Behaviour analysis across different types of Enterprise Online Communities

Matthew Rowe, Miriam Fernandez, Harith Alani
Knowledge Media Institute
The Open University,
Milton Keynes, MK7 6AA
United Kingdom
{m.c.rowe,m.fernandez,h.alani}
@open.ac.uk

Inbal Ronen
IBM Research Lab
Haifa 31905, Israel
inbal@il.ibm.com

Conor Hayes and Marcel Karnstedt
Digital Enterprise Research Institute
NUI Galway, Ireland
{firstname.lastname}@deri.org

ABSTRACT

Online communities in the enterprise are designed to fulfil some economic purpose, for example for supporting products or enabling work-collaboration between knowledge workers. The intentions of such communities allow them to be labelled based on their type - i.e. communities of practice, team communities, technical support communities, etc. Despite the disparate nature and explicit intention of community types, little is known of how the types differ in terms of a) the participation and activity, and b) the behaviour of community users. Such insights could provide community managers with an understanding of *normality* and a diagnosis of *healthiness* in their community, given its type and corresponding user needs. In this paper, we present an empirical analysis of community types from the enterprise social software system IBM Connections. We assess the micro (user-level) and macro (community-level) characteristics of differing community types and identify key differences in the behaviour that users exhibit in these communities. We further qualify our empirical findings with user questionnaires by identifying links between the objectives of the users and the characteristics of the community types.

Author Keywords

Community Analysis, User Behaviour, Enterprise Communities, Web Science

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors, Measurement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.
Copyright 2012 ACM 978-1-4503-0267-8/11/05...$10.00.

INTRODUCTION

Being a member of an online community is now an everyday experience for Web users. Communities serve multiple purposes, such as the collaborative editing of Wikipedia articles, photo-sharing, online gaming and personal exchanges. The study of online communities originally tended to focus on single online channels such as Usenet groups or bulletin boards, where communities were fairly well defined [11]. As such, there have been several studies on the functions of communities, quality of online ties, the role of identity and the incentive for users to participate [11, 5, 2]. With the growth in popularity of social networking platforms such as Facebook, Twitter and Reddit, the notion of community has become somewhat eclipsed by the notion of the social network. Recent research has tended to focus on the challenges and opportunities brought by the scale of these social networks, e.g. pattern detection and prediction of popularity and user actions [12, 13], influence detection [3], etc.

In this paper, we revert to the original notion of an online community as a group of people who use online communication systems to pursue mutual interests. Online communities have become an essential tool through which knowledge-workers collaborate and share information in the Enterprise. Unlike most typical public communities, Enterprise communities are organised around multiple channels, each of which offers different types of interaction. Common examples of such Enterprise community platforms include IBM Connections,[1] Cisco Quad,[2] Jive,[3] and Socialtext.[4] What channels are used in the communities on these platforms and how users interact will depend on the goals of each community. For example, a community focussed on the generation of new ideas may use wikis and blogs. A support community may rely on discussion fora. Likewise, a community dedicated to a product build will support itself through a different set of standard media channels [10].

[1] http://www-01.ibm.com/software/lotus/products/connections/
[2] http://www.cisco.com/web/products/quad/
[3] http://www.jivesoftware.com/
[4] www.socialtext.com/

Enterprise communities represent a significant investment in (social) capital and it is in the interests of community owners (and members) that they are adequately maintained and supported. Despite this, there is little research on how such multi-channel communities can be understood in terms of the typical interaction patterns and behaviour of members. Such insights would provide a basis to define typical and normative functionality and a means of detection and diagnosis of communities that are failing and may require additional support.

To address this, we carry out a quantitative and qualitative analysis of communities and community members from the IBM Connections platform, the current market leader in Enterprise Social Software. Our work explores the question: *How do enterprise community types differ from one another?*

We perform a quantitative macro (community-level) and micro (user-level) time-series analysis of the characteristics of many IBM communities of different types. Our objective is to profile each community type in terms of the behaviour of its users over time in order to understand how activities within communities differ in accordance to their type - defining behaviour as the tangible attributes of a user relative to the community that he participates in (e.g. engagement with other users, initiation of content, etc.). Through statistical analysis, we identify significant differences in user behaviour between the community types. Using statistical clustering, we assess the overlap of community types when partitioned by their features. Following our quantitative analysis, we qualify our findings by assessing the responses to user questionnaires from the different community types and identify links between the needs users have for the communities and the behaviour evident within the community.

The paper has been structured as follows: in section 2 we describe existing work intended to differentiate community types and social web systems from one another by user and community behaviour. In section 3 we describe the provided IBM Connections dataset and in section 4 we present the features, both macro and micro, used for our analysis. Section 5 describes our analyses and the results and findings gleaned from the results. Section 6 relates the findings from our empirical analysis with the questionnaire responses from the various community types. Section 7 details the conclusions drawn from this work.

RELATED WORK

The early research on online communities tended to focus on the properties of these new types of communities in relation to the relatively well understood behaviour of real world communities [11]. In this context, there have been several studies on the functions of communities, quality of online ties, the role of identity and the incentive for users to participate [11, 5, 2]. This research has tended to focus on single online channels such as Usenet groups or bulletin boards, where a community is understood as a group that uses an online communication channel to pursue mutual interests [11]. A similar definition of community underlies our work. However, our research focuses in multi-channel communities. Furthermore, our approach, while informed by previous work in the social sciences, takes a strong data analytics approach in order to uncover patterns of behaviour from user interaction data.

As such, our approach is strongly motivated by the recent data-driven approaches being applied to large scale online social network systems. For example, a number of recent papers provide comparative analysis at the system rather than the community level (e.g. comparing Digg with Youtube). The learning and prediction approach proposed in [8] assesses conversation datasets from Usenet, Yahoo! groups and Twitter as to how the platforms differ. Based on the gained insights, the proposed approach learns models to predict branching and authorship. [9] compares the social network properties of Flickr, YouTube, LiveJournal and Orkut. Interestingly, the authors find consistency across the platforms in terms of the distributions. However, compared to our work, the choice of features is rather limited as is the number of different community types. Tan et al. [13] predict social actions on the three different platforms (i.e. communities) Twitter, Flickr and Arnetminer. However, they consider only one representative type of social action in each platform. Similarly, [14] aims at predicting the volume of community activity on eight different news platforms and finds consistent patterns across platforms. Each of these approaches analyses user behaviour at the platform level, but does not take into account the notion of community. In contrast, our work provides an empirical analysis of user and community behavioural dynamics across several community types on a single platform.

There have been several studies on behavioural roles in online communities, motivated in part by the utility of summarising complex social systems in terms of well-founded behavioural signatures that allow the comparative study of different communities [7]. These roles are typically inferred from various features derived from the ego-centric network of users [4, 6, 16]. While this approach inspired our choice of features, it is usually based on a static view of the network whereas we take into account the time variant nature of behaviour.

Fernanda [15] proposed two novel visualisation tools that are suited for categorising (conversational versus non-conversational) Usenet forums. Although this work is mainly focusing on the underlying user roles, it takes a first step towards the suitability of such roles for categorising communities by their type. However, in contrast to our work, the considered types are rather limited and not confirmed by the actual users and owners of the communities. In a similar line, [1] proposed different methods for forum grouping based on communication patterns. The authors observed that users tend to have consistent conversational behaviour over time and used this knowledge to hierarchically cluster forums. However, the presented analysis is restricted to the post-reply behaviour as a single feature.

Our work continues a recent study of Muller et al. [10], who identified five different community types in the enter-

prise. Whereas Muller et al. identified different patterns of social media tool usage within each community type, we focus more on the behavioural aspect within those types.

DATASET: IBM CONNECTIONS

To perform our empirical analysis of community types, we were provided with a dataset from the enterprise social software suite IBM Connections. IBM Connections is used to promote and enable online communities within the enterprise. The software suite includes person profiles, collaborative bookmarking, wikis, blogs, file sharing, activities and discussion forums. The communities service in IBM Connections supports the collaboration of employees through the sharing of all these activities. IBM Connections is used by IBM's employees as well as by customers.

IBM Connections: Dataset

Under a non disclosure agreement, IBM provided the public data from the IBM Connections Intranet deployment that is accessible to all employees, to create the dataset analysed in this paper. Data of private communities were not imported. The dataset includes communities with their ID, creation date, members (with their roles) and applications such as blog entries, wiki pages etc. For forums it includes all the threads, comments, dates and related people, such as the initial author and responders.

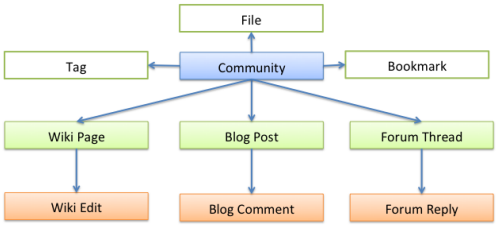


Figure 1. The structure of the IBM Connections dataset .

Figure 1 explains how the provided data is structured. As the diagram shows, a community can include several services simultaneously to be used by its members in the context of the community, including a community blog, a forum and a wiki. The community blog is composed of several blog posts, created by different authors. Each of these posts may receive multiple comments by different community members. In the same fashion, the community forum contains multiple forum entries. Each of these entries may be followed by a chain of replies, representing the discussion that has taken place within a thread. Wikis are slightly different in the sense that, although they are composed of multiple pages, wiki pages are neither commented on nor replied to but are instead edited multiple times by different community members.

IBM Connections: Community Types

In recent work, Muller et al. [10] analysed IBM Connections to distinguish how the functionalities and technologies adopted by community types differ. Having assessed the literature related to types of communities, the authors listed five distinct community types, each differing in their role and intention. These were defined as follows:

- *Communities of Practice (CoP)* A group of people with a common interest or practice who share information and/or network.
- *Teams (Team)* Communities working on a shared goal for a particular client, project or business function.
- *Technical Support Groups (Tech)* Providing technical support for a particular technology.
- *Idea Labs* Communities in which members brainstorm around a set of questions or issues for a limited period of time, usually as part of a client engagement.
- *Recreation* Communities devoted to recreational activities unrelated to work.

Muller et al. dispatched a questionnaire to community owners on IBM Connections asking them to categorise their communities.[5] The options were multi-label, such that a community owner could select multiple types for her community - e.g. the owner could label her community as being both a *Community of Practice* and an *Idea Lab*. In order to provide a unary relation between a community and its type a *majority opinion* was taken between three raters.[6] We were provided with a mapping of the 186 most active communities to their types, covering three of the above types. These are, with the number of communities of each type in brackets: *CoPs* (100), *Teams* (72) and *Techs* (14). We use these 186 communities and their data for our analysis, and use the predefined community types to complement the work of Muller et al. [10].

FEATURE ENGINEERING

Our approach towards understanding how community types differ requires defining common community attributes and then assessing how community types differ across those attributes. To achieve such comparisons, we implemented two feature sets, one to capture a community's *macro* attributes and another to capture *micro* attributes of each community. As the names imply, in the former case we assess the general properties of the community, while in the latter case we assess the low-level behaviour within the community - i.e. measuring the behaviour exhibited by each community user.

In order to measure the macro and micro features of a community we need to engineer the required features from our provided data. As we mention in the previous section, IBM Connections allows users to create a community based around a central topic or goal. Within the community users may *initiate* content by creating a blog post or a forum thread, or adding a page to a wiki. Users may also *contribute* by commenting on a blog post, replying to a forum thread or editing a wiki page. Although other actions are possible, for example sharing a file or a bookmark, we utilise the forum, blog and wiki data from which to compile the micro (i.e. user-level) features, given the interactive qualities of such data items.

[5]N.b. The respondents to this questionnaire are distinct from the community users that replied to the later described user needs survey. In this instance the owners of the communities were asked specific questions of their community's intentions and types.

[6]The interrater agreement was 0.84 ($F = 11.53$, $p < 0.001$) using the Cronback alpha measure.

Macro Features

For the macro features we wish to describe the attributes of each community through the use of statistical features. To engineer the features we use a *sliding window* method by beginning at a *collect date*, in our case the 1st January 2010, and deriving a *feature window* that extends 6-months (184 days) prior to the *collect date*. Within the *feature window* we then measure the community features given below. The *collect date* is then moved forward one month and the *feature window* is therefore applied over a different time period. Once again we measure the features for the community within the time period. This process of moving the *collect date* and *feature window* is repeated until we reach the end of the data (April 2011).

In vector form we produce an instance x_t for each time step t throughout the range of the 16 *collect dates*. The resulting dataset, for each community $c \in C$, is a 16×3 matrix with 16 rows, one for each of the 16 time steps (i.e. collect dates), and 3 columns, one for each of the three macro features measured over those time steps. We chose three macro features for our analysis, each of which was common across the community types and their inherent functionality (wikis, forums, blogs, etc.) and would allow comparisons with data from other platforms in the future. These were defined as follows:

1. **Seed Post Count:** We define a *seed post* as any piece of initiated content within a given community that receives an additional contribution. For instance, a blog entry is a *seed* if it receives a comment, an initial post in a forum thread if it receives a reply and a wiki page if it is edited. We included this feature in order to measure the extent to which content yields contributions by community users. To measure the **Seed Post Count** for a given community, we count how many blog entries, forum threads and wiki pages are seeds within the given *feature window*.
2. **Non-seed Post Count:** Converse to *seed posts* we define *non-seed posts* as those initiations that receive no contribution - e.g. a blog entry with no comments, a forum thread with no replies or a wiki page with no edits. We included this feature to measure the number of initiations that incur no contributions from the community. To measure the **Non-seed Post Count** for a given community, we count how many blog entries, forum threads and wiki pages are non-seeds within the given *feature window*.
3. **Users:** This feature measures the number of users who have participated with the community within a given time period. Using this feature, we can investigate whether different community types have differing user numbers. To gauge the **User Count**, we count how many users have participated with the community, either by initiating content or contributing to existing content, within the allotted *feature window*.

Micro Features

For the micro features our goal is to capture the behaviour of individual users within a given community. If we measure the behaviour of the community users over time, we can then gather a longitudinal perspective of behaviour. To derive the features described below, we used the same *sliding window* approach as above for the macro features, by starting with the *collect date* of 1st January 2010 and using the 6-months prior to this date as the *feature window*. Within the *feature window*, we list all the users who participated within the community and then use all the past actions by each user within the window to derive the micro features - the derivation of these features is described below. We do this until the end of the dataset - i.e. April 2011 - over 16 time steps.

In vector form at each time step we produce an instance x_i for each user v_i that has participated in community $c \in C$. The same user may appear in multiple time steps, therefore we may have multiple instances for the same user, but with different behaviour features - given that their behaviour is likely to change over time. The resulting dataset is an $n \times 5$ matrix for each community $c \in C$, n rows corresponding to the number of unique user instances produced over the collect dates and five features in the columns - e.g. if there are 10 users each of whom appear in the 16 time steps then the dimensionality of the matrix will be: 160×5. We derive the features as follows:

1. **Focus Dispersion:** The focus of a user v_i is a measure of her concentration of activity across multiple communities on IBM Connections. If the user is *focussed* then she will have a lower value, whereas if her activity is broad and she interacts with many communities then focus will be *distributed*. To gauge this feature for a given user (v_i), we take all the initiations and contributions that the user has made within the *feature window* as the combined set of posts (P_{v_i}) and assess the community in which each post has been made. Let C_{v_i} be all the communities that user v_i has posted in and $p(c_.|v_i)$ be the conditional probability of v_i posting in community $c_.$. We can derive this using the post distribution (P_{v_i}) of the user, therefore we define the **Focus Dispersion** as the community entropy (H_C) of a given user:
$$H_C(v_i) = -\sum_{j=1}^{|C_{v_i}|} p(c_j|v_i) \log p(c_j|v_i) \qquad (1)$$
2. **Initiation:** This feature measures how many pieces of content the user creates for the community. Initiations can be the creation of a forum thread, the creation of a blog entry or the creation of a wiki page. In each case we differentiate the initial creation stage from replying to such content in order to assess the extent to which a user proactively engages with the community. This feature is derived from summing the number of forum threads, blog entries and wiki pages created by the user of a given community (c) within the *feature window*.
3. **Contribution:** The contribution of a user is the extent to which the user interacts with existing content by contributing to it. We use the abstract notion of contribution in order to encapsulate the actions of replying to a forum thread, commenting on a blog post or editing a wiki page. Therefore, this feature is derived from the sum of all thread replies, blog comments and wiki edits by the

community user within the *feature window*.

4. **Popularity:** The popularity of a user assesses the extent to which other users interact with the user via her initiated content. For instance, if a user υ_i initiates a blog post, then the number of unique users replying to that blog entry gauges the popularity of that piece of content. Assessed over all initiations by υ_i, this provides a summary metric of the user's popularity. In essence popularity is the *in-degree* of υ_i measured through content interactions. As the dataset provided for our experiments does not contain explicit social network connections (e.g. the addition of a *friend* link between two users), we must rely on the interaction graph.
   Let $\Upsilon_{in,i}$ be the set of users that have contributed to content initiated by υ_i within the *feature window*, we therefore derive the **Popularity** of υ_i as $|\Upsilon_{in,i}|$.

5. **Engagement:** Converse to the popularity of each community user - i.e. the extent to which other community users interact with them - we also wish to measure how many users a given user has interacted with through content initiated by them. In essence this measures the *out-degree* of υ_i. To do this, we gather all the contributions that user υ_i has made to the community within the *feature window* and then identify the authors of the initiated content that those contributions were towards.
   Let $\Upsilon_{out,i}$ be the set of users that υ_i has interacted with through their initiated content within the *feature window*, we therefore derive the **Engagement** of υ_i as $|\Upsilon_{out,i}|$.

ANALYSIS OF BEHAVIOUR DYNAMICS

The differing intentions of enterprise communities allow them to be categorised based on their type - i.e. communities of practice, team communities, technical support communities. Despite the disparate nature and explicit intention of community types, little is known of how the types differ in terms of activity patterns and user behaviour. In this section, we use the provided dataset from IBM Connections and the aforementioned macro and micro features to empirically analyse how community types differ.

Experimental Setup

For each community within the provided community-to-type mapping, we measured the macro and micro features over the allotted *collect dates*, thereby building a dataset for each community for each feature set. These datasets were used for two analysis tasks geared towards exploring the central research question defined within the introduction: *How do enterprise community types differ from one another?* We now explain the experimental setup and motivation behind each task:

1. *Analysis of Behaviour Distributions:* The first task assessed how the macro and micro features differed between community types. We generated a single macro and micro feature dataset for each of the three community types: *Community of Practice (CoP)*, *Team* and *Technical Support (Tech)*, by taking all the communities for each type, based on the community-to-type mapping, and combining the individual community datasets into a single community-type dataset. From these datasets we then assessed the distribution of each feature, both macro and micro, within the different community types by: a) examining the mean and standard deviation and using Wilch's t-test to assess the statistical significance of the differences, b) plotting the empirical cumulative distribution function of each feature within each community type, and c) measuring the deviance between the feature distributions and communities using the Kolmogorov-Smirnov test. In doing so, we could identify how features differed between community types in terms of variance and skew.

2. *Analysis of Community Partitions:* The second task assessed the partitioning of communities into the clusters of their respective types. To do this, we generated a *community motif* for each community across both the micro and macro behaviour features by averaging the feature values of every instance in the micro and macro datasets, respectively. For each community we produced a 1×5 and 1×3 vector for the micro and macro features respectively. Using the *community motifs*, we then assessed the partitioning of the communities into their respective type clusters, investigating the purity of the segmentation and whether there was a noticeable overlap. For this we used principal component analysis and assessment of the partitioning within an n-dimensional vector space - where n denotes the dimensionality of the data under inspection, being $n = 5$ and $n = 3$ for micro and macro features, respectively.

Results: Behaviour Distributions

Macro Features

Inspecting the distribution properties (the mean and standard deviation) of macro features - i.e. the community-level dynamics - in Table 1, we see that *Team* communities have the most **Seeds** - i.e. pieces of initiated content that yield interactions - while also having the highest mean for **Non-seeds**, however the differences in the means are not found to be significant. In such communities the need to collaborate as part of a team requires content to be shared with community members and develop collaborations. The magnitude at which such content is created with respect to the remaining community types indicates the extent to which this occurs. We also find, in Table 1, large standard deviations in the communities for the various types. This is due to the differences in collaborative environments and communities, we find the deviation to be greatest for seeds. Table 1 also indicates that the *Tech* communities have the highest average number of **Users** - this is found to be significantly higher than the other communities at $\alpha < 0.05$ - suggesting that in such communities the creation of content is dispersed much more evenly across individual members, with fewer posts on average per user.

To provide a greater insight into how the distributions differ between the community types, we induced empirical cumulative distribution functions (ECDFs) for each of the macro features in each community type's respective dataset. The empirical cumulative distribution function derives the probability distribution of univariate data, representing a single

feature in our case, as the value limit is iteratively increased. In essence, it allows us to see how a feature is distributed across its values and whether there is a skew towards the feature being lower or higher in different community types. Let $I_{x_i \leq t}$ define an indicator function that returns 1 if the value of x is less than or equal to t and 0 otherwise, then the ECDF is defined, using an increasing value range for t between the features *minimum* and *maximum* value, as:

$$\frac{1}{n}\sum_{i=1}^{n} I_{x_i \leq t} \tag{2}$$

We omit the ECDF plots for the macro features, as they show no clear differences between the distributions. Indeed, we found the behaviour of the different community types to appear consistent when observing the macro-level attributes of a community. To assess the quantitive differences between the distributions, we used the Kolmogorov-Smirnov two-sample test to compare, in a pairwise fashion, the induced ECDFs - e.g. comparing the distribution of **Seeds** between *CoP*, *Team* and *Tech*. This test returns the *maximum deviation* between the distributions and the *p-value* of the divergence, thereby allowing us to gauge the significance of the divergence.

Table 1. Mean and Standard Deviation (in parentheses) of macro-features within the different community types

Feature	CoP	Team	Tech
Seeds	7.094 (15.601)	**7.128 (15.622)**	6.680 (13.076)
Non-seeds	3.298 (9.418)	**3.397 (9.594)**	3.390 (8.896)
Users	4.041 (6.669)	4.024 (6.616)	**4.172 (6.767)**

The differences between the feature distributions are presented in Figure 2, where we compare the ECDF of each of the macro features. The bar charts indicate the lack of deviance between the distributions. The largest appears to be where **Seeds** are concerned, as there is a marked difference between the *Tech* communities and the two other types - this difference is found to be significant at $\alpha = 0.05$. We also find the difference between Tech communities and the other types to be significant, again at a significance level of $\alpha = 0.05$, when assessing the **Non-seeds** distribution. As we have demonstrated, the differences between the communities in terms of their macro features are minimal, in particular when considering the empirical cumulative distribution functions. In the next section, we extend this analysis to the behaviour exhibited by community users and how that differs between the types of communities, thereby delving deeper into the implicit dynamics of the communities.

Micro Features

We now inspect the differences between community types in terms of the micro features. Table 2 contains the mean and standard deviation for each community type and feature. For **Focus Dispersion** we find that *CoP* has the highest value - significant at $\alpha < 0.001$ - indicating that users of that type of community tend to disperse their activity across many different communities. Conversely, for *Tech* communities this value is lowest, where users are focussed on just participating in a selection of communities. For **Initiation**, Table 2 indicates that *Team* communities have a much higher mean (and standard deviation) than the other community types -

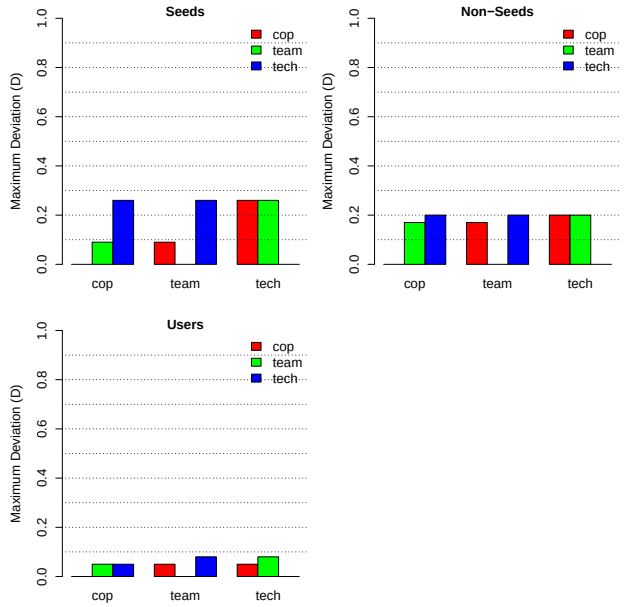


Figure 2. Maximum Deviation (D) between ECDFs from disparate community types and macro features, measured using the Kolmogorov-Smirnov test

also significant at $\alpha < 0.001$. This could be due to such communities requiring users to work together, often on a shared goal, such as developing a product for a client, therefore more ideas are shared through forum posts and blog entries.

The mean of the third micro feature, **Contribution**, is highest for *CoP* (but not significantly higher than the others) indicating that more initiated content is interacted with than in the other communities. **Popularity** is higher in *Team* and *Tech* communities, but not significantly, than in *CoP*, suggesting that although users of the latter community provide more contributions, it is with content published by fewer users. For **Engagement** the mean is significantly highest - at $\alpha < 0.001$ - for *Team* indicating that users tend to participate with more users in these communities than the others.

Table 2. Mean and Standard Deviation (in parentheses) of the distribution of micro features within the different community types

Feature	CoP	Team	Tech
Focus Dis'	**1.682 (1.680)**	1.391 (1.581)	1.382 (1.534)
Initiation	7.788 (21.525)	**13.235 (23.361)**	3.088 (6.676)
Contribution	**26.084 (77.607)**	21.130 (72.298)	11.753 (17.182)
Popularity	1.660 (3.647)	**2.302 (2.900)**	2.286 (3.920)
Engagement	1.016 (1.556)	**1.948 (2.324)**	1.036 (1.575)

We induce an empirical cumulative distribution function (ECDF) for each micro feature within each community and then qualitatively analyse how the curves of the functions differ across communities. For instance, in the case of Figure 3 we see that for **Focus Dispersion** *Tech* communities have the highest proportion of focussed users (i.e. where entropy is 0). This indicates that users are interested in concentrating in those communities alone for discussing support requests and asking/answering questions to specific topics. For *CoP* the users are more dispersed, indicated by the low proportion of users who have an entropy of 0 and the low curve of this

community's ECDF as entropy increases. For **Contribution** the probability mass for *Tech* communities is skewed to lower values than the two remaining community types, indicating that initiated content is interacted with less. As expected, the distribution for *Team* communities is skewed towards higher values, given that users contribute to existing content in order to achieve a shared goal and work together.

The distribution for the micro feature **Initiations** shows that more users create content in *CoP* and *Team* communities, while for *Tech* communities the majority of users initiate less content compared to the other two communities - i.e. the ECDF reaches 1 earlier. For **Popularity** we find similar curves for the distributions, while for **Engagement** the *Team* communities' users are found to engage with more users than the other two types of communities, as would be expected in a community driven by collaboration.

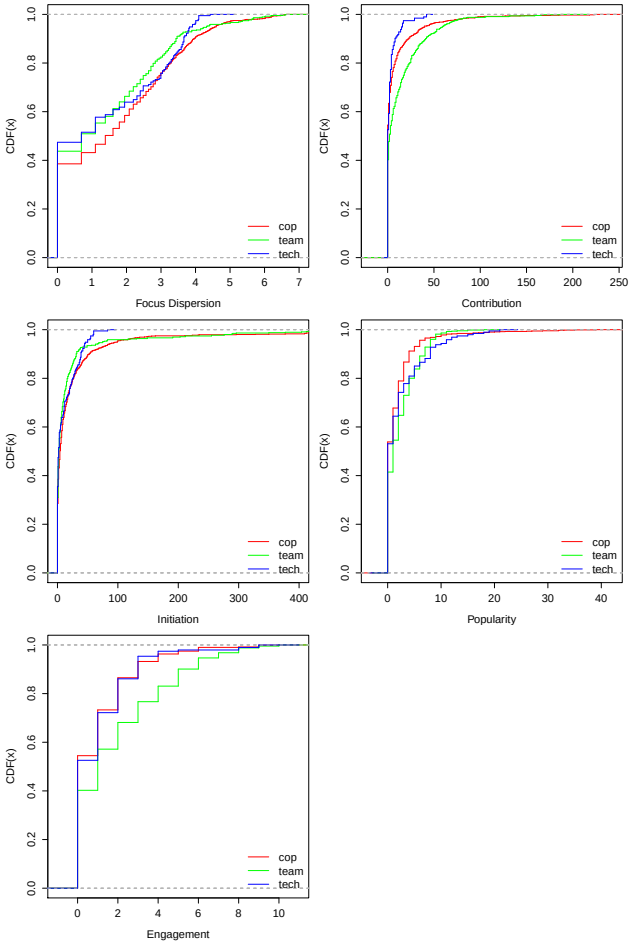


Figure 3. Plots of the Empirical Cumulative Distribution Function for each community type and micro feature

Figure 4 presents bar charts of the pairwise deviations between the community types for the five different micro features using the Kolmogorov-Smirnov test. For **Focus Dispersion** we find *Tech* community users to be distinct from other types, where the distribution is skewed to lower values (see Figure 3) while having more focussed users (the *deviation* between the *Tech* community type and the others is significant at $\alpha = 0.001$). We also find that users of *Tech* communities differ significantly (also at $\alpha = 0.001$) for **Contribution** and **Initiation**, where such users do not contribute as much as the other community types, nor do they initiate as much content. For **Popularity** and **Engagement** all community types begin to differ. For instance, Figure 4 indicates that users of *CoP* are distinct from two other community types in that their users interact with fewer other users (the divergence of **Popularity** is significant at $\alpha = 0.001$), while *Team* and *Tech* community users are distinct from one another in terms of **Engagement**, where users of *Team* communities engage more with other users.

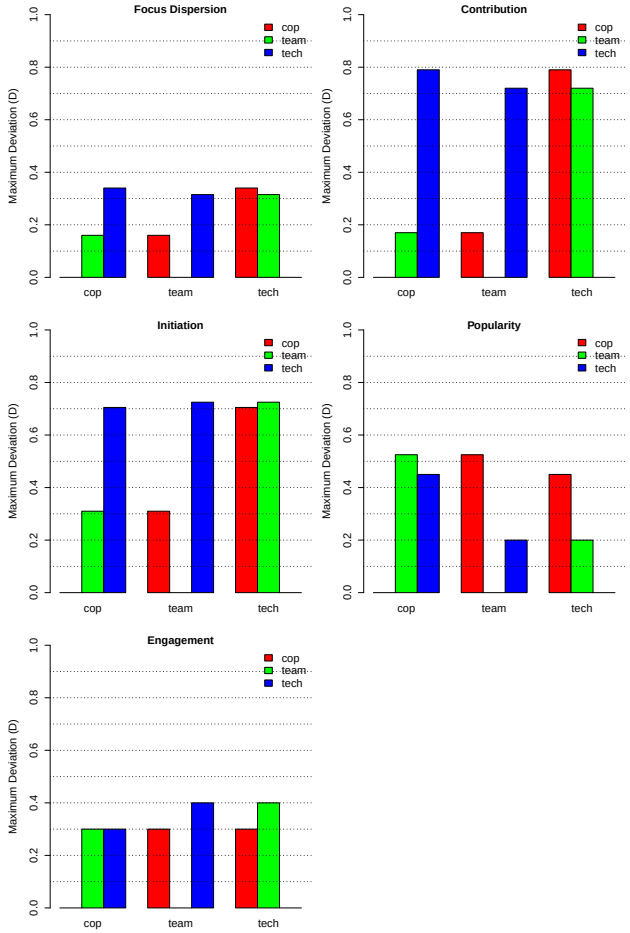
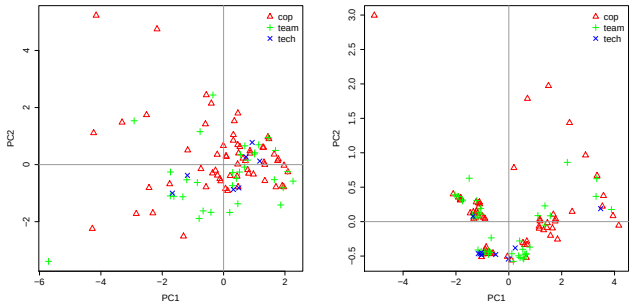


Figure 4. Maximum Deviation (D) between ECDFs from disparate community types and micro features, measured using the Kolmogorov-Smirnov test

The higher-level assessments achieved through the macro-features do not reveal large differences between the community types. Indeed, the empirical cumulative distribution functions have similar curves that, according to the Kolmogorov-Smirnov tests, have only small deviances from one another. By analysing the micro-features in different community types we reveal significant differences in how users behave across all the examined features. This deeper exploration into the dynamics of community types could be driven by the different needs that users of disparate types of communities bestow upon them. We explore this thesis through users questionnaires later in the paper.

Results: Community Grouping

For the second analysis task, we sought to group clusters into their respective types and then assess the *purity* of this grouping. To do this, we constructed a *community motif* for each community, where one *motif* was built from the micro features and a second from the macro features. To build the motif, we took the mean for each feature in each community's dataset, thereby producing a single vector representation of a given community. To begin with we performed Principal Component Analysis (PCA) in order to generate a qualitative view of how the communities were grouped.



(a) Micro (b) Macro

Figure 5. Principal Component Analysis plot of the communities distributed based on a) micro features and b) macro features

Figure 5(a) and Figure 5(b) show the PCA clustering for the micro and macro features respectively. We note that for the micro features the *CoP* and *Team* communities are more dispersed than the *Tech* communities, indicating that there are intra-type differences in the behaviour of users within the two former community types. For macro features, in Figure 5(b), we find that the *Tech* communities are largely contained within a central cluster with a few outliers. As with the micro features, *CoP* communities are more dispersed in the plot rather than being clustered together.

The qualitative perspective that PCA plots afford provides an insight into the dispersion of community types. However, we do not know the *quality* of the clustering - i.e. how well the communities are partitioned based on their types. To quantify this, we use the *silhouette coefficient* produced when the communities are grouped into their respective type clusters using the micro and macro features. We define the silhouette coefficient (s_i) for a given community as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{3}$$

Where a_i denotes the average distance between all items in the same cluster and b_i is given by calculating the average distance to all 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 clustering where distinct items are grouped together, and therefore *misclassifications* would be made, and the latter indicates perfect cluster cohesion and separation. To derive the silhouette coefficient for the entire clustering, we take the average silhouette coefficient of all items. When calculating the silhouette coefficient with the *Euclidean distance* as the distance measure, using the micro features we yield -0.441 and for the macro features we yield -0.130. These numbers indicate that using macro features achieves a *purer* distinction between community types, while there are commonalities in terms of user behaviour between the community types - characterised by the low silhouette coefficient for the micro features.

To further assess the differences between the community types, we constructed centroid vectors for each community type from the *community motif* of each of their communities. The similarity between cluster centroids was then gauged by measuring the *Euclidean distance* between the centroids, where a lower distance indicates a greater similarity. The results from the centroid comparisons are presented in Table 3, demonstrating that the centroid for *CoP* appears further away from the two other community types, based on micro features, while the *Tech* centroid is placed further away in terms of macro features. This latter finding confirms what is shown in the PCA plot based on macro features, in Figure 5(b), where *Tech* communities are, largely, contained within a central cluster, while the communities for *CoP* and *Team* are more dispersed.

Table 3. The Euclidean distance between the community type centroids using the micro and macro features

(a) Micro Features

	CoP	Team	Tech
CoP	0.000	14.534	21.149
Team	14.534	0.000	7.652
Tech	21.149	7.652	0.000

(b) Macro Features

	CoP	Team	Tech
CoP	0.000	0.281	0.569
Team	0.281	0.000	0.365
Tech	0.569	0.365	0.000

USER NEEDS AND COMMUNITY DYNAMICS

In order to further qualify the results of the aforementioned analysis, we used a questionnaire designed to collect insights into the usage of the online communities on IBM Connections. This questionnaire was designed in the context of the ROBUST[7] project and is aimed at understanding the users' personal needs for using an online community and what they value in a community and its members.

The complete questionnaire can be found online[8] and consists of 20 carefully selected questions. In questions where users had to express the degree to which they perform an activity or agree with a statement, we used the five point Likert-type scale to capture the responses. To analyse the results, we translated these options to a numeric scale of 1 to 5 with 1 representing *Never*, *Strongly disagree* and *Completely irrelevant*, and 5 representing *Very Often*, *Strongly agree* and *Very important*. The questionnaire was circulated to nearly 4,000 users of IBM Connections communities. We received 186 responses[9], of which 150 were complete, covering 53 different communities. The complete responses were divided according to their community type resulting in 95 completed questionnaires for *CoP* communities, 33 for *Team* commu-

[7] http://robust-project.eu/

[8] http://socsem.open.ac.uk/limesurvey/index.php?sid=55487

[9] N.b. Originally we received 197 responses but for 11 we could not identify the community type due to the usage of the prior mapping file

nities and 58 for *Tech* communities. We did not circulate the questionnaire to any *Idea lab* or *Recreation* communities as they were extremely sparse in our dataset.

To contrast the results of the previous empirical analysis with the user needs extracted from the questionnaire, we first performed a mapping between each micro feature - e.g. **Initiation**, **Contribution**, etc. - and a subset of questions describing the feature. We chose to omit macro features due to the limited insights that such features provided. We have placed the mapping online for the reader's benefit[10]. For example, **Initiation** was described using questions like: *How often do you ask a question? How often do you create content? How often do you announce work news and events?*, etc. Given this mapping we could then derive an average score for each micro feature based on the questionnaire responses. Due to our use of the Likert-type scale such averaging was feasible by taking the response values (given that these ranged from 1 to 5) and taking the mean over those for all community type responses - e.g. taking the mean of the responses for all **Initiation** questions for the 95 *CoPs*. The set of results can be seen in Table 4.

Table 4. Mean and standard deviation (in parentheses) values of micro-features obtained using the questionnaires for the different community types

	CoP	Team	Tech
Focus Dis'	4.019 (0.093)	3.055 (0.426)	**4.070 (0.070)**
Initiation	2.483 (0.838)	**2.587 (0.838)**	2.243 (0.873)
Contribution	**3.239 (0.926)**	3.202 (1.016)	3.158 (0.945)
Popularity	2.875 (0.070)	**3.084 (0.168)**	2.104 (0.173)
Engagement	2.844 (0.539)	**3.027 (0.588)**	2.406 (0.522)

As Table 4 demonstrates, the findings from the analysis highly correlate with what users expressed to be relevant for each community type. We previously found that high levels of **Initiation** and **Contribution** are discriminative factors of *Team* and *CoP* communities with respect to *Tech* communities. Additionally, by looking at the behaviour distributions of these features, we find that higher levels of **Initiation** are more common for *Team* communities, while higher levels of **Contribution** are more common for *CoP* communities. Collaboration is a strong element for both community types, either for sharing common interests as in the case of *CoPs* or for sharing a common task or goal in the case of *Teams*. However, in *Team* communities the collaboration is driven by the task, and this may require frequent uploads of pieces of work to the community (in the form of wiki pages, blog entries or forum announcements) given the higher level of **Initiation**. On the other hand, *CoPs* are driven by the need to share common interests or practices and therefore discussions about the content posted in a blog, wiki, or forum thread constitute a more relevant factor. This correlates with our findings from the macro features analysis, where *Team* communities have the highest levels of **seed** and **non-seed** posts (i.e. posts that do not generate a reply). As mentioned before, in these communities content initiations may be done as part of the task, but not with the aim of generating a discussion. As Table 4 describes, user needs corroborate these facts. Average numbers for **Initiation** and **Contribution** are higher for *CoPs* and *Team* communities than for *Tech* communities. Additionally, we also see that users consider **Initiation** a more relevant factor for *Team* communities, while **Contribution** is considered a more relevant factor for *CoP* communities.

Another insight that emerged from the analysis, and is corroborated by the user questionnaires, is the fact that, over the three different community types, *Team* communities show the highest levels of **Initiation**, **Popularity** and **Engagement**. By intuition, in *Team* Communities each member needs to interact with other members of the team in order to achieve their common goal, a key collaborative property that is missing from the two remaining community types. These interactions across team members make **Popularity** and **Engagement** discriminative factors of *Team* communities. Moreover, as shown in Figure 4, while **Contribution** and **Initiation** discriminate *CoPs* and *Team* communities from *Tech* communities, **Popularity** is the factor that better discriminates *CoPs* from *Team* communities.

For **Focus Dispersion** the findings from the analysis and the questionnaire differ slightly. Our analysis and the users' opinions agree on the fact that **Focus Dispersion** is a discriminative factor for *CoP* communities, i.e. users of *CoP* community tend to disperse their activity across may different topics. In the *Tech* communities the diversity of expertise was found to be a valuable community attribute in the questionnaire responses, one would therefore have anticipated that the mean of **Focus Dispersion** would be higher than for other community types in our previous empirical analysis task - see Table 2. However, this was not the case. The reason for this is the derivation of **Focus Dispersion**, given that this feature was engineered by using all posts by a user such that the content she initiated - e.g. creating a wiki page - and contributed to - e.g. editing a wiki page - was pooled together. As a consequence, initiations could bias the mean of the distribution for *Tech* communities. For example, it is common that users who initiate a forum thread are asking for information, but do not share the knowledge of the community - i.e. users who are novices for the particular community topic. As future work we plan to divide the distributions explored previously into technology-dependent micro features, thereby yielding a **Focus Dispersion** measure for forum replies that captures the diversity of topics that users responding to forum threads have - such replies often denote answers in *Tech* communities.

CONCLUSIONS AND FUTURE WORK

Enterprise communities are provided to support a variety of purposes with the common ground of economic benefit. Previous work by Muller et al. [10] divided enterprise communities on IBM Connections into distinct types, finding that each community type had a specific intention and pattern of social media tool usage. In this paper, we explored the question: *How do enterprise community types differ from one another?* We performed both quantitative and qualitative analyses and in doing so have provided insights into the differences between community types and how those are re-

[10] http://socsem.open.ac.uk/WebScience2012/Association-of-microfeatures-with-questions.html

lated to the needs that community users have for the differing types.

Empirical analysis of the micro - i.e. user level - and macro - i.e. community level - features of three different community types (*CoP*, *Team* and *Tech*) identified differences in the characteristics of those types. We found that users of *CoP* communities were more dispersed in their activity, by visiting and interacting with more different communities than the two other community types. We also found that users of *Team* communities initiated more content than other community types, suggesting that the creation of new content in such communities is more commonplace.

Through a questionnaire disseminated to users from the three community types we enquired as to what needs users had for the online community in which they participated in on IBM Connections. The questionnaire responses showed common patterns between the empirical analysis of the micro features and how the users used the communities and what they required. For instance, we found that the ability to initiate content was most important in *Team* communities, while initiation behaviour was highest among users of the community in the analysed data. Likewise, we found the ability to contribute to existing content to be the most important in *CoPs*, while the contribution behaviour - i.e. editing a wiki, commenting on a blog posts or replying to a forum thread - was highest in the empirical analysis for that community type.

Our future work will explore two avenues. The first concerns the inference of community types given new emerging communities. We can utilise the knowledge gained from the presented work combined with the insights of [10] to identify discerning features for specific community types, for instance by comparing the micro feature distributions of a given community and using the behaviour of its users to infer the communities type, and therefore implicit needs. The second avenue of work will be to explore how the satisfaction of the needs of community users can be measured. We touched upon this in the discussion section in which we identified incongruity between the empirical observations for the focus dispersion of community users and the prevalent need for diverse expertise in *Tech* communities. Understanding how the needs of different community types can be assessed will ultimately allow the success or failure of communities to be judged, a concept which, at present, still remains fuzzy - particularly when considering community *health*.

ACKNOWLEDGEMENTS

The work of the authors was supported by the EU-FP7 project Robust (grant no. 257859) and the Science Foundation Ireland Grants LION-2 (SFI/08/CE/I1380) and Clique(08/SRC/I1407). Authors would like to thank Dr Bernie Hogan from the Oxford Internet Institute for his valuable feedback and help with the survey questionnaire.

REFERENCES

1. Andrey Kan, Jeffrey Chan, Conor Hayes, Bernie Hogan, James Bailey, C. L. A Time Decoupling Approach for Studying Forum Dynamics. *World Wide Web Internet And Web Information Systems In press* (2011), 1–24.
2. Butler, B., Sproull, L., Kiesler, S., and Kraut, R. Community effort in online groups: Who does the work and why? *Human-Computer Interaction Institute* (2007), 90.
3. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2010).
4. Chan, J., Hayes, C., and Daly, E. M. Decomposing Discussion Forums and Boards Using User Roles. *Forum American Bar Association* (2010), 215–218.
5. Cummings, J., Butler, B., and Kraut, R. The quality of online social relationships. *Communications of the ACM 45*, 7 (2002), 103–108.
6. Fisher, D., Smith, M., and Welser, H. 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 (2006), 59b–59b.
7. Gleave, E., Welser, H. T., Lento, T. M., and Smith, M. A. A Conceptual and Operational Definition of 'Social Role' in Online Community. *Sciences-New York* (2009), 1–11.
8. Kumar, R., Mahdian, M., and McGlohon, M. Dynamics of conversations. In *SIGKDD*, KDD '10 (2010).
9. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. In *SIGCOMM conference on Internet measurement*, IMC '07 (2007), 29–42.
10. Muller, M., Ehrlich, K., Matthews, T., Perer, A. A., Ronen, I., and Guy, I. Diversity among enterprise online communities: Collaborating, teaming, and innovating through social media. In *The 2012 ACM SIGCHI Conf. on Human Factors in Computing Systems* (2012).
11. Smith, M., and Kollock, P. *Communities in cyberspace*. Psychology Press, 1999.
12. SzaBo, G. Predicting the popularity of online content. *communicaTionS of The acm 53*, 8 (2010).
13. Tan, C., Tang, J., Sun, J., Lin, Q., and Wang, F. Social action tracking via noise tolerant time-varying factor graphs. In *SIGKDD*, KDD '10 (2010), 1049–1058.
14. Tsagkias, M., Weerkamp, W., and de Rijke, M. Predicting the volume of comments on online news stories. In *CIKM*, CIKM '09 (2009), 1765–1768.
15. Vi, F. B. Visualizing the Activity of Individuals in Conversational Cyberspaces. *Source 00*, C (2004), 1–10.
16. Welser, H. T., Gleave, E., Fisher, D., and Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. 1–32.