Here, have an upvote: communication behaviour and karma on Reddit

Donn Morrison and Conor Hayes

Digital Enterprise Research Institute
National University Ireland, Galway
`first.last@deri.org`

Abstract. In this paper we cluster users of discussion forums into behavioural roles based on features derived from their egocentric reply-graphs. We then analyse the user's popularity with respect to the prototypical behaviour the user demonstrates. We perform this analysis on a selection of discussion forums ("subreddits") from the popular social news aggregation website Reddit.com and show that while many of the forums exhibit similar compositions in user behaviour, different behavioural roles yield different distributions of popularity ("karma") among the community. We uncover a small core of highly active users who play an instrumental role in initiating and sustaining communication throughout each community.

1   Introduction

Users participate in online and enterprise communities for a variety of reasons ranging from social to informational to support. Some communities cater to specific user needs such as the question and answering forums of Stack Exchange,[1] while other more general communities such as Reddit.com[2] offer a wide range of forums (called "subreddits") covering political, personal and leisure discussion. These communities are rich sources of data for the study of users' social behaviour.

Automatic analysis of behavioural patterns in online communities has many applications for administrators, moderators and community owners. The ability to quickly visualise and summarise discussion-based communities permits efficient management of the community structure. Behavioural changes of individuals or groups of influential users may help to indicate levels of community health [AR11], predict user churn [KRC+10] and motivate changes to overall structure such as the creation, division or combination of new or existing forums.

There have been several efforts to group users into roles according to their observed behaviour [FSW06,WGFS07,GWLS09,CHD10]. Studies in this domain typically use unsupervised clustering algorithms and rely on measures such as cluster quality due to the lack of a reliable ground truth. [CHD10] defined a set of eight prototypical behavioural roles and introduced a method of hierarchically clustering users accordingly, demonstrating that different forums had different compositions of user behaviour.

---

[1] `http://www.stackexchange.com/`
[2] `http://www.reddit.com/`

[AR11] showed how the presence of certain roles defined by [CHD10] was correlated with changes in levels of forum activity. In conducting an analysis over time, it was found that certain role compositions remained relatively stable.

Although these and other studies have provided a basis for understanding behaviour in online forums and managing the practical issues associated with such analyses, there remain many aspects of the dynamics of user behaviour that are not yet fully understood. Namely, how do behavioural roles in online forums correlate with community-defined measures of popularity? Is certain behaviour linked to more or less popular users, or do different behaviour types encompass both popular and unpopular users? We construct a behaviour analysis framework and in order to answer these questions by first clustering users into prototypical roles and then studying the distribution of user popularity within each role. In the case of Reddit.com, user popularity is measured in *karma* - users can vote on the submissions or comments of other users.

The remainder of this paper is organised as follows. Section 2 presents related research in clustering and tracking user behaviour in online forums with a focus on popularity as a feature. In Section 3 the methodology is outlined: we describe the dataset and features and then introduce the behavioural roles and rule-based cluster labelling. In Section 4 we present and discuss the results. Finally, conclusions and future research directions are presented in Section 5.

2   Related work

The analysis of user behaviour in online community settings has been partially addressed in several previous works. Some studies have looked at defining behavioural roles according to qualitative and visual analysis of the communication structure and message content [Vi04,FSW06,WGFS07,GWLS09] while others have attempted to automate the analysis in a firmly bottom-up approach using machine learning and data mining techniques [ARSX03,CHD10,AR11,KCH$^+$11]. In such analyses, behavioural roles are typically inferred from the volume of communication (number of messages posted) and various features derived from the egocentric network of the poster (e.g. the frequency a message posted by a user is replied-to by other users) [CHD10].

Chan et al. [CHD10] studied a set of 20 forums from the Boards.ie dataset and demonstrated an analysis technique that derived 8 behavioural roles they believed to underlie the user base. While we base our work on this initial work, the authors did not consider the explicit popularity of users in the context of their behaviour, only the implicit popularity derived from the reply-graph.

Research on popularity in social networks is usually limited to measures of centrality, authority, or similar. The in-degree of a user on a reply-graph, i.e. the amount of incoming communication, can provide an indication of popularity [CHD10], but this is only inferred. Popularity measured from the community by means of direct feedback (e.g. up- and downvotes on Reddit.com), is a much more reliable gauge of popularity because users are motivated to explicitly provide these ratings.

The authors of [Mie11] analysed the relationship between up- and downvotes on Reddit.com and found that the overall score (the upvotes minus the downvotes) could be modelled by a general random walk with a boundary equal to the number of users

that are able to vote on the submission. In [SH10], the authors introduced a method for predicting long-term popularity of content (user submissions) on Digg based on early measurements of user assessment (votes). The authors showed that they could accurately predict the popularity of content 30 days in advance by examining user-provided votes in the first two hours following the submission of new content. However, neither of the above works consider the link with user behaviour.

Within the context of a larger analysis, [RAAF12] studied the popularity of a user's post (points awarded by other users) and its contribution to behavioural roles in enterprise discussion forums. They found that popularity was highly correlated with features for engagement and contribution. However, although popularity was found to be important for at least one type of user behaviour (the *focussed expert participant* role), the information was used as a feature. A more thorough analysis of this feature is desirable, especially with respect to different behaviour types.

To our knowledge, no research has explicitly examined the link between user popularity and the behavioural role to which a user belongs. Studying this particular aspect of user behaviour could lead to the prediction of popular users based on a sample of communication behaviour. In effect, this experimental study can be seen as a precursor to prediction of popular users based on previous behaviour.

3 Methodology

3.1 Dataset and features

We analyse forums ("subreddits") from the popular social news aggregation website Reddit.com. The data was crawled over a period of three months from August 2011 to October 2011. The Reddit.com API allows the retrieval of the last 1,000 posts per subreddit and 500 comments per post. The subreddits studied in this paper are 20 of the most active: *politics, gaming, pics, worldnews, technology, videos, AskReddit, aww, Music, atheism, funny, movies, bestof, WTF, askscience, todayilearned, fffffffuuuuuuuu-uuuuu, IAmA, trees, science*. The names of each forum are generally self-explanatory. For example, the *politics* subreddit is concerned with the submission and discussion of politics-related news articles and other media. The submissions to *AskReddit* are in the form of a question a user wants answered and the ensuing discussion centres around that question and its answers. The submissions to the *pics* subreddit consist exclusively of links to photographs. For brevity, we omit the descriptions of each subreddit and allow the reader to visit `http://www.reddit.com/r/<subreddit-name>` for more information. In our dataset, *IAmA* was the most active, with 24,450 users. *aww* was the least active, with 3,093 users. The mean number of users per subreddit was 10,378.

The users of Reddit.com can accrue karma through both submission of posts and through comments associated with those posts. Karma is allocated by other users through the action of "upvoting" and "downvoting" these entities. To an extent, karma motivates and facilitates participation in the community, with many users eager to increase their karma by appealing and catering to the interests of others, both in opinion and content. In this study we only consider karma associated with the commenting system of Reddit.com. We do not consider the karma awarded to submissions (threads).

In addition to karma, the features used in this study are based on communication behaviour and the egocentric reply-graphs of the users in each forum. Specifically, we make use of the following 10 features:[3] *# of submitted posts (new threads)* (**th**), *mean # of comments per post that the user engages in* (**mpth**), *standard deviation of comments per post that the user engages in* (**spth**), *# of comments submitted by the user that received at least one reply* (**pr**), *proportion of a user's peers where bi-directional communication exists* (**bin**), *proportion of posts participated in by the user where bi-directional communication exists* (**thbi**), *forum focus dispersion (entropy - lower if a user participates in few forums, higher if the user participates in many forums)* (**ent**), *in-degree (number of incoming edges from a user's peers)* (**ind**) *and out-degree (number of outgoing edges to a user's peers)* (**outdeg**). The features were chosen to capture behaviour such as initiation (th), engagement (mpth, spth), reciprocity (pr, bin, thbi), focus or breadth of interest (ent), and popularity (ind, outdeg) [CHD10,RA11].

3.2 User clustering with K-means

We cluster users based on the features listed above using K-means. K-means is an unsupervised clustering algorithm that converges on a local minimum by a two-step iterative process. The only parameter it requires is the number of initial centroids k. K-means finds a partition $\{\mathcal{V}_1, ..., \mathcal{V}_k\}$ that minimises:

$$KM = \sum_{l=1}^{k} \sum_{i \in \mathcal{V}_l} ||\boldsymbol{v}_i - \boldsymbol{\mu}_l||^2, \tag{1}$$

where $\boldsymbol{v}_i \in \mathbb{R}^m$ is the feature vector and $\boldsymbol{\mu}_l$ denotes the centroid for cluster l. Stability between independent clusterings is often problematic with K-means. In order to address this we randomly sample the set of centroids and choose the set that maximises the inter-cluster distances.

3.3 Behavioural roles

The roles we defined for this study were adapted from those introduced by [CHD10] within the context of general purpose online discussion forums. This adaption attempts to encompass as many differing types of user behaviour as possible while maintaining broad enough definitions such that we can have a clear understanding of how they contribute to the community.

The choice in the number of behavioural roles came from a model selection using K-means and evaluated using the silhouette coefficient. We performed successive clusterings where we varied the number of clusters from two to 30 for each forum. The results overwhelmingly pointed to an optimal clustering at $K = 4$ clusters.

Informed by the model selection, we manually inspected the cluster centroids for all subreddits and labelled each cluster with the high-level characteristics (initiation, engagement, reciprocity, focus, popularity). We then derived a decision rule to map each cluster to a one of the semantic labels defined in Table 3.3. This decision rule permits automatic cluster labelling and is presented as Algorithm 1.

---

[3] Features are referred to later in the text only by the abbreviation in parentheses.

Table 1: Role definitions, characteristics and correspondence to the roles defined by Chan.

Role label	Features	Chan's roles	Characteristics
Contributor	High initiation, medium engagement, high reciprocity, high popularity (in- and out-degree)	Supporter/popular participant/popular initiator	Stable backbone of the forum; contributes useful content that yields replies
Ignored	High engagement, low reciprocity, low popularity	Ignored/taciturn	Generally ignored by other users; in unmoderated forums spammers would fall into this role
Lurker	Low engagement, low reciprocity, low popularity	Grunt/joining conversationalist	Low communication with few users
Casual commentator	Medium engagement, high reciprocity	Elitist	High communication with few users

Prior to the mapping, the centroids are scaled into the range $[0, 1]$. First, **Contributors** are separated by finding the highest combination of *spth, pr, ent, ind, outdeg* (high initiation, medium engagement, high reciprocity and high popularity). Next, **Ignored** users are isolated by finding in the remaining clusters the combination of highest *mpth* (high engagement) and lowest *bin* & *thbi* (low reciprocity). **Lurkers** are then determined by the lowest *mpth* (low engagement). Finally, **Casual commentators** constitute the remaining cluster with medium engagement and high reciprocity.

Algorithm 1 Role labelling decision rule set.

Input: Set of K clusters $\mathcal{C}$ with cendroid means {mpth, spth, pr, bin, thbi, ent, ind, outdeg}
Output: Set of K labels $\mathcal{L}$

    contributor $\leftarrow c \in \mathcal{C} : \text{argmax}_c \text{ spth}_c + \text{pr}_c + \text{ent}_c + \text{ind}_c + \text{outdeg}_c$
    $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
    ignored $\leftarrow c \in \mathcal{C} : \text{argmax}_c (1\text{-bin}_c) + 1\text{-thbi}_c) + \text{mpth}_c$
    $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
    lurker $\leftarrow c \in \mathcal{C} : \text{argmax}_c (1\text{-mpth}_c)$
    $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
    casual commentator $\leftarrow c \in \mathcal{C}$
    $\mathcal{L} \leftarrow$ {contributor,ignored,lurker,casual commentator}

3.4 Experimental setup

The analysis carried out in this study was conducted as follows. First, the reply-graph features were extracted from the forum data. Next, for each subreddit, the features were clustered using K-means with $k = 4$ cluster centroids (derived empirically through model selection). Finally, the behavioural roles defined in Table 3.3 are mapped to the clusters based on the values of each clusters' feature means (see Algorithm 1). With the clusters mapped to roles, we can then examine the distribution of popularity between the different behavioural roles.

4 Results and discussion

The results of the behaviour decomposition per subreddit are presented in Figure 1 and 2. The results show that most of the subreddits have a high amount of **Ignored** and/or **Lurker** users (more than 50% in some cases) who but receive little or no replies and are generally ignored by other users. The main difference between these two behaviours is that when **Lurkers** choose to participate, they engage in a higher amount of reciprocal communication, whereas **Ignored** users are unable to elicit responses from other members of the community. **Casual commentators** make up the next largest component, showing that a substantial part of the user base communicates with only a handful of other users. A small core population of **Contributors** appears to initiate and sustain the majority of the communication in all of the subreddits.

Looking at the karma distributions for each role (cumulative number of upvotes minus downvotes per comment), we can see that different roles yield different distributions of karma. Users with the highest variability in karma exhibit **Contributor** behaviour. **Contributors** also have the highest mean karma, suggesting that while they may have more controversial comments and/or submissions (more extremes), they are typical of the community and generally submit content that appeals to other users. We can examine this role by looking more closely at the data of individual users. We isolate a user (whose account has since been deleted) who posted frequently in two topically related subreddits, *worldnews* and *politics*, and has been assigned to the **Contributor** role in both communities. The user has a mean karma of -1.1 with standard deviation 6.2. Looking at the comments posted by the user, we find that the numbers of up- and downvotes are almost equally distributed - a clear indication that the user posts content that other users find controversial.

Ignored and **Lurker** users have the lowest mean karma. This shows that users with this behaviour type are ignored both in the reply-graph and in the karma assigned by other users. When they do accumulate karma, it remains low. We examine these roles by randomly sampling users categorised as **Ignored** in the highly active subreddit *AskReddit*. We find that these users have low activity (only one or two posts in total) and that 10% of our sample had karma scores greater than 10 for their contributions. Performing the same analysis for the **Contributor** in *AskReddit* yields similar activity levels but with 37% of the sampled users having karma scores greater than 10.

Figure 3 overlays the behavioural roles onto the reply-graph for the *politics* subreddit. The graph layout applied is the Fruchterman Reingold algorithm. **Contributors** and some **Lurkers** are evident as comprising the core, while the remaining roles **Lurkers, Ignored** and **Casual commentators** comprise the periphery. This shows a small core of users who initiate and steer the conversation associated with each thread. This also supports the findings in Figures 1 and 2 in that these users are recognised by the community for their effort, hence the higher mean karma. Table 2 shows the cluster centroids and sizes for each behavioural role for the *politics* subreddit.

Finally, we would like to comment the large proportion of **Ignored** and **Lurker** users on Reddit.com. This is interesting for community owners, administrators and moderators because low activity users have been associated with negative forum health [RA11]. One possible explanation is in the way the comment threads are presented to the user. On Reddit.com, comments in threads are sorted by popularity by default. In the

(a) AskReddit

(b) askscience

(c) atheism

(d) aww

(e) bestof

(f) fffffffuuuuuuuuuuuu

(g) funny

(h) gaming

(i) IAmA

(j) movies
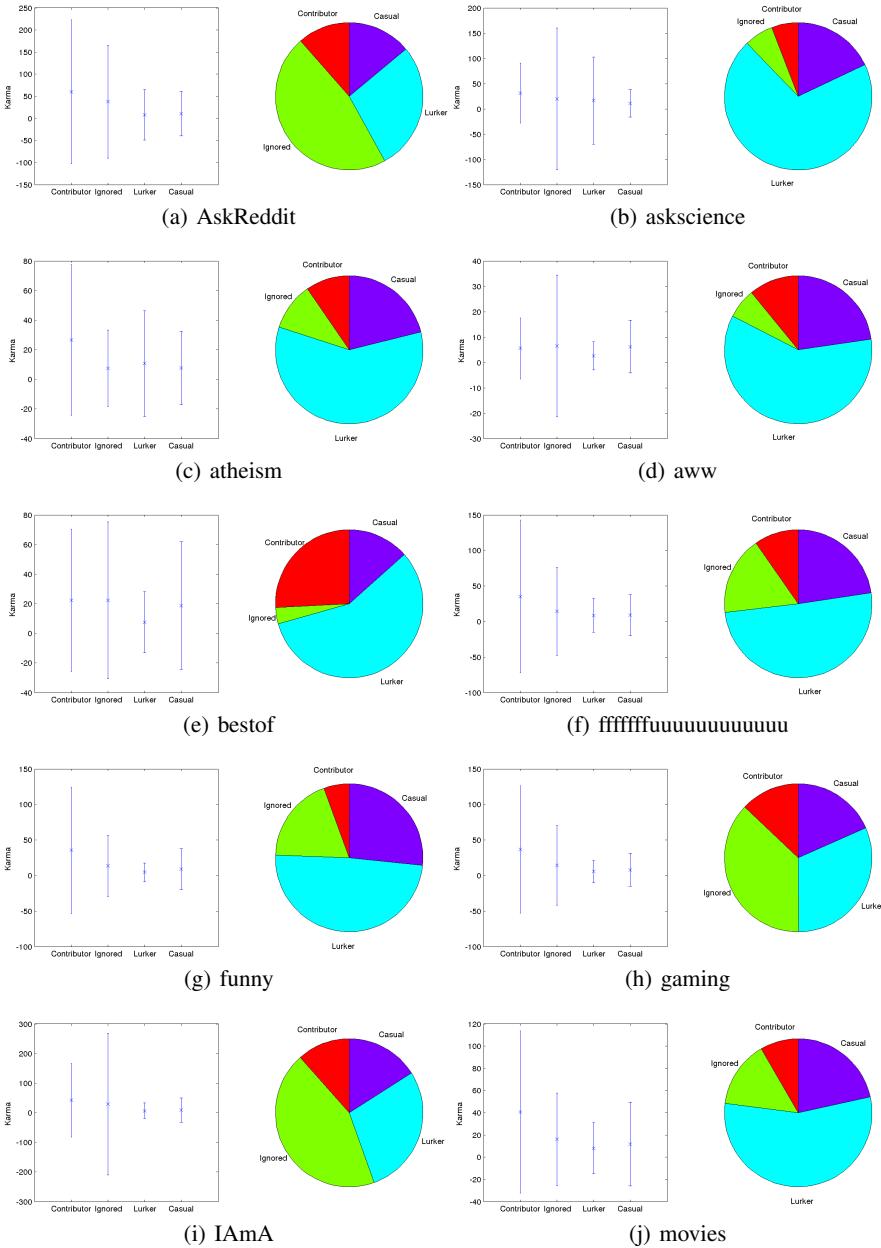
Fig. 1: Role compositions and karma boxplots for users of the selected set of subreddits from Reddit.com. The majority of the communities have high proportions of **Lurkers** and **Ignored** users. **Lurkers** have the lowest overall karma. Users with **Contributor** behaviour are evident as valued to the wider community by the higher mean karma.

2264
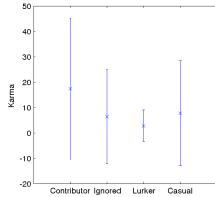
(a) Music
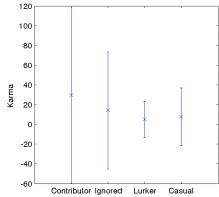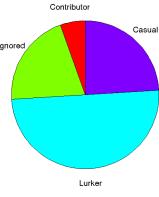
(b) pics
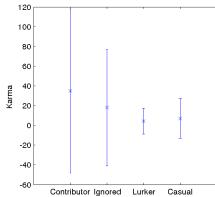
(c) politics
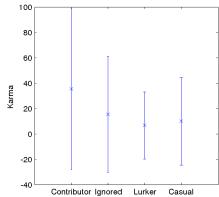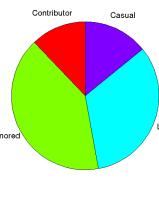
(d) science
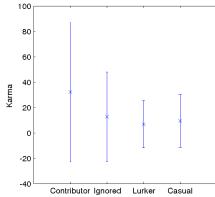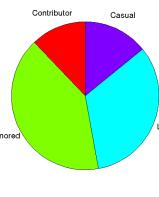
(e) technology
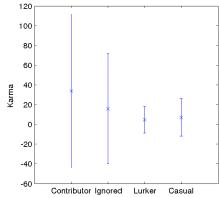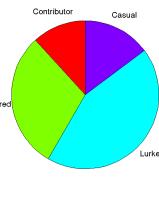
(f) todayilearned
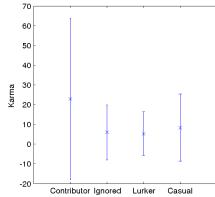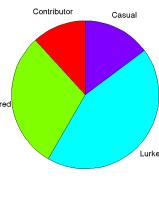
(g) trees
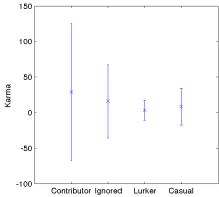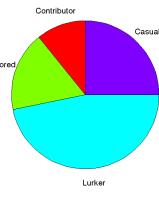
(h) videos
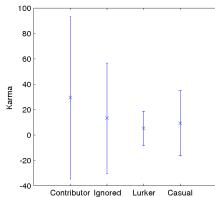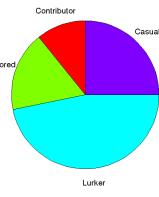
(i) worldnews
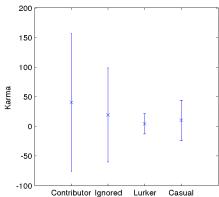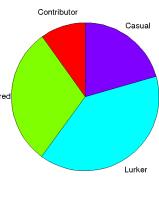
(j) WTF
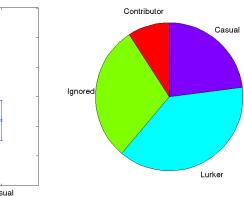
Fig. 2: Role compositions and karma boxplots for users of the selected set of subreddits from Reddit.com. The majority of the communities have high proportions of **Lurkers** and **Ignored** users. **Lurkers** have the lowest overall karma. Users with **Contributor** behaviour are evident as valued to the wider community by the higher mean karma.
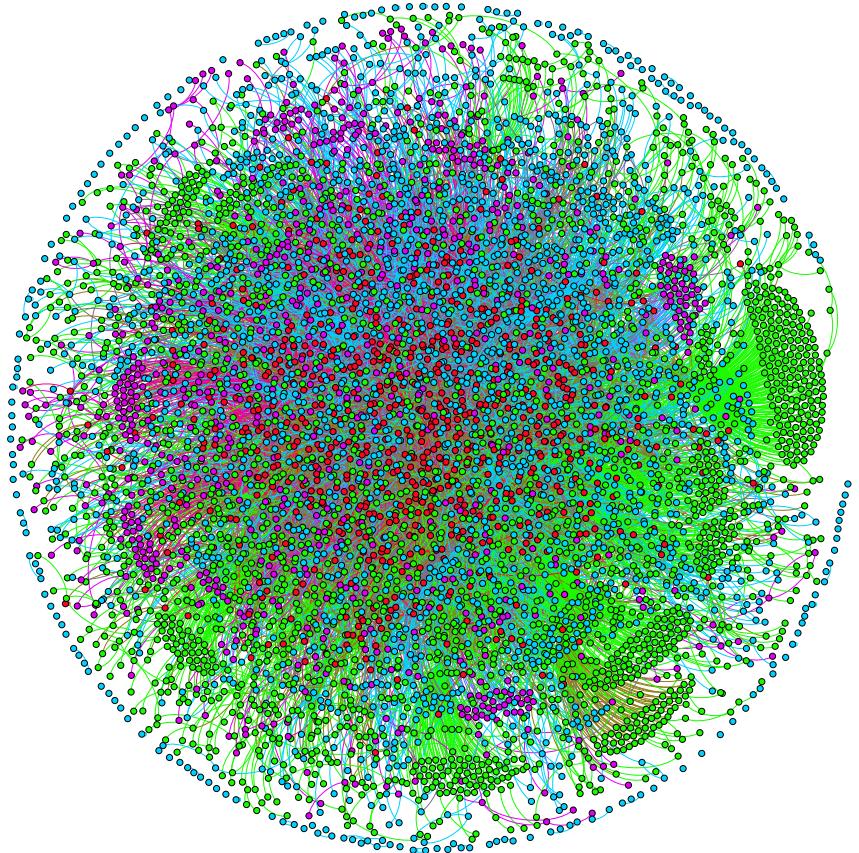
Fig. 3: The role-coloured reply-graph for the *politics* subreddit. The layout is the Fruchterman Reingold algorithm. Key: Red: **Contributor**, Green: **Ignored**, Turquoise: **Lurker**, Violet: **Casual commentator**. **Contributors** and some **Lurkers** are evident as comprising the core, while the remaining roles comprise the periphery.

Table 2: Cluster centroids for the users of the *politics* subreddit.

Role	th	mpth	spth	pr	bin	thbi	ent	ind	outdeg	Cluster size
Contributor	0.0046	0.5189	0.7362	0.0729	0.1735	0.2719	0.6016	0.0185	0.0875	827
Ignored	0.0001	0.9793	0.0157	0.0189	0.1162	0.1958	0.5436	0.0062	0.0279	2375
Lurker	0.0116	0.0769	0.0338	0.0293	0.1529	0.2186	0.5683	0.0083	0.0304	1926
Casual c.	0.0000	0.5387	0.0114	0.0202	0.1649	0.2672	0.5515	0.0044	0.0254	708

forums studied in previous work, comments are often presented to the user in chronological order. In other words, comments on Reddit.com *may* exhibit a "rich-get-richer" phenomenon where popular comments are more likely to be seen and therefore rated and replied to by other users. This leads to the question of what role user interface design plays in the perceived behaviour of users.

5   Conclusions

In this study we clustered users of Reddit.com into behavioural roles using features derived from their egocentric reply-graphs. We explored the link between the distribution of karma (i.e. popularity measured by the number of up- and downvotes) and the behavioural role to which that user belongs. Overall, users with **Contributor** behaviour (those having higher engagement, initiation, reciprocity, focus and popularity in terms of node degree) prove to be more popular, however the variance in karma in this role is high, which indicates that this role encompasses both popular and unpopular users (i.e. users who post content deemed controversial by other users).

This study is preliminary in nature but provides a basis for exploring methods to predict user popularity based on behavioural features extracted from a user's egocentric reply-graph. By predicting high-karma users, community owners or moderators, or even the users themselves, could more efficiently manage and maintain behaviour in the forums under analysis. Although our framework has the capability to analyse the evolution of user behaviour, the dataset did not permit such an analysis. We are currently crawling the subreddits of Reddit.com on a periodic basis to gain a picture of the evolution of user behaviour. Initial results on this temporal data support the findings presented in this paper.

Acknowledgements

References

[AR11]   Sofia Angeletou and Matthew Rowe. Modelling and analysis of user behaviour in online communities. 2011.

[ARSX03]   Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, pages 529–535. ACM, 2003.

[CHD10]   Jeffrey Chan, Conor Hayes, and Elizabeth M Daly. Decomposing Discussion Forums and Boards Using User Roles. In *International AAAI Conference on Weblogs and Social Media*, pages 215–218, 2010.

[FSW06]   D. Fisher, M. Smith, and H.T. Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, 00(C):59b–59b, 2006.

[GWLS09]   Eric Gleave, H.T. Welser, T.M. Lento, and M.A. Smith. A conceptual and operational definition of'social role'in online community. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–11. IEEE, 2009.

[KCH⁺11]   Andrey Kan, Jeffrey Chan, Conor Hayes, Bernie Hogan, James Bailey, and Christopher Leckie. A Time Decoupling Approach for Studying Forum Dynamics. *World Wide Web Internet And Web Information Systems*, In press:1–24, 2011.

[KRC⁺10]   Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The Effect of User Features on Churn in Social Networks. *Human Factors*, 2010.

[Mie11]   Piet Van Mieghem. Human Psychology of Common Appraisal: The Reddit Score. *IEEE Transactions on Multimedia*, 13(6):1404–1406, 2011.

[RA11]   Matthew Rowe and Sofia Angeletou. Predicting discussions on the social semantic web. *Web: Research and Applications*, 2011.

[RAAF12]   Matthew Rowe, Harith Alani, Sofia Angeletou, and Miriam Fernandez. Community Analysis through Semantic Rules and Role Composition Derivation. *Journal of Web Semantics*, 2012.

[SH10]   Gabor Szabo and Bernardo a. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80, August 2010.

[Vi04]   Fernanda B Vi. Visualizing the Activity of Individuals in Conversational Cyberspaces. *Source*, 00(C):1–10, 2004.

[WGFS07]   H.T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, 8(2):564–586, 2007.