show that accuracy of health prediction increases when using community-specific prediction models, rather than using a one-model-fits-all approach.

I. INTRODUCTION

Online communities form one of the primary pillars of the social Web. Many businesses nowadays invest in supporting an online community to increase customer loyalty, brand awareness, spread of word-of-mouth, and idea generation [1], [2]. Understanding the socio-technical parameters and dynamics that drive the evolution of these communities has been the topic of many scientific works in recent years, and forms one of the core objectives of the Web Science and Social Computing research fields.

Managers and owners of such communities need to continuously monitor and maintain the health of their communities to ensure their longevity and value generation. Community health is subject to numerous internal and external factors, such as topic popularity, competing communities, community policies and services, and behaviour of its members. Although a comprehensive model for measuring community health does not exist yet, many metrics have been used or proposed for measuring this health (see Section II).

One of the new challenges associated with managing online communities is the ability to predict change in community health. Providing community managers and owners with such early warnings can support their decision making to safeguard their communities. In this paper we focus on detecting and understanding the correlation between community social behaviour and its overall health. The assumption is that the type and composition of behaviour roles exhibited by the members of a community (e.g. experts, novices, initiators) can be used to forecast change in community health. Hence the main research question this work is targeting is Can we accurately and effectively detect positive and negative changes in community health from its composition of behaviour roles? Subsequently, we make the following main contributions in this paper:

1) Identification and implementation of a range of online community health indicators from the literature.
2) A role mining approach to measure the behaviour role composition of a significant business community platform.
3) Analysis of community health using role compositions to detect health changes and forecast community health.
4) Demonstration of the need for community-specific models to achieve higher prediction accuracy of health.

This paper is structured as follows: Section II describes related work in community health assessment and in user behaviour and role analysis. Section III details the dataset used for our experiments from the SAP Community Network (SCN). Section IV presents the health indicators we implemented and their relation to existing work, and section V presents our approach in deriving the role compositions of online communities. Section VI presents our experiments and findings, and section VII discusses our results in the wider context of community health analysis. Section VIII finishes the paper with our conclusions.

II. RELATED WORK

A. Measuring Community Health

Defining ‘health’ within the context of online communities involves assessing the ‘condition’ or ‘state’ of a given community at a given point in time. Various metrics have been proposed for quantifying and measuring health [3], with multiple health indicators being commonly used to provide an overview of a community’s current condition [4].

For measuring community health and success, Preece [5] propose using measures of sociability (number of participants, number of posts, members’ satisfaction, and reciprocity) and usability. Similarly, Lu et al. [6] use the number of active users for measuring of community health. Lithium Technologies [7] has combined multiple indicators into a single metric: ‘Community Health Index’, which includes number of new members, post count, page views, time to responses, and number of times activity exceeds a critical threshold.
In addition to number of active users and number of posts, Marin et al [4] also calculate the network cohesion of the community, which they found to be positively correlated with increases in participants who stimulate discussions. According to [8], Social Capital provides an assessment of community health by gauging the extent to which users can be connected to one another (e.g. network cohesion). Chen et al. [9] argue that the ‘sense of virtual community’ that users feel in an online community correlates with community loyalty, where this sense is heightened when the community is better connected and social capital is increased. Lin et al. [10] also use member loyalty to quantify success of online communities. Work by Iriberri and Leroy [11] found that successful communities had a high post count and their users were close to one another, thereby allowing relationships to develop.

In summary, existing health metrics cover four factors: loyalty (retention of users), participation (active contributors), activity (number of posts) and social capital (connectivity). In this paper we propose Churn Rate as a measure of loyalty - thereby gauging the proportion of the community that leaves - and the graph-based metric Clustering Coefficient, that accounts for community cohesion and the dynamics of information flow through a community. Our focus is not on identifying the most effective health metrics, but rather on correlating and predicting their change based on user behaviour.

B. Assessing Behaviour and Role Compositions

The behaviour that users exhibit within online communities is associated with their actions and interactions with other community users while the role that a user assumes is the label associated with a given type of behaviour. Roles are identified by a set of behaviours, (or behaviour dimensions), such as engagement, contribution, popularity, participation, etc. The general procedure to model these behaviours in online communities is by translating them into measurable behavioural features from the social network graph with an associated intensity level (e.g. low, medium, high). For example Hautz et al. [12] measure in-degree, out-degree and number of content uploads, Nolker and Zhou [13] measure; spreading knowledge, and length and volume of conversations, while [14], [15] measure social network features such as: in-degree, out-degree, in-length (total duration of calls received), out-length (total duration of calls made to others), and more complex social network graph measures such as InnerPageRank and OuterPageRank.

A wide number of studies from different research communities (sociolinguistics, social psychology, ethnography communication, etc.) have aimed to capture the set of roles and behaviours present in online communities: captain, pillar, moderator and mediator [16], [17], celebrity [18], popular initiator, popular participant and joining conversationalist (who have medium initiation and participation) [19], lurker (consume but not contribute) [16–18], and content consumer [20], grunt and taciturn [19] (contribute with low intensity). Although there is no standard subset of roles and associated behaviours across communities, there is a clear tendency in the literature to use certain behaviours like: popularity, engagement, contribution, initiation and focus.

According to [21] role mining can be divided into two general methodological approaches: interpretive analysis and structural analysis. Interpretive analysis approaches (e.g. [18]) employ methods like ethnography, content analysis, and surveys to capture behaviours and relations within groups. Structural analysis approaches [12]–[15], [19] use formal methods like clustering or network structure analysis to identify relevant roles within the community.

Nolker et al. [13] assume the existence of: a) roles identified from the literature (leaders and motivators) and; b) a set of behavioural features identified from the social network graph, and then associate the features and their intensity level (high, moderate, low) to the preselected roles. In [12] the assumed behavioural features are used to mine eight different roles: motivator, attention attractor, idea generator, passive user, etc. [15], [19] assume a set of initial behavioural features and then perform cluster analysis to identify the set of roles that emerge from the community. Each cluster approximately corresponds to one role. In this paper we describe a role identification step that uses a maximum-entropy decision tree to empirically generate the role labels without the need for a pre-conceived role collection.

III. Dataset: SAP Community Network (SCN)

To ground our work we use the SAP Community Network (SCN) for role identification and role composition analysis. The SAP Community Network is a collection of online forums hosted by SAP in which users can discuss SAP-related issues including software development, SAP products and usage of SAP tools. Points can be awarded by question posters to the answers that they deem to be the best. Over time users therefore build up a reputation on the platform as being knowledgeable about certain subjects by their ability to provide highly rated answers.

We were provided with a subset of the SCN covering 33 communities, listed in Table I. The dataset contained 95,200 threads, 421,098 messages of which 78,690 were allocated points, and 32,942 users. As the post counts within Table I indicate, there is a large variance in activity between the communities, with community 264 having the highest number of posts with over 85K, and community 486 having the lowest with only 7 posts.

IV. Community Health Indicators

Existing work towards measuring community health demonstrates the multi-faceted nature of its assessment. The differences between online communities based on their type and nature means that the importance of one health measure for one community may differ from another [3]. As a consequence we adopt four health indicators, covering the four factors that we have identified from the literature: loyalty, participation, activity and social capital. These are explained below.
TABLE I
COMMUNITIES AND THEIR IDS WITHIN THE SCN DATASET

| ID | Name                                           | Posts | Unique
|----|------------------------------------------------|-------|-------
| 1  | Best Practice and Benchmarking                 | 367   | 163   
| 161| SAP Business One Integration Technology        | 3163  | 812   
| 197| Business Process Project General Discussion     | 7464  | 2699  
| 198| Business Process Modelling Methodologies        | 950   | 305   
| 200| Organizational Change Management                | 230   | 47    
| 201| Standards                                       | 367   | 163   
| 210| Analytics                                      | 488   | 170   
| 226| SAP Discovery System for Enterprise SOA          | 1105  | 408   
| 252| SAP Business One E-Commerce and Web CRM         | 4487  | 1389  
| 256| Governance, Risk and Compliance                  | 19092 | 4279  
| 264| SAP Business One Core                           | 85057 | 17838 
| 265| SAP Business One Product Development            | 2624  | 1127  
| 270| Financial Performance General                   | 8904  | 2482  
| 281| Sustainability                                  | 190   | 42    
| 319| SAP Forum                                      | 483   | 214   
| 338| SAP Business One Reporting & Printing           | 30854 | 7344  
| 344| SAP Business One Partner Solutions (Add-ons)    | 665   | 104   
| 400| International Financial Reporting Standard (IFRS)| 291   | 78    
| 411| Operational Performance Management General      | 299   | 89    
| 412| B2B' Planning & Consolidations: SAP NetWeaver   | 14439 | 5462  
| 413| B2B' Planning & Consolidations: Microsoft Platform | 18859 | 4245  
| 414| SAP Strategy Management                         | 1954  | 399   
| 419| SAP Business One - SAP Add-ons                 | 19656 | 3899  
| 420| SAP Business One Training                       | 16113 | 3222  
| 429| SAP Business One Training                       | 481   | 119   
| 44  | Process Integration                             | 27566 | 4907  
| 466| Green IT                                       | 90    | 8     
| 470| Manufacturing Execution (ME)                    | 1442  | 301   
| 482| ASAP Methodology and Project Management         | 138   | 36    
| 485| OSI Standards and SAP                           | 44    | 14    
| 486| Enterprise Social Systems                       | 7     | 3     
| 50  | ABAP, General                                  | 54718 | 13262 
| 56  | SAP Business One SIK                           | 79090 | 15802 |

A. Churn Rate

The first health indicator concerns community loyalty by measuring the proportion of users who are churners in a given time segment. This indicator is similar to criteria described in [10] which measures the propensity of users to remain active. We define this indicator formally as the Churn Rate, where \( \Upsilon \) is the number of users who have posted in the community and \( \Upsilon_c \) is the number of users who have posted in the community for the last time: \( \text{Churn Rate} = \frac{\Upsilon_c}{\Upsilon} \).

B. User Count

Participation is often regarded as a key indicator of community activity and success and is normally quantified by the number of users [4]–[6]. We define User Count as the number of users who posted at least once: \( \text{User Count} = \Upsilon \).

C. Seeds / Non-seeds Proportion

The third factor pertains to the activity that communities experience. Our previous work [22] used the number of posts within a community as a basic signifier of health, similarly to [4], [5] and [7]. Here we introduce a new metric that covers community activity by measuring the proportion of seed posts \( (P_s, \text{thread starters that yield at least one reply by another user}) \) to non-seed posts \( (P_n, \text{thread starters that yield no replies}) \). Our intuition is that in a support-oriented community a signifier of good health is a high ratio of seeds to non-seeds, as this demonstrates activity and engagement, where fewer thread starters fail to get a reply. We formulate the Seeds / Non-Seeds Proportion as: \( \text{SeedsToNonSeeds} = \frac{P_s}{P_n} \).

D. Clustering Coefficient

Our fourth health indicator is related to the social capital factor, similarly to [8], [9]. These works consider the average network degree of users in the graph to gauge, on average, how well connected users are. We go one step further by measuring the Average Network Clustering Coefficient of users within the community’s largest connected component. This measure gauges the cohesion of the community and the extent to which it forms a clique, it also allows for the assessment of possible information flow through the community as a high clustering coefficient indicates that there are many possible paths for information to pass through the network.

In order to build the network’s directed graph \( G = (V,E) \) we take all users who participated within the community during a time window, thereby returning the set of vertices \( V \), and form the set of edges \( E \) between users such that an edge denotes a reply from one user to another - e.g. \( e_{ij} \) where \( v_i \) replied to \( v_j \). We then identify the largest connected component \( (G_{cc} = (V_{cc},E_{cc}) \) in the graph and measure the local clustering coefficient \( (C_i) \) of each node \( v_i \in V_{cc} \) as follows: \( C_i = \frac{2|E_i|}{k_i(k_i-1)} \). To derive the average network clustering coefficient we take the average local clustering coefficient of each user in the graph. This provides our measure of social capital that we will refer to hereafter as Clustering Coefficient for brevity.

V. Measuring Role Compositions

Online communities exhibit differing behavioural patterns as users interact with one another in a disparate manner and participate within the community in a unique way. Understanding the behaviour of community users and how that relates to community health indicators, could provide community managers with information of healthy and unhealthy behavioural traits found in their communities. In this section we describe the numerical representation of community users’ behaviour, the inference of a community’s role composition (i.e. the percentage breakdown of users assuming different roles) and the mining of roles for a specific platform, in our case SCN.

A. Modelling and Measuring User Behaviour

In accordance with related work (Section II), we describe six general dimensions for measuring user behaviour. Similarly to [12], we ground each behaviour dimension with a specific feature that could be measured on the platform of our SAP community dataset:

1) Focus Dispersion: the forum entropy of a user, where a high value indicates that the user disperses his/her activity across many SAP forums, while a low value indicates that the user concentrates his/her activity in a few forums. Let \( F_{v_i} \) be all the forums that user \( v_i \) has posted in and \( p(f_j|v_i) \) be the conditional probability of \( v_i \) posting in forum \( f_j \) - we can derive this using the post distribution of the user - therefore we define the Forum Entropy \( (H_F) \) of a given user as:

\[
H_F(v_i) = - \sum_{j=1}^{|F_{v_i}|} p(f_j|v_i) \log p(f_j|v_i)
\]  

2) Engagement: the proportion of users that the user has replied to. A larger value indicates that the user has
roles and derive the community composition.

Fig. 1. Overview of the approach to analyse user behaviour, label users with roles and derive the community composition.

3) **Popularity**: the proportion of users that have replied to the user. A larger value indicates that the user is popular within the platform. Let $\Upsilon$ be the total number of users and $\Upsilon_{in,i}$ the users that have replied to $v_i$, then the popularity of a user is defined as $\Upsilon_{in,i}/\Upsilon$.

4) **Contribution**: the proportion of thread replies that were created by the user. This measures the extent to which the user contributes replies to threads. Let $P_r$ be the total set of replies authored by all users and $P_{r,i}$ be the set of replies authored by $v_i$, we define the contribution of $v_i$ as $P_{r,i}/P_r$.

5) **Initiation**: the proportion of threads that were started by the user. This gauges how much the user instigates discussions and asks questions. Let $P_s$ be set of thread starters authored by all users and $P_{s,i}$ be the set of thread starters authored by $v_i$, we define the initiation of $v_i$ as $P_{s,i}/P_s$.

6) **Content Quality**: the average points per post awarded to the user. This provides a measure of expertise of the user. Let $P_{i,j}$ be the set of posts authored by $v_i$ and points$(p)$ to be a function that returns the points awarded to post $p$, we define the content quality of $v_i$ as:

$$\frac{\sum_{j=1}^{P_{i,j}} \text{points}(p)}{|P_{i,j}|}$$  \hspace{1cm} (2)

**B. Inferring Roles**

Our approach to derive the role composition functions by taking the users who participated in the community over a given period of time and inferring the role of each user in the community, thereby providing a measure of the role composition - e.g. 10% roleA, 20% roleB, etc. We can then derive the role composition repeatedly over incremental time periods and capture how the composition changes in the community.

The third stage of our approach compiles the rule base from the **Skeleton Rule Base**, which is platform-dependent and set according to the analysis that is to be performed. It contains a single rule for each behaviour role. The antecedent of each rule contains a mapping between a feature and the level that that feature should be:

- **popularity=low**, **initiation=high** -> **roleA**
- **popularity<0.5**, **initiation>0.4** -> **roleA**

The rules are constructed from the **Skeleton Rule Base** and the bins derived for each feature such that level boundaries are set within the rule:

- **popularity<0.5**, **initiation>0.4** -> **roleA**

The final stage of the approach is to apply the rules to the community’s users and infer each user’s role. Once every community user has been labelled with a role we can then derive the community’s composition by the percentage of users that each role covers. The process of deriving the composition of a community can be repeated over time to detect changes in how the community evolves.

**C. Mining Roles from SAP Community Network**

Compilation of the **Skeleton Rule Base** for our approach is platform-dependent. The roles present on one platform differs from another. We therefore need to mine the roles that are present on the SAP Community Network and compile the **Skeleton Rule Base** accordingly. We select a tuning segment of the dataset, choosing the first-6 months of 2008 over which to mine roles and then use the remaining data that follows this tuning segment as our later analysis window.

1) **Discovering Correlated Behaviour Dimensions**: The aforementioned behaviour dimensions, although intended to be distinct, may in fact be strongly correlated, thus reducing their value in describing unique behaviours. Hence we need to detect and remove these correlated dimensions thereby reducing the dimensionality of our dataset and aiding discrimination between roles. To do this we built the above behaviour dimensions, assigned features for each user in our tuning dataset and then measured the Pearson correlation coefficient ($r$) between each dimension. In order to filter out the highly correlated dimensions that were significant we ran the Pearson correlation coefficient significance test where $r > 0.75$. We found that engagement, contribution and popularity were all highly correlated with one another. Therefore we removed the first two dimensions from our dataset, resulting in the following dimensions remaining: focus dispersion, initiation, content quality and popularity.

2) **Clustering Users**: Following the filtration of the initial dimensions we use the remaining dimensions to cluster users,
thereby separating users based on their behaviour and discovering distinct roles on the platform. We ran three different unsupervised clustering algorithms: Expectation-Maximization (EM), K-means and Hierarchical Clustering, over the 6-months’ tuning segment. The model selection phase not only requires choosing the correct clustering method but also selecting the optimum number of clusters to use - providing this value as a parameter $k$. To judge the best model - i.e. cluster method and number of clusters - we measure the cohesion and separation of a given clustering as follows: For each clustering algorithm $(\Psi)$ we iteratively increase the number of clusters $(k)$ to use where $2 \geq k \geq 30$. At each increment of $k$ we record the silhouette coefficient produced by $\Psi$, this is defined for a given element $(i)$ in a given cluster as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$  \hspace{1cm} (3)

Where $a_i$ denotes the average distance to all other items in the same cluster and $b_i$ is given by calculating the average distance with all other items in each other distinct cluster and then taking the minimum distance. The value of $s_i$ ranges between $-1$ and 1 where the former indicates a poor cluster where distinct items are grouped together and the latter indicates perfect cluster cohesion and separation. To derive the silhouette coefficient $(s(\Psi(k)))$ for the entire clustering we take the average silhouette coefficient of all items. We find that the best clustering model and number of clusters to use is K-means with 11 clusters. We found that for smaller cluster numbers $(k = [3, 8])$ each clustering algorithm achieves comparable performance, however as we begin to increase the cluster numbers K-means improves while the two remaining algorithms produce worse cohesion and separation.

3) Deriving Role Labels: Provided with the most cohesive and separated clustering of users we then derive role labels for each cluster. Role label derivation first involves inspecting the dimension distribution in each cluster and aligning the distribution with a level mapping (i.e. low, mid, high). This enables the conversion of continuous dimension ranges into discrete values which our rule-based approach requires in the Skeleton Rule Base. To perform this alignment we assess the distribution of each dimension and derive boundary points for the three feature levels using an equal-frequency binning approach. The distribution of each dimension is shown in Figure 2 for each of the 11 induced clusters together with the level boundaries. We assess the distribution of each feature for each cluster against the levels derived from equal-frequency binning of each feature, thereby generating a feature-to-level mapping. This mapping is shown in Table II where certain clusters are combined together as they have the same feature-to-level mapping patterns - i.e. 2,5 and 8,9.

In order to derive the role labels for each cluster we use a maximum-entropy decision tree to divide the clusters into branches that maximise the dispersion of dimension levels. Figure 3 shows the separation of the clusters from a complete grouping into a single cluster, or merged clusters in the case of 2,5 and 8,9, in each leaf. To perform the separation at a given decision node, we measure the entropy of the dimensions and their levels across the clusters, we then choose the dimension with the largest entropy. This is defined formally as:

$$H(dim) = -\sum_{level} p(level|dim) \log p(level|dim)$$  \hspace{1cm} (4)

![Figure 2: Boxplots of the feature distributions in each of the 11 clusters. Feature distributions are matched against the feature levels derived from equal-frequency binning](image)

![Figure 3: Maximum-entropy decision tree used to segment the clusters into minimal-distance paths. The paths are used to generate the role labels for each respective cluster](image)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Dispersion</th>
<th>Initiation</th>
<th>Quality</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>1</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>4</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>H</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>7</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
</tr>
</tbody>
</table>

We perform this process until single clusters, or the previously merged clusters, are in each leaf node and then use the path to the root node to derive the label. For instance, for cluster 0 the path from the root node to the leaf node is quality=high, dispersion=low, initiation=medium, thereby deriving the role label Focussed Expert Participant for the cluster. In the label, focussed describes the focus dispersion of the role - i.e. it is low and therefore not distributed, expert...
Rule Base

labels for each cluster from Table II, these role labels and their labels using dimension splits we produced the following role and an answerer. Based on this method of deriving the role being high given the quality of their answers - and describes the level of expertise that a user will have - i.e. being high given the quality of their answers - and participant denotes the extent to which this role starts threads - i.e. being in the middle in this case and thus being both an initiator and an answerer. Based on this method of deriving the role labels using dimension splits we produced the following role labels and their feature-to-level mappings are used to compile our Skeleton Rule Base:

- **1 - Focussed Novice**: this user is focussed within a few select forums but does not provide good quality content.
- **0 - Focussed Expert Participant**: provides high quality answers but only within select forums. They also mix between asking questions and answering them.
- **6 - Knowledgeable Member**: medium-level expertise (neither an expert nor a novice) and medium popularity
- **10 - Knowledgeable Sink**: user who has medium-level expertise but many from the community reply to them - hence a sink. Differs from cluster 6 in terms of popularity.
- **4 - Focussed Expert Initiator**: similar to cluster 0 in that this type of user is focussed on certain topics and is an expert on those, but to a large extent starts discussions and threads, indicating that his/her shared content is useful to the community
- **2, 5 - Mixed Novice**: is a novice across a medium range of topics
- **8, 9 - Mixed Expert**: medium-dispersed user who provides high-quality content
- **7 - Distributed Novice**: participates across a range of forums but is not knowledgeable on any topic
- **3 - Distributed Expert**: an expert on a variety of topics and participates across many different forums

VI. EXPERIMENTS

Exploring the relation between role compositions and health identifies patterns that explain the relation between a degradation or improvement in a community’s health and the behaviour of its members. In this section we demonstrate the efficacy of this approach by first assessing the patterns that relate role compositions and health indicators before demonstrating how role composition information can be used to predict changes in community health.

A. Experimental Setup

To demonstrate the utility of our approach we analysed 25 of the 33 SAP communities from 2009 through to 2011, removing 8 communities with <100 threads in the analysis window. Figure 4 shows how our dataset was divided into the tuning section - i.e. the first half of 2008 in which we derived our clusters and aligned them to roles (as described in Section V-C) - and the analysis section. We began with 1st January 2009 as our collect date by taking a feature window 6 months prior to this date (going back to the 2nd half of 2008) in which we measured the behaviour dimensions for each community’s users. In order to gauge the role composition in a community over time we move our collect date on one week at a time and use the 6-months prior to this date as our feature window. As Figure 4 demonstrates we repeat this process until we reach 2011.

The role compositions of individual communities are measured in a time-ordered, iterative manner, with the health indicators being measured in the same way. For each community we use the same feature window of 6-months to measure the Churn Rate, User Count, Seeds to Non-seeds Proportion and Clustering Coefficient. For the latter of these features we used the Jung Graph Framework\(^1\) to construct the user network from replies within the community, derive the strongest connected component in the community and then measure the clustering coefficient. To demonstrate the application of such analysis we performed two distinct experiments:

1) **Health Indicator Regression**: induces linear regression models for each of the analysed communities by using the role composition in each community as predictors for the health indicator, thereby performing four individual analyses, one for each of the four health indicators. We perform Principal Component Analysis (PCA) by extracting the regression coefficients from each community’s model and using those as a community motif (or vector) for plotting in the PCA space, thereby observing which communities are clustered together and which are distinct. By selecting 6 communities from the PCA plots, 3 within the central cluster and 3 that are outliers, we then assess the coefficients in the linear regression models and their coefficient of determination \((R^2)\) to judge the goodness of fit.

2) **Health Change Detection**: performs a binary classification task to detect changes in community health from one time step to the next, exploring: Can we accurately and effectively detect changes to communities that could result in bad health? This experiment is formulated such that at time step \(t = k + 1\) we predict whether the health indicator of a community has increased or decreased since \(t = k\). We create an instance for each time step \((t = k+1)\) such that the features are the 9 roles with their composition proportions as values and the class label as either positive for an increase in the health indicator or negative for a decrease. We divide the dataset for each community up into an 80:20 split for training and testing, while maintaining time-ordering, and apply the logistic regression classifier to detect changes. We

\(^1\)http://jung.sourceforge.net/
report the Matthews Correlation Coefficient (MCC), to demonstrate the improvement over the random classifier as our baseline, and precision, recall, f-measure (setting $\beta = 1$) and the area under the Receiver Operator Characteristic Curve (AUC).

B. Results: Health Indicator Regression

Our goal for this first experiment was to assess the differences between communities as to how role compositions and health indicators are related, thereby identifying different health composition patterns. Therefore we begin by focussing on a common health pattern for SCN before investigating patterns in outlier communities.

1) A Common Health Composition Pattern: The PCA plot in Figure 5 indicates that for each health indicator there are common health composition patterns for certain communities, given the consistent central cluster in each plot close to the origin. There are also several outlier communities which are consistently separate across all health indicators. We begin by first analysing the health composition patterns of three randomly selected forums that are contained within this central cluster across the four health indicators: 252 (SAP Business One E-Commerce & Web CRM), 412 (Business Planning & Consolidations: SAP NetWeaver), 414 (SAP Strategy Management), the coefficients for which are presented in Table III - together with three forums that are outliers in the PCA space which we later explore in greater detail. We hypothesise that because each of these three central forums appear close to one another in the PCA space that they will exhibit similar health composition patterns, and that as they are similar we can learn a general pattern that explains the relation between role compositions and health indicators for the majority of SCN communities.

Beginning with Churn Rate and inspecting the regression coefficients of roles in Table III, we find that each of the forums within the central cluster exhibit a slightly different behaviour. For instance, we find that the coefficients and signs differ between the forums for Focussed Expert Participant and Mixed Expert, suggesting that each forum has idiosyncratic dependencies between the churning of its users and the appearance of experts who distribute their activity across forums. We find, however, that a decrease in Focussed Expert Initiators is associated with an increase in Churn Rate across the three communities, indicating that, in general, communities on SAP experience the leaving (perhaps moving to other forums) of users if expert users who start discussions leave.

For the User Count we once again find unique health composition patterns in each of the communities within the central group. An increase in Focussed Expert Initiators is correlated with an increase in users for 252, while the contrary is true for 412 and 414. The models do indicate common patterns across the communities for certain roles, for Knowledgeable roles - i.e. Member and Sink - we find that a decrease in either role is associated with an increase in user counts. The third health indicator, Seeds to Non-seeds Proportion, indicates that there are common patterns across the central communities as a decrease in Focussed Expert Participants and Initiators, and Distributed Experts is associated with an increase in the proportion of seeds to non-seeds. This is expected, as the two former roles are synonymous with the creation of content and not just merely replying.

Measuring the interactivity of the community through the Clustering Coefficient we see marked differences between the three communities, in Table III. For instance, for 252 an increase in Knowledgeable Members is associated with an increase in the clustering of the network, while a decrease in this role for 412 and 414 is associated with increased community interaction, we also see a similar effect for Focussed Expert Initiators. These findings indicate that possible information flow through a community’s network is affected by different roles in different communities.

The findings from analysing the heath composition patterns for each of the four health indicators negate our earlier hypothesis that a general pattern exists for describing the relation between role composition and community health. Although there are certain cases where a role is common across a given indicator - e.g. Churn Rate and Focussed Expert Initiators - these cases are spurious.

2) Idiosyncratic Health Composition Patterns: The non-existence of a common health composition pattern suggests that each community is unique and that the relation between community behaviour and health is down to the environment in which the users are participating. As Figure 5 indicates, the central cluster for each of the four health indicators is surrounded by several outlier communities, each of which exhibit unique, and possibly extreme, health composition patterns. We now explore the idiosyncratic nature of three of these communities by examining their regression models in more detail, focussing on: 353 (SAP Business One Reporting & Printing), 419 (SAP Business One System Administration) and 50 (ABAP, General), the regression coefficients for which are presented in Table III.

We find that for Churn Rate the coefficients in the models of the outlier communities are more extreme than for those communities within the central PCA cluster. For instance a large increase in Focussed Novices is associated with an increase in the churning of community users, while for Distributed Novices the forums differ drastically with large differences in the magnitudes of the coefficients. Interestingly, despite each community being an outlier in the PCA space, we find that for the User Count there are common relations between roles: Focussed Novice, Mixed Novice and Knowledgeable Sink, and increases in the user counts, although the coefficients differ for each of the communities.

For the third health indicator, Seeds to Non-seeds Proportion, we find distinct, and extreme, patterns between the forums. A decrease in the Distributed Experts is associated with an increase in the proportion of seeds to non-seeds for the three central communities, which is the same for the outlier communities.

2N.b. we only comment on roles that are statistically significant within the regression model.
353, but not for 50. The coefficients also demonstrate the divergent patterns in the outlier communities, particularly for the role of Distributed Novices in forum 50. For the Clustering Coefficient, however, we observe similarity between the three outliers in terms of the relation between the Knowledgeable Sink and increases in this health measure. Once again the magnitudes of the coefficients largely differ, however the signs remain the same, indicating that users who assume these roles help to bind the community together, thereby adding to the social capital of the community and the flow of information through users in these forums.

C. Results: Health Change Detection

The results from the regression analysis demonstrate the idiosyncratic nature of the SCN communities and the disparity between their health composition patterns. This suggests that a general model describing the health composition of every community on the platform would perform poorly in comparison with a model that describes a single community. To test this we detected changes in health indicators between time steps based on a community’s role composition. Table IV shows the performance of a logistic regression model induced for the different health indicators. We tested the models when trained (a) across the entire platform and (b) per-forum (using data from only one forum). We find that for each health indicator learning community-specific patterns outperforms the platform-level models. This demonstrates the need to assess individual communities and understand what works best in those forums given the dynamics at play.

Focussing on the MCC values we find that the per-forum models significantly outperform the random classifier - using the sign test to assess the significance of the results - while only the platform-level model for Clustering Coefficient significantly outperforms the baseline. For the f-measure levels we achieve relatively similar levels across the four health indicators for the per-forum models.3

Focussing on the forums within the central cluster and as outliers in the earlier PCA plot (Figure 5) we present the results from each forum’s detection model in Table IV(c). We anticipated that the extreme health composition patterns that we found within the previous experiment would render the induction of a classification model difficult, given the large variation in the earlier coefficients. However, as the

---

### TABLE III

**COEFFICIENTS AND SIGNIFICANCE TEST RESULTS FOR PER-FORUM LINEAR REGRESSION MODELS. EACH MODEL PREDICTS THE HEALTH INDICATOR USING THE COMMUNITY’S ROLE COMPOSITION.**

<table>
<thead>
<tr>
<th>(a) Churn Rate</th>
<th>(b) User Count</th>
<th>(c) Seeds / Non-seeds Prop</th>
<th>(d) Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Role</strong></td>
<td><strong>Central</strong></td>
<td><strong>Outliers</strong></td>
<td><strong>Central</strong></td>
</tr>
<tr>
<td></td>
<td>252</td>
<td>412</td>
<td>353</td>
</tr>
<tr>
<td>Focussed Expert Initiator</td>
<td>27.21***</td>
<td>-37.43***</td>
<td>43.59***</td>
</tr>
<tr>
<td>Focussed Expert</td>
<td>28.508***</td>
<td>-26.812***</td>
<td>45.197***</td>
</tr>
<tr>
<td>Mixed Expert</td>
<td>25.264***</td>
<td>-31.364***</td>
<td>42.525***</td>
</tr>
<tr>
<td>Knowledgeable Initiator</td>
<td>28.508***</td>
<td>-26.812***</td>
<td>45.197***</td>
</tr>
<tr>
<td>Knowledgeable Sink</td>
<td>1.311***</td>
<td>3.532***</td>
<td>6.363***</td>
</tr>
<tr>
<td>Knowledgeable</td>
<td>3.301***</td>
<td>7.519***</td>
<td>11.348***</td>
</tr>
</tbody>
</table>

---

3We also verified the performance of the tested models against the null logistic model and found performance to be significantly better at α < 0.001.
results indicate using the role composition information, even in the outlier communities, provides sufficient information to outperform the random guesser baseline for all health measures except the User Count for forum 353. We also find that for the 412 and 414 central forums we achieve poorer performance than the baseline for the User Count and Clustering Coefficient.

### TABLE IV

**PERFORMANCE OF DETECTING HEALTH CHANGES USING A LOGISTIC REGRESSION MODEL INDUCED ACROSS THE ENTIRE PLATFORM (FIGURE IV(A)), PER-FORUM (FIGURE IV(B)) AND FOR SPECIFIC CENTRAL AND OUTLIER FORUMS (FIGURE IV(C)). IN THIS LATTER CASE WE REPORT THE MATTHEWS CORRELATION COEFFICIENT AND THE F1 SCORE.**

<table>
<thead>
<tr>
<th>(a) Platform</th>
<th>Class</th>
<th>MCC</th>
<th>Pre</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Count</td>
<td>0.033</td>
<td>0.89</td>
<td>0.86</td>
<td>0.025</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Seeds/Non-seeds</td>
<td>0.078</td>
<td>0.62</td>
<td>0.56</td>
<td>0.065</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>0.077</td>
<td>0.59</td>
<td>0.56</td>
<td>0.051</td>
<td>0.56</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>(b) Per-forum</th>
<th>Class</th>
<th>MCC</th>
<th>Pre</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Count</td>
<td>0.199</td>
<td>0.88</td>
<td>0.86</td>
<td>0.022</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Seeds/Non-seeds</td>
<td>0.185</td>
<td>0.87</td>
<td>0.85</td>
<td>0.047</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>0.106</td>
<td>0.68</td>
<td>0.65</td>
<td>0.056</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>(c) Forum Specific Results. MCC / F1</th>
<th>Class</th>
<th>Central</th>
<th>414</th>
<th>412</th>
<th>419</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Churn Rate = Increase, User Count = Decrease, Seeds to Non-seeds Proportion = Decrease and Clustering Coefficient = Decrease.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 shows the Receiver Operator Characteristic Curves for the per-forum logistic regression models. The curves indicate we can outperform the random classifier for all forums apart from 5 for the Churn Rate and User Count, all but 6 for the Seeds to Non-seeds Proportion and all but 8 forums for the Clustering Coefficient, demonstrating the variation in performance that we achieve across the communities. The forums that we consistently performed poorly on were 265 (SAP Business One Product Development) and 319 (Best Practice and Benchmarking), achieving worse performance than the random baseline for all health indicators. The reduction in accuracy could be caused by the roles detected in the community not fitting its nature, where instead, conversation and discussion-driven roles are assumed - similar to the roles used in our previous work [22]. In general, the results demonstrate the effectiveness of using role composition information to detect when a forum’s health will degrade. Using this information the managers of such forums can now identify when the users of their communities change their behaviour in a way that could negatively affect the health of their community.

## VII. DISCUSSION AND FUTURE WORK

The findings from our analyses have identified interesting health composition patterns and the lack of a global composition pattern for the entirety of SCN. Although it has been argued that the choice of metrics should be dependent on community objectives [3], there are hardly any studies that demonstrate this. SCN is a Question Answering platform, and hence its objective is providing responses to questions. When inspecting the patterns learnt for the Seeds to Non-seeds Proportion indicator we find that across communities a decrease in Focussed Experts is correlated with an increase in the indicator. This implies that users knowledgeable on specific topics actually increase the number of unanswered posts. It could be that the questions asked by such experts users cannot be solved by most community users.

Marin et al. analysed 11 Linux support communities and found a global positive correlation of code users and network cohesion [4]. They defined core users as those whose out-degree is greater than the community average out-degree plus one standard deviation. In our work we account for a multitude of roles that users may assume, rather than just one set. Our analysis also showed that such global patterns are harder to identify when the communities differ in topic and nature. For instance an increase in Knowledgeable Member for 252 and 353 was linked with an increase in the Clustering Coefficient, while being the converse for 412 and 414, we also found that Knowledgeable Sink was associated with increased social capital for 252, 353, 419 and 50. This latter role is closest to the notion of core users that we find on SCN given its high Popularity.

In this paper we focused our study on 4 popular health indicators, and several more can be added next. However, one pertinent question is whether codedependencies exist between the health indicators. Our future work will explore the correlation between these indicators. It could be the case that certain metrics are redundant, while others are salient. Similarly, it is possible that some of these metrics are more representative of health of some communities than others. Linked to this avenue of exploration is the creation of a single health metric, or index similar to [7], that provides community managers with a basic observable indicator.

Our analysis identified various correlations between behaviour roles and health metrics. Next we need to study causation dynamics to better understand the influence and sequence of events that lead to a health metric, or a behaviour role, changing.

## VIII. CONCLUSIONS

Assessing the health of online communities provides managers and operators with information about the condition of their community and how it is acting. Tying such assessments to the implicit behaviour within online communities, and the...
behavioural roles that users assume, enables health composition patterns to be gleaned from the communities and enables understanding of what works for one community and what does not.

In this work we explored the question: Can we accurately and effectively detect positive and negative changes in community health from its composition of behaviour roles? To perform this investigation we presented four health indicators derived from the literature, enabling the assessment of community health in terms of: loyalty (Churn Rate), participation (User Count), activity (Seeds / Non-seeds Proportion) and social capital (Clustering Coefficient). We described an approach to derive the role composition of online communities - i.e. the percentage breakdown of users assuming different behavioural roles - that (a) represents the behaviour dimensions of users numerically, (b) mines roles for a given platform, and (c) infers the roles that users assume and derives the community role composition as a result.5 Using data from the SAP Community Network - a Q&A community for SAP products, services and technologies - we examined the relation between the role composition in disparate communities and their health along the four indicators, demonstrating the disparate nature of communities and the difference in what makes communities tick. Armed with such insights we were then able to accurately detect changes in community health, along the four indicators, using role composition information alone. Testing platform-level models for each health indicator against the forum models we found significantly better results for the latter, indicating the lack of general health patterns across the differing communities.

ACKNOWLEDGMENT

The work of the authors was supported by the EU-FP7 project ROBUST (grant no. 257859). We would also like to thank SAP for the provision of the dataset for our analyses.

REFERENCES


5 Should the reader require a more detailed explanation of our role derivation approach then please refer to [23]