Decomposing Discussion Forums using User Roles

Jeffrey Chan
Digital Enterprise Research Institute
NUI Galway
Ireland
jkc.chan@deri.org

Conor Hayes
Digital Enterprise Research Institute
NUI Galway
Ireland
conor.hayes@deri.org

ABSTRACT

Discussion forums are a central part of Web 2.0 and Enterprise 2.0 infrastructures. The health and sustainability of forums is dependent on the information exchange behaviour of its members. Such behaviour needs to be better understood and characterised so that forums can be better managed, new services delivered and opportunities and risks detected. In this paper, we present a method for analysing user communication roles in discussion forums. We analyse the composition of several forums from a medium-sized national bulletin board in terms of these roles, demonstrating similarities between forums based on underlying user behaviour rather than topic. We suggest that analysing the evolution of role composition is an important step in developing a predictive model of forum health.

Keywords

discussion forums, roles, behaviour, social network analysis

1. INTRODUCTION

In recent years, the use of Web 2.0/3.0 applications for personal and professional purposes has grown exponentially, enabling users to readily exchange ideas, knowledge and opinions[1]. Discussion forums form a central part of Web 2.0 infrastructure, even though they been around for many years in the form of newsgroups [10]. Commerical organisations are increasingly using forums to extend their off-line technical support and manage customer relationships as part of the so-called Enterprise 2.0 methodology. Boardtracker.com estimates that there are tens of millions of public forums with an average daily posting rate of 3-4 million which, with current growth rates, is conservatively estimated to exceed 20 million posts per day by 2013 [2].

However, on-line forums require a better understanding and characterisation of member communication behaviour so that forums can be better managed, new services delivered and opportunities and risks detected. The health and sustainability of a forum is dependent on the information exchange behaviour of its members. While forums haves been the topic of several studies [7], their composition in terms of behavioural roles has up until now remained unexplored. For example, roles might include a topic instigator, who tends to initialise popular threads, or a taciturn contributor who tends to ask questions but only engages in limited conversion. Forum hosts may wish to offer topic instigators incentives to continue contributing. On the other hand, a forum might be dominated by non-communicative, non-social members, or worse still, spammers, whose effect may cause user dis-satisfaction and churn.

Manual role identification of large-scale data is time consuming and infeasible. Some work has been carried out, including the characterisation of users that are influential [2, 13] in disseminating opinions and ideas and characterising users based on their interaction patterns [1, 8]. However there has been little work in profiling the constituent features of user behaviour in forums and examining similarities between forums based on such features. Such analysis will enable host organizations to assess the health of their forums, and make decisions on resources such as moderator-ship or additional support.

This paper contributes an automated forum profiling technique to capture and analyse user behaviour, which is empirically evaluated using a medium sized national discussion board dataset. Our analysis found that forums are typically composed of eight behaviour types such as 'popular initiators', 'grunts' and 'taciturns'.

The remainder of this paper is organised as follows. Section 2, presents related work in forum and role analysis. Section 3 describes the data set and its representation. In Section 4, we describe how user roles are identified. Section 5 discusses the roles identified followed by Section 6 which decomposes forums into their role composition for comparison analysis. We discuss future work in Section 7. Section 8 concludes the paper.

2. RELATED WORK

In this section, we review some relevant literature in three related areas. The most relevant approach to ours is feature-based profiling where features are selected and used to profile users and forums. Then there is research on role equivalence in social network analysis. The third area is the specific analysis of discussion boards.

The majority of feature-based profiling involves identifying user roles either by visualisation or by utilising a set of identified user features and measures in order to capture user behaviour. Visualisation techniques include Netscan [5], AuthorLines [16] and TreeMap [15]. While these methods are powerful for exploring a small number of forums, threads and users manually, they cannot be extended to automatic analysis of user roles and forums.

A number of different features and measures have been suggested to profile user behaviour. The features used in this paper are most related to those presented by [8]. The authors analysed a users ego-centric network and the out-degree distribution along with visual representations of the network. However, they did not try to fit a distribution to the out-degree plots. Fitting the distribu-

[1] http://www.forrester.com/rb/Research/global _enterprise_web_20_market_forecast_2007/q/id/43850/t/2
[2] Private communication from www.boardtracker.com

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.
.

tions enables the incorporation of out-degree distribution as a feature into an automated approach. In [18], the authors extended the ego-centric network analysis to analysing the roles of Wikipedia. The authors illustrate some interesting roles, like technical editors and substantive experts, but again, the manual approach is not scalable. Ellison et al [6] analysed Facebook communities using a combination of demographic features and survey based questions. Utilising regression techniques they determined the dependency of these features and relationship strength. Similarly, Barash et al. [4], used regression to learn the linear relationships features, like number of threads and posts contributed by a user and the tags of their posts, to classify whether messages are factual (e.g., technical) or relational (e.g., opinion and support). In both papers the value of the target function/feature is known and the goal is to learn the relation between their target and the features. However, we do not know our target measure/grouping, and we are interested in determining discrete groupings of features to discover common roles. In [11], Himelboim et al. analysed the social roles in political forums. Three different ratios based on the amount of replies to posts and threads initiated by a user were proposed. Although they were able to distinguish between social leaders and the rest of the users in the forum, in our analysis, we found the ratios are most likely a result of the interaction graph been scale-free; i.e., you have a very small set of users who communicate/link to most of the other users.

In role equivalence [14, 17], users are modelled as vertices in a graph, and edges, usually undirected and unweighted, represent some relationship between the two users. There are various definitions of role equivalence, but in the strictest sense (structural equivalence), a set of users play the same role if they are linked to the same set of users. A looser definition of role equivalence and one more relevant to our work is regular equivalence. Here, two users play the same role if they are connected to the same types of users. Regular equivalence can be used to discover relational social roles, but it is difficult to incorporate non-binary features like number of replies between two users into the equivalence model and techniques. Therefore, role equivalence cannot be used to discover roles currently.

Finally, a number of previous works have focused on analysing question and answer style discussion forums. In [7], the aim was to determine the informative part of answer posts. and in [12], a combination of natural language techniques and reputation measures was used to classify question-answer type threads. Similar to our work [1] analysed the communication graph of the forums of the Q&A website Yahoo!Answers and examined the thread length, amount of replies to user questions, in and out degree of users, etc. to classify a forum. Unlike our work, they do not break down a forum into the composition of user roles.

In summary, all the presented related work either used manual methods to analyse user roles and forums, which are not scalable, limited to unweighted relations (role equivalence), or are focused on Q&A forums only. In this paper, we present our forum analysis approach that is automated and scalable to larger forums, can analyse weighted features, and can be applied to any type of discussion forums.

3. DATA SETS

In this section, the Boards.ie data set is discussed. Boards.ie is the largest general topic discussion board in Ireland. In the last 12 months, there were 596 forums, 244850 threads, 75400 users and over 4.3 million posts. The Boards.ie dataset is different from other publicly available discussion board datasets in that it contains the internal reply structure. As described in the introduction, a key challenge in analysing discussion forum interaction patterns is knowing which post is replying to which other post. Without access to the boards database, it is difficult and highly error prone to infer the reply structure. This is important in extracting the interaction between users, which we show is vital in profiling users.

To represent the communication interaction between users, we model the interaction as a weighted, directed graph. Each vertex represent a user in a forum, and a directed edge exists from user v_i to user v_j if user v_i has replied to a post of user v_j in thread t_k in the forum. We also associate the number of posts between two users as the edge weight. Note that from this definition, multi-edges can exist between two users, with each directed edge representing reply-behaviour from one user to another in a particular thread. We call this graph the *reply graph*. The *collapsed reply graph* aggregates all the multi-edges into a single edge, with the weight of the resultant edge being the sum of the weights of the multi-edges. The reply graph is used to analyse reciprocity of communications between users and which types of users are communicating.

To demonstrate the profiling technique, in this paper we focus our analysis to 20 different forums from boards.ie from the period 01/07/2006 to 31/12/2006, inclusive. The forums represent a range of topics from discussion to technical to advertisement. The method is general and can be applied to any number of forums. See Table 1 for a list of the 20 forums.

4. FORUM COMPOSITION APPROACH

In this section, we describe the approach we used to decompose forums into a set of user roles. We first describe the set of features used to build the user roles. Then we present our method to group the similar users, with each group representing common user roles.

4.1 Features

In this section, we describe and explain a selection of features we use. Some of the features are discriminating within a forum, while others are useful for analysing users across forums. We have analysed approximately 50 different features, but many of these were highly correlated with each other, hence redundant for grouping purposes. In the following, we will present and explain the features features we used in this paper.

4.1.1 Structural Features

Structure features provide an indication of the communication between users. They do not take into account the actual amount of communication, but examine who is replying to who, how many users reply to a user, etc. These can be derived from the properties of the unweighted, directed graph. We wish to understand the type of users each user interacts with. Users can be characterised by the interactions of their neighbours. For example, an elitist is a user who tends to only talk to their own, small clique. Therefore, he can be characterised by his neighbours who are elitist as well – their neighbours would have low degree in general, but high degree among themselves. To characterise the neighbours, we study the ego-centric networks of each user [17]. An ego centric network is the sub-graph consisting of a user, their neighbours they interact with and the edges between the neighbours themselves. It provides a view of the local interaction pattern around a user and the type of users (e.g., users with many neighbours) a user interacts with.

In our analysis of the Boards.ie data, we found the ego-centric networks follow a power-law distribution. This is further confirmed by the low clustering coefficient of all ego-centric networks. Therefore, to represent the ego-centric networks and the type of neighbours, we analysed the in- and out-degree distributions of the neighbours. The distributions form a power law, which is a heavy-tail distribution and can be parametrised by its exponent. Therefore,

Title	Type	Users Num	Threads Num	Posts Num	Edge Num
Poker	Hobby	760	2059	34726	14389
Soccer	Hobby	767	902	31523	15587
Martial Arts	Hobby	548	751	12614	5252
Personal Issues	Support	2234	1311	25715	14991
Politics	Discussion	1012	405	15002	8607
Christianity	Discussion	236	73	5867	1548
Paranormal	Discussion	1236	72	4873	1112
Humanities	Discussion	567	102	3813	2365
UCD	General	467	425	11552	5148
TCD	General	341	243	8584	3066
Real World Tournament and Events	Advertisement	426	425	8567	4295
Accommodation & Property	Advertisement/Discussion	859	451	4582	3060
Gigs and Events	Advertisement	1028	350	5122	3546
Playing instrument	Technical	422	567	7610	3045
Overclocking Logs	Technical	291	295	4379	1806
Windows	Technical	694	492	3488	2319
Development	Technical	528	465	3014	1881
Travel	Advice	1370	903	6342	4380
Thunderdome	Flame	438	37	5378	2034
Weather	Misc	215	43	3273	1045

Table 1: Forums selected for detailed analysis.

we represent the neighbourhood distributions by the exponent of the power law (**in-degree exponent**, **out-degree exponent**).

4.1.2 *Reciprocity Features*

The feature **% of bi-directional neighbours** represents the percentage of the neighbours of a user where there is both in and out edges (i.e. they have replied to each other). In addition, we analysed the percentage of *threads* in which a user has reciprocal communication with at least one other user (the two users have replied to each other's post in the thread). A user can have a low percentage of bi-directional neighbours but a high percentage of threads in which there is at least one reciprocal communication.

4.1.3 *Persistence Features*

Persistence features measure the length of the conversations a user typically engages in in. We measure the mean and standard deviation of the posts per thread (**average post/thread**, **std. dev. post/thread**). We included the standard deviation because it suggests the spread of posts per thread that the mean hides.

4.1.4 *Popularity Features*

These features measure how popular a user is. The more popular a user, the more likely are they to be replied to. We use two measures for this category. The first one is the ratio of a user's in-neighbours (i.e., those that replied to the user) compared to all users that have replied to someone else (**in-degree %**). This measures popularity based on the number of repliers. The second measure measures the percentage of posts where there is at least one reply to the user. This measures popularity based on the number of replied posts. These two measures are complimentary, as a user can have many repliers but only a low percentage of her posts actually receive replies.

4.1.5 *Initialisation Features*

initiated % measures what percentage of threads are initiated by a user. It can distinguish users who initiate many threads from those that just replies. We also computed the percentage of threads initiated that have at least one reply as a measure of the popularity of threads initiated by a user. However, we found that this feature was not very discriminating, as most users have very high scores. Therefore, we excluded this measure.

We have presented nine different features used for grouping users into common roles. Next we describe how we perform the grouping to find the common roles.

4.2 User Role Discovery Approach

Analysing the reply graphs of each forum, we found that the in-degree, out-degree, post/thread distributions and the distributions of many other features have a heavy tail distribution - a power-law in many cases, indicating that the reply graphs have scale-free properties [3]. Thus, most highly connected and highly communicative users are connected to many low degree and one-post users. As a result, most users have a star shape ego-centric network. As these low-degree, one-post users have similar characteristics across all forums, we removed them from the analysis. Such users add noise to the profile of other users. Due to their uniform nature, they do not add much insight, apart from what fraction of the forum they constitute. The star shape hypothesis is reinforced by a very low clustering coefficient for the high degree users.

We filter out the low-degree, low posting activity users. We do not set a hard threshold because what constitutes a low-degree user varies across forums. Instead, we use clustering techniques to partition the users of each forum into three bands. Each band will constitute a group of users with similar attributes within a forum. Most of the features are correlated strongly with size, or do not vary much within the forum.

Using principle component analysis to analyse the features, we found that the the amplitude of the largest principal component constituted more than 95% of the variance in the features, and the size of the ego-centric networks was the dominant feature in the largest component. Hence, we use the size of the ego-centric networks as our feature to partition the users into the three bands.

We discard the lowest band, which consists one-post users, and the middle band, which does not have enough neighbours to have an accurate power law exponent fit. Using agglomerative hierarchical clustering, we cluster the feature profile data of the remaining

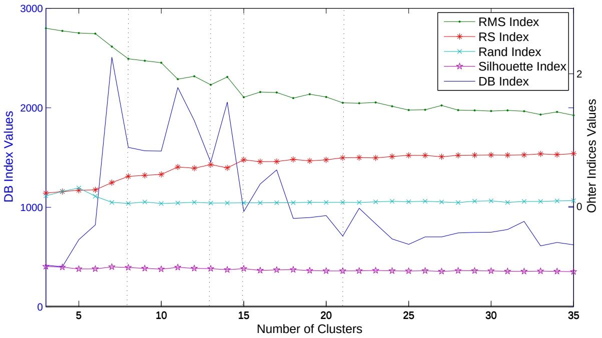


Figure 1: Plot of cluster validation indices versus the number of clusters. The dotted lines are the possible optimal number of clusters.

top band users from all forums. We use agglomerative hierarchical clustering because it doesn't make assumptions about the cluster shape like *k*-means and is a good method for data exploration. To determine the optimal number of clusters, we use five different validation techniques. Rand, Silhouette, RS, Root mean square and DB Index [9]. Internal validation techniques involve measuring the discovered clusters against a predetermined criterion, and are used when we do not know the actual clustering and hence, cannot compare against them.

As these indices are increasing or decreasing with the number of clusters, we look for kinks or knees in the plots of these measures against the number of clusters to estimate the optimal number of clusters. Figure 1 shows the plot of these measures against the number of clusters. We found that the optimal number of clusters was either 8, 13, 15 or 21. After manual inspection, we selected 8 and 15 as the best numbers of clusters.

5. USER ROLES IN BOARDS.IE

Table 2 shows the averages of the measures for cluster for the *k*=15 partitioning. Each row represents one cluster, with its unique Id, the average value of the nine features and the number of users in each cluster (size). Each cluster approximately corresponds to one user role type. The average value of the nine features and the number of users in each cluster are used to build a quantitative description of the clusters/user role types. We manually grouped the 15 types into 8 role categories. While this process was informed by the *k*=8 partitioning, we also noted that this partitioning would not have discovered all of the roles summarised in Table 3. In the next section, we describe how we classify forums based on their role composition.

Clusters 1 and 2 can be considered as forming the role *joining conversationalists*. These users do not initialise any threads, have extremely high average and standard deviation of posts per thread, a high percentage of posts that are replies, but only communicate with a relatively number small number of users. Additionally, their in and out degree exponent is relatively low, as a fair amount of communication occurs among the neighbours themselves. For example, an user called "CreepingDeath" (Cluster 1) exclusively posts on a sticky thread called the "The Insults Thread" in the Thunderdome forum. This user does not initiate any threads during the period we analysed, and communicates extensively with a small but dedicated group of users. Clusters 3 and 13 formed the role *popular initiators*. These users have very high levels of thread initialisation and relatively high popularity (high in-degree %). One example is the user "6th" on the paranormal forum, who regularly starts new threads about ghost sightings and has lengthy discussions with other users.

Clusters 5 and 6 form the role of *taciturns*. *Taciturns* have extremely low reciprocity, suggesting they rarely get involved in two way conversations. In addition, they do not communicate with many users (low in-degree %), and when they do, they post only a few posts per thread. For example, "gerryk" answers a question in the windows forum in 2006 (posting to the forum once every 6-18 months), hence having no two way conversation and one post per thread.

Cluster 9 are the *elitists*. *Elitists* are characterised by low percentage of neighbours with 2-way communications, yet have many threads with 2-way communications. Combined with a low in-degree %, this suggest that *elitist* prefer to carry on a conversation with a small set of users. An example elitist is the user "Akrasia" in the humanities forum. This user gets involved in long conversations with a few users in a thread about The US and Iraq.

Clusters 4 and 7 constitute the *supporters*. These users have middle of the range statistics for all features. These users are most likely the transition stage from grunts to the highly communicative roles like the popular initiators. For example, user "nicnicnic" is a frequent poster in the poker forum, often communicating with other frequent posters.

Clusters 8, 12 and 14 form the role of *popular participant*. These users do not initiate many threads, but are involved with a large percentage of the users of a forum. They can be considered the intermediate role between *joining conversationalist* and *popular initiators*. An example of such type of user is "Son Goku", who frequently chats to other users in the Christianity forum, but does not initiate many threads.

Clusters 10 and 11 constitute the role of *grunts*. *Grunts* have similar profiling as *taciturn*, but are distinguished by their relatively higher levels of reciprocity. It can be argued to merge these two roles, but as we shall show in Section 6, one of the forums is mainly consisting of *taciturns* but few *grunts*, hence both roles are required. An example is the user "Furious-Red", who posted a question on the gigs & events forum to which he got a reply.

Cluster 15 form the role *ignored*. These users are characterised by having an extremely low % of their posts been replied to, suggesting they are not very popular in the forum. Note that they are not spammers because all the forums in Boards.ie are moderated, hence spamming would have been removed – an example is the user "Anti" in the thunderdome forum.

6. FORUM COMPOSITION IN BOARDS.IE

In this section, we analyse the 20 selected forums using the roles discovered in the previous section. We will show that we obtain different and unexpected groupings of the forums using the role compositions.

Figure 2 shows the role composition of the 20 forums. The same colour/shading is used in each pie chart to illustrate the same role across the forums. Visually, we can see some forums are distinctly different from the others, such as the personal issues forum. But there are also some forums that have similar compositions, such as the soccer and poker forums. Using an unweighted Euclidean distance, we cluster the forums into groups (see Table 4).

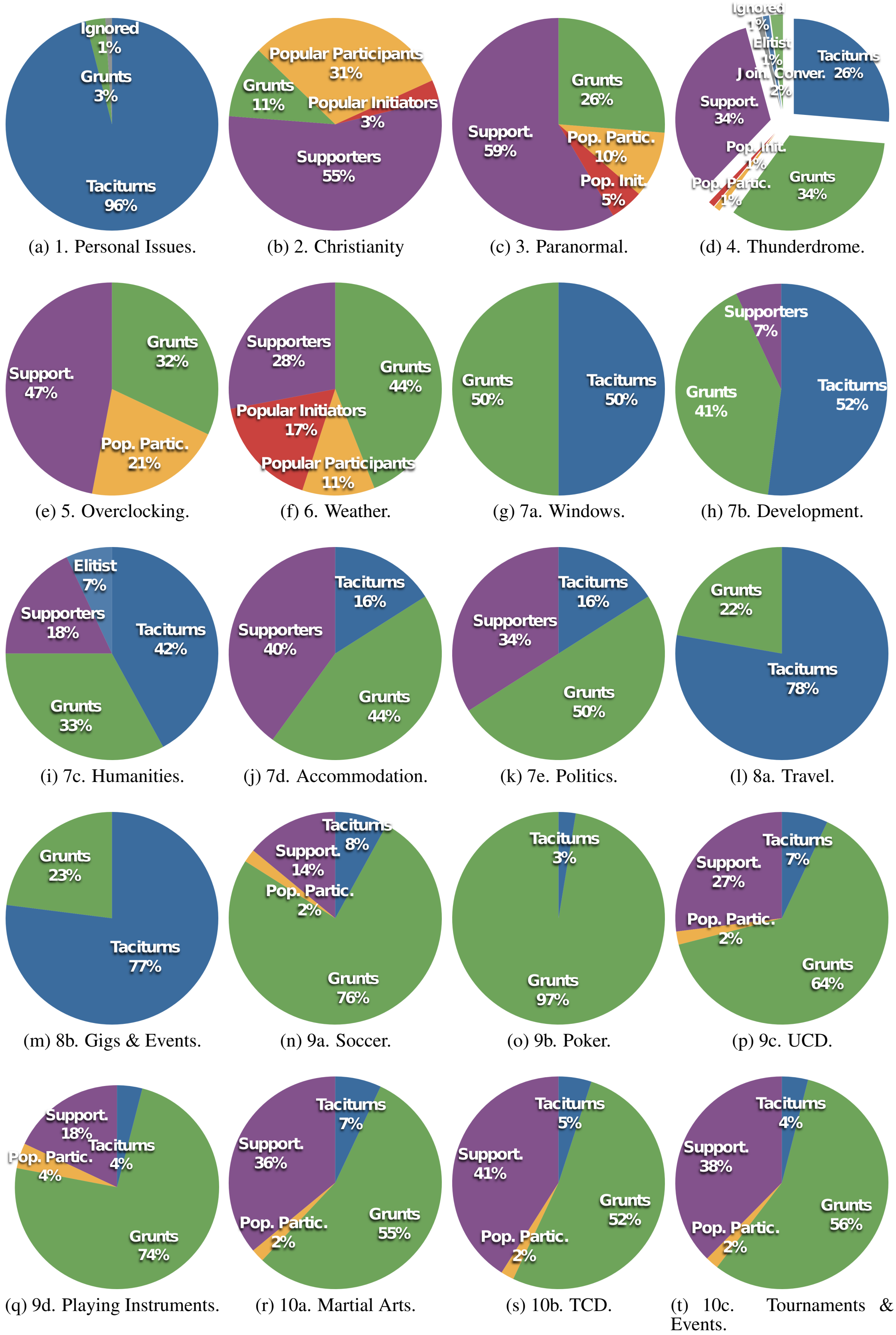
We can clearly see the single groupings of 1 to 6 have very different composition to all other forums. For example, the *taciturn* role makes up 95% of all users in the personal issues forum (grouping 1). This suggests that most users in the personal issues forum go to rant and complain, not really expecting replies. The thunderdome forum (grouping 4) is composed of 12 of the 15 clus-

Cluster	**In-deg. %**	**In-deg. Exp.**	**Out-deg. Exp.**	**% Init.**	**% of Posts Replied**	**% of Bi-dir Neighs**	**% of bi-dir thrs.**	**Avg. post/thr**	**Std. post/thr**	**Size**
1	0.1080	-1.5189	-1.5101	**0.0000**	**0.8417**	0.7857	1.0000	**14.7027**	**26.0061**	1
2	0.0526	-1.4895	-1.4929	**0.0000**	**0.8333**	0.6667	1.0000	**9.2381**	**18.0341**	1
3	**0.4171**	-1.5893	-1.5854	**0.1991**	0.7162	0.5573	0.8648	3.1309	7.4597	3
13	**0.1142**	-1.5240	-1.5189	**0.1029**	0.7835	0.4023	0.8053	1.8421	1.5059	3
5	**0.0345**	-1.6100	-1.6279	0.0020	0.6283	**0.1963**	**0.2126**	**1.0944**	**0.3156**	237
6	**0.0233**	-1.9321	-1.9422	0.0036	0.7731	**0.1904**	**0.2869**	**1.1098**	**0.3098**	98
4	0.1389	-1.4153	-1.4203	0.0111	0.7168	0.4958	0.4842	1.3959	0.9791	221
7	0.0948	-1.5431	-1.5448	0.0049	0.7685	0.5147	0.6926	2.7566	4.1268	31
9	**0.0344**	-2.4132	-2.0564	0.0122	0.7361	**0.1876**	**0.8667**	1.5124	0.6982	5
8	**0.2121**	-1.4329	-1.4293	**0.0079**	0.6871	0.6293	0.8745	**4.8848**	**10.3483**	5
12	**0.2593**	-1.4522	-1.4552	**0.0557**	0.7394	0.5494	0.5971	**1.6025**	**1.4121**	20
14	**0.1743**	-1.5067	-1.4967	**0.0253**	0.7401	0.6298	0.6955	5.9383	17.4885	2
10	**0.0344**	-1.7530	-1.7741	0.0026	0.7337	0.3857	0.7590	**1.4299**	**0.8749**	32
11	**0.0781**	-1.4261	-1.4321	0.0048	0.7442	0.3597	0.3693	**1.2021**	**0.5515**	517
15	0.0321	-1.5487	-1.5636	0.0000	**0.1302**	0.1688	0.3788	1.5583	1.2591	2

Table 2: Cluster statistics. Columns 2 to 10 are the means of the nine features for each the 15 clusters. The size column is the number of users in each cluster.

Name	**Clusters**	**Comments**
Joining Conversationalists	1, 2	No initialisation. High levels of communications with a relatively small set of users.
Popular Initiators	3, 13	Very high levels of thread initialisation, coupled with relatively high popularity (high in-deg %).
Taciturns	5, 6	Very low reciprocity, volume of communication and few neighbours suggest limited conversation with a few users. The main difference between clusters 5 and 6 is their exponents, suggesting they communicate with different types of neighbours.
Supporters	4, 7	Relatively middle of the road statistics, suggesting the users form the backbone of the forums. Difference between clusters 4 and 7 is the amount of communications.
Elitists	9	Characterised by very low percentage of neighbours with bi-directional communications but high percentage for bi-directional threads. Combined with low in-deg percentage, these users prefer to carry on conversation with a very small set of users.
Popular Participants	8, 12, 14	Do not initiate much threads, unlike the popular initiators, but are involved with a large percentage of users on forums. They can be considered a cross between joining conversationalist and popular initiators. The difference between clusters 8 and 12 is the volume of communications.
Grunts	10, 11	Low volumes of communications to a few users. Different from taciturns by the relatively higher levels of reciprocity.
Ignored	15	Very low percentage of posts get replied to

Table 3: Summary of the common user roles.



(a) 1. Personal Issues.

(b) 2. Christianity

(c) 3. Paranormal.

(d) 4. Thunderdrome.

(e) 5. Overclocking.

(f) 6. Weather.

(g) 7a. Windows.

(h) 7b. Development.

(i) 7c. Humanities.

(j) 7d. Accommodation.

(k) 7e. Politics.

(l) 8a. Travel.

(m) 8b. Gigs & Events.

(n) 9a. Soccer.

(o) 9b. Poker.

(p) 9c. UCD.

(q) 9d. Playing Instruments.

(r) 10a. Martial Arts.

(s) 10b. TCD.

(t) 10c. Tournaments & Events.

Figure 2: The user role composition of the 20 forums.

Id	Forums
1	Personal Issues
2	Christianity
3	Paranormal
4	Thunderdome
5	Overclocking
6	Weather
7	Windows, Development, Humanities, Accommodation, Politics
8	Travel, Gigs & Events
9	Soccer, Poker, UCD, Playing Instruments
10	Martial Arts, TCD, Tournaments & Events

Table 4: Forum groupings.

ters, which reflects its nature as the forum to which flame wars are moved from all different forums. Grouping 2 (the Christianity forum) has a strong component of cluster 13 (popular initiators), suggesting that a few users regularly initiate threads that subsequently generate discussion (large percentage of popular participants and supporters). Grouping 6, the weather forum, is also strongly constituted by popular initiators and popular participants, but it has a larger portion of grunts and supporters than the Christianity forum, suggesting that lengthy discussion is not as widespread as in the Christianity forum.

Grouping 8 consists of the travel and gigs forums. A large percentage of their users assume the *taciturn* role, illustrating that these forums tend to be where events and travel advertisements are posted and long conversations are rare.

Groupings 7, 9 and 10 consists of four forums each. However, the crucial difference between the three groupings is the relative proportions of grunts, popular participants, supporters and taciturns. Forums in grouping 9 have the largest portion of grunts (role cluster 11) compared to the other two groupings. In addition, it has no users from the role cluster 10, suggesting its in and out-degree exponent is low, around -1.4. Furthermore, forums in grouping 9 are typically lead by a small number of popular participants supported by a sizeable band of supporters. In contrast, forums in grouping 7 have fewer portion of grunts but a larger portion of supporters (role cluster 4) and a significant portion of taciturns. This suggest these forums have a larger number of regular participants than forums from grouping 9, and also a sizeable number of users who post questions (Windows and Development forums) or advertisements/ideas (Accommodation, Humanities and Politics) that are not answered or ignored. The taciturn role is mostly absent from forums in grouping 9, which indicates that there are less question type or controversial type of threads in these forums, leading to increased chance of getting a reply. Finally, forums from grouping 10 consists of roughly 50% of grunts with a significant portion of supporters and almost no taciturns. Forums in grouping 10 differ from those in grouping 7 in that they do not have any taciturns, suggesting that most threads and posts will get a reply. The forums also are less likely to touch on controversial topics, increasing the chance of a reply. Finally, we note that the two university forums, TCD (Trinity College of Dublin) and UCD (University College of Dublin) are in different groupings. UCD has a larger portion of grunts and a smaller portion of supporters than TCD, suggesting UCD has more one-off users (UCD has more users, posts and threads than TCD in the six month period analysed).

7. FUTURE WORK

In comparison to the Weather or Christianity forums, the forums in grouping 7 appear to be much less social, dominated by taciturn and grunt roles. At face value this may suggest that these forums are not functioning well. However, this may not be the case as two of these forums are technical support forums (Windows and Development) where a question-and-answer format may be the most usual form of communication. Similarly, the accommodation forum tends to be made up of personal notices seeking or advertising accommodation rather than discussion. The humanities forum has a clique of highly social elitists and a sizeable number of supporters. As with the politics forum, it is difficult to say without further analysis whether these two forums are functioning to the satisfaction of their participants. Future work will involve examining what are the behavioural norms for different types of forums (e.g. Q&A, hobbyist, social, announcement forums etc). As the objective of this work is to be able to predict the health and sustainability of forums, we plan to examine the dynamics of role evolution and forum composition over time. For example, we will examine failed forums and forums that have continued to function successfully over time in order to develop predictive models of forum health. Finally, we plan to investigate the types of policies that forum owners might introduce to rescue a forum that is in danger of failing.

8. CONCLUSION

In this paper, we have presented a novel method to analysis forums, namely categorising the roles played by users in the forums, then analysing and comparing the forums using their composition of roles. We used nine different features to profile the user roles, including popularity, reciprocity, length of interaction, initialisation, neighbour's roles and volume of communication measures. Then applying a two stage clustering approach, we group the users of the forums into 15 groups and eight roles. Using these roles, we describe the forums based on their composition of these discovered roles. We showed how the forums can be clearly compared, analysed and grouped based on their composition, and how it is not possible to do this by analysing the forums by averaging out the features across their users. In further work, we plan to analyse the role composition across time. At the moment, we take one snapshot of a discussion forum or board and analyse the composition, but we can learn how forums change and how users move between roles by including time in our analysis. In addition, we would like to extend our role composition technique to other domains with public publishing, e.g., weblogs.

9. ACKNOWLEDGEMENTS

This work was carried out in the CLIQUE Strategic Research Cluster which is funded by Science Foundation Ireland (SFI) under grant number 08/SRC/I1407. We would like to thank John Breslin for kindly providing the Boards.ie data and Ron Kass of Boardtracker.com for insightful discussions and board statistics.

10. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of WSDM '08:*, pages 207–218, New York, NY, USA, 2008. ACM.

[3] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[4] V. D. Barash, M. Smith, L. Getoor, and H. T. Welser. Distinguishing knowledge vs social capital in social media with roles and context. In *Proceedings of the ICWSM 09*, May 2009.

[5] A. B. Brush, X. Wang, T. C. Turner, and M. A. Smith. Assessing differential usage of usenet social accounting meta-data. In *CHI*, pages 889–898, 2005.

[6] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4), 2007.

[7] D. Feng, E. Shaw, J. Kim, and E. Hovy. Learning to detect conversation focus of threaded discussions. In *Proceedings of HLT/NAACL 2006*, pages 208–215, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[8] D. Fisher, M. Smith, and H. T. Welsher. You are who you talk to: detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference*, volume 3, pages 59–68, January 2006.

[9] J. Handl, J. Knowles, and D. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201, 2005.

[10] M. Hauben. *Netizens: On the History and Impact of UseNet and the Internet*. Peer-to-peer Communi, 1996.

[11] I. Himelboim, E. Gleave, and M. Smith. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication*, 14(4):771–789, 2009.

[12] L. Hong and B. D. Davison. A classification-based approach to question answering in discussion boards. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval*, pages 171–178, New York, NY, USA, 2009. ACM.

[13] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[14] J. Lerner. Role assignments. In *Network Analysis*, pages 216–252, 2004.

[15] T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4), 2005.

[16] F. B. Viégas and M. Smith. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, page 40109.2, Washington, DC, USA, 2004. IEEE Computer Society.

[17] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1 edition, 1994.

[18] H. Welser, G. Kossinets, M. Smith, and D. Cosley. Finding social roles in wikipedia. In *Presented at the Annual Meeting of the American Sociological Association Annual Meeting*, pages 1–11, July 2008.