Mixed Membership Models for Exploring User Roles in Online Fora*

Arthur White
School of Mathematical Sciences, University College Dublin, Ireland

Jeffrey Chan
Department of Computing & Information Science, University of Melbourne, Australia

Conor Hayes
Digital Enterprise Research Institute, NUI Galway, Ireland

Thomas Brendan Murphy
School of Mathematical Sciences, University College Dublin, Ireland.

Abstract

Discussion boards are a form of social media which allow users to discuss topics and exchange information in a complex manner, in a number of different settings. As the popularity of such message boards has increased, communities of users have emerged, and several prominent types of social role have been identified, such as Question Answerer, Celebrity, Discussion Person and Topic Initiator. Recent studies have noted the structural similarity of the egocentric network of users assigned the same role by qualitative criteria. In this paper a methodology is developed with which to cluster together users with similar ego-centric network structures. This is achieved using a mixed membership formulation which allows for the fact that different groups of users may have characteristics in common. The method is then applied to data taken from boards.ie, a medium sized message boards website. Prominent clusters of users are identified and discussed, and illustrative examples of user behaviour provided. The type of interaction, both locally and globally, taking place within forums is examined.

Introduction

In recent years, the substantial increase in usage of Web 2.0 applications has facilitated digital interaction between users in an unparalleled manner, providing those who were formerly mass information consumers with the means to become information providers (Agarwal et al. 2008). Many types of information exchange can take place in such a setting: for example, users can socialise and exchange ideas about topics of interest, and the medium can also be utilised by companies to extend off-line customer support and manage knowledge. Examples of such settings include wikis, blogs, message boards, social media and many others. As applications have developed, communities of participatory users have emerged, with users displaying distinctive patterns of behaviour in media such as Usenet newsgroups (Golder and Donath 2004; Fisher, Smith, and Welser 2006), boards.ie (Chan, Hayes, and Daly 2010), and wikipedia (Welser et al. 2011). While such behaviour may be observed directly (Golder and Donath 2004), the self-documenting, electronic nature of the medium means that the communication patterns of users can be empirically analysed using tools from social network analysis (Fisher, Smith, and Welser 2006; Chan, Hayes, and Daly 2010; Welser et al. 2011).

*This work is supported by Science Foundation Ireland under the Clique Strategic Research Cluster (08/SRC/I1407).

The User Role in Electronic Media

We investigate and develop methods to identify and cluster the social roles occupied by users participating in online discussion. Even in an online setting in which formal roles exist, informal yet distinct social roles have been identified (Golder and Donath 2004; Welser et al. 2011). In attempting to discover social roles in communities, two general methodologies, interpretative and structural, have been developed (Gleave et al. 2009).

Interpretative metholodologies emphasise the qualitative analysis of social interaction, whereby a social role is defined as a a combination of factors which place constraints on behaviour (Golder and Donath 2004), with the expectations of other individuals sharing their social context considered alongside a person's own skills and social abilities. Within the paradigm of structural methodology, a person's role is defined by their network structure (Wasserman and Faust 1994, Chapters 9 and 10). Methods such as blockmodelling can be used to describe the interaction between clusters of users partitioned together with respect to some class of equivalence, such as structural (Lorrain and White 1971) or stochastic (Holland, Laskey, and Leinhardt 1983).

While the coupled-relations definition of role is focused on finding particular structural patterns between groups, we are interested in finding the generative behaviours exhibited in common by groups of individual users that may produce the coupled-structural relations underpinning the more formal notion of role. Our methodology consists of two stages. We firstly cluster users based on structural information obtained from their egocentric network, which we consider to be the key social structure with which to identify user roles (Fisher, Smith, and Welser 2006). We then interpret user clusters qualitatively before assigning them social roles.

The forum setting provides a novel challenge to the analyst wishing to use clustering methods. Typically, a majority of users provide only a minor contribution to a forum, while a relatively small collection of individuals exert major influence, and are thus of particular interest. Fur-

thermore, distinct roles occupied by users can have features in common (Fisher, Smith, and Welser 2006; Chan, Hayes, and Daly 2010) we account for this fact by developing a mixed-membership algorithm, while also exploring the multiple roles which users can inhabit within different fora.

Boards.ie Data

We investigate data gathered from boards.ie, a medium-sized bulletin board and the largest general topic discussion board in Ireland. We focus our analysis to 13,416 users participating in twenty forums over a six month period, from the 1st of July to the 31st of December, 2006.

Six structural features are investigated, five describing user interaction at the dyadic level and a sixth describing user's proactivity in attempting to generate discussion. These features have previously been effective at distinguishing user patterns (Chan, Hayes, and Daly 2010):

Indegree The number of distinct users who reply directly to a user comment.

Outdegree The number of distinct users whose commment a user directly replies to.

Weighted Indegree The total number of replies a user receives.

Weighted Outdegree The total number of times a user replies to user comments.

Reciprocity The total number of times user interaction on a thread is reciprocal, i.e., whenever a user both quotes and is quoted by another user on a thread.

Threads Initialised The number of threads initialised by a user.

Note that from a modelling perspective, user behaviour within each forum is considered to be entirely separate. While this may seem like a restriction on the model, it allows us to explore the different roles users perform in different fora, as will be discussed later.

Rudimentary analysis reveals the data to be highly correlated and strongly skewed. While visually no obvious clusters are apparent there is clearly wide-ranging and different behaviour within all forums. Initially, hierarchical clustering methods were employed to cluster users. While some clusters were obtained successfully, inspection of the model at various cutpoints revelealed several smaller clusters displaying overlapping features, a phenomenon outside the range of the standard clustering framework. The highly skewed nature of the data also meant that standard attempts to scale the data, such as principal component analysis, were somewhat unsatisfactory.

Model Specification

Due to the overlapping characteristics of the data, we model user behaviour within a framework which incorporates individual-level membership into the standard model-based clustering approach (Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Joutard 2007; Rogers et al. 2005). While the formulation of this model may be complex, its interpretation is somewhat simpler. Essentially, individual users are modelled as possessing multiple characteristics from different groups: this allows us to identify the social traits which different social roles share from directly within the models framework.

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ denote our dataset, consisting of N observations of user behaviour. Each $\mathbf{X}_n = (X_{n1}, \ldots, X_{nM})$ in turn consists of M recorded network statistics.

Assuming the data to be generated by a fixed number G of *extreme profiles*, the probability of profile membership for each user n in the dataset is defined by a G- dimensional parameter $\boldsymbol{\tau}_n$, drawn by a Dirichlet prior probability distribution with hyper-parameter $\boldsymbol{\delta}$. For each network statistic m, membership to profile g is drawn with probability τ_{ng} and the network statistic is then generated by a Poisson distribution with associated parameter θ_{gm}. Inference for this model is simplified by the introduction of the indicator variable $\mathbf{Z}$, where

$$Z_{nmg} = \begin{cases} 1 & \text{user } n \in \text{profile } g \text{ for statistic } m; \\ 0 & \text{otherwise.} \end{cases}$$

We wish to find the set of parameters which maximise our posterior, defined to be

$$p(\boldsymbol{\tau}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\delta}) \propto p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z}|\boldsymbol{\tau}) p(\boldsymbol{\tau}|\boldsymbol{\delta}) \qquad (1)$$

where

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) = \prod_{n,m,g=1}^{N,M,G} \left(\frac{\exp(-\theta_{gm}) \theta_{gm}^{X_{nm}}}{X_{nm}!} \right)^{Z_{nmg}},$$

$$p(\mathbf{Z}|\boldsymbol{\tau}) = \prod_{n,g=1}^{N,G} \tau_{ng}^{\sum_{m=1}^{M} Z_{nmg}}, \qquad p(\boldsymbol{\tau}|\boldsymbol{\delta}) \propto \prod_{n,g=1}^{N,G} \tau_{ng}^{\delta_g - 1}.$$

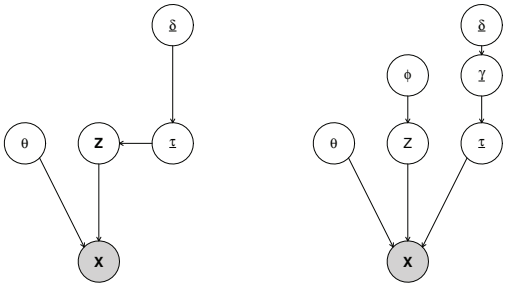

Figure 1: Graphical depiction of the mixed-membership formulation. The second figure depicts the variational Bayes approximation of the model.

Variational Bayes

We approximate (1) by employing a variational approach, namely approximating $p(\mathbf{Z}, \boldsymbol{\tau})$ with an independent distribution $q(\mathbf{Z}, \boldsymbol{\tau}|\boldsymbol{\phi}, \boldsymbol{\gamma})$ such that

$$q(\mathbf{Z}, \boldsymbol{\tau}|\boldsymbol{\phi}, \boldsymbol{\gamma}) = q(\boldsymbol{\tau}|\boldsymbol{\gamma}) \prod_{m=1}^{M} q(\mathbf{Z}_m|\boldsymbol{\phi}_m) \qquad (2)$$

where $\boldsymbol{\phi}, \boldsymbol{\gamma}$ are free variational parameters of the multinomial and Dirichlet distributions $q(\boldsymbol{\tau}|\boldsymbol{\gamma})$ and $q(\mathbf{Z}|\boldsymbol{\phi})$ respectively. Parameter updates are as follows:

$$\phi_{nmg} \quad \propto \quad \exp(-\theta_{gm})\theta_{gm}^{X_{nm}} \times \exp\left(\Psi(\gamma_{ng})\right)$$

$$\gamma_{ng} = (\delta_g - 1) + \sum_{m=1}^{M} \phi_{nmg}, \quad \hat{\theta}_{gm} = \frac{\sum_{n=1}^{N} \phi_{nmg} X_{nm}}{\sum_{n=1}^{N} \phi_{nmg}}$$

where Ψ denotes the digamma function (Abramowitz and Stegun 1965). Graphical model representations of equations (1) and (2) are shown in Figure 1.

While model assumptions require the number of profiles G to be fixed and known, in reality this is not the case. We therefore run the model over a range of values of G, and compare the models post-hoc. Note that the variational approximation provides only a lower bound to the model posterior, making the use of criteria such as the BIC difficult to interpret. We can, however, evaluate the hold-out likelihood of the model by integrating out $\boldsymbol{\tau}$ using Monte Carlo methods (Rogers et al. 2005).

Results

Mixed membership models were fitted to the data over between one and fifteen profiles, with the integrated likelihood providing clear evidence that an eight profile model optimally fits the data. Inspection of Table 1 indicates that the profiles are well seperated across network statistics, and that profile membership can be straightforwardly interpretable. While it is tempting to view each profile in the table as a separate behaviour cluster, it must be noted that a substantial number of users display membership across one or more profile type. This can be shown by calculating the extent of profile membership (EoM) of each user, where

$$\text{EoM}_n = \exp(-\sum_{g=1}^{G} \log \hat{\tau}_{ng}).$$

In this instance, about 30% of users have membership across more than one profile.

Parameter Estimates

We can impose a hard clustering by mapping users to their most probable profile memberships for each statistic. Here we introduce a convenient notation to denote a user's cluster membership: a six digit number, where each digit denotes profile membership with respect to the six network statistics described previously. For example, an individual characterized as belonging to cluster 222244 is characterized by extreme profile 2 for the first four descriptors and extreme profile 4 for the fifth and sixth descriptors. Table 2 provides a summary of the thirty most prominent such clusters. Typical to many message boards, many users display very low activity. This behaviour is summarised by the nine thousand strong cluster 111111, who could be described as engaging in no more than one or two conversations. The more active subgroups profiled display a richer set of behaviours, with several clusters differing in only one or two aspects, such as clusters 333332 and 333336. Users in these clusters have similar levels of activity but differ by reciprocity, indicating a difference in conversational engagement. While some clusters have extremely small memberships, this is a reflection of the more extreme behaviour taking place within boards.ie. For example, for the weighted in degree statistic, only twelve users map to profile 4, with these users clearly displaying prominent roles in their respective forums.

Role Identification

The problem of role interpretation to some extent remains. As an illustrative example, we consider the cluster 443332. This small cluster consists of 14 users, participating in in the Humanities, Development, Travel, Accommodation, Gigs/Events and Personal Issues forums. The cluster's behaviour can be described as low-volume, usually reciprocated interaction with a relatively high number of users. By inspecting user interaction in their respective forums during the time frame in question, it becomes clear that users in the cluster may be characterised as advice-givers, or problem solvers. Typically, the members of this cluster provide advice that helps to fix a problem, providing a resolution to the conversation and thus ending the thread.

Note that the roles users inhabit within the discussion board may vary depending on the forum in which discussion is taking place. As noted previously, users in different forums were treated as distinct within the model framework. As a result, users participating in multiple forums thus received multiple cluster assignations in our analysis. Comparing cluster assignments gives an indication of the different roles the user occupies within each forum.

Of our dataset, 21% of users on boards.ie converse with other users in more than one forum. Of these, roughly 1,000 users, about 6.8% of those in the study, have multiple "nontrivial" roles within different forums. By this we mean that these users were assigned to clusters other than cluster 111111, that is, the low-activity, casual user. A table showing the number of forums in which users have nontrivial participation is shown in Table 3.

Table 3: This table shows the number of forums to which unique users are assigned a profile cluster other than 111111.

No. of Fora	0	1	2	3	4	5	6	7	9
No. of Users	8337	4019	526	299	141	37	36	9	12

As an illustrative example, we consider one particular user. In the Development forum this user was assigned to the previously discussed cluster 443332. Within the Politics forum, he is assigned to cluster 777755, indicating that he is a highly active participant in the forum, engaging in detailed discussions with many users, and initiating several discussions himself.

Discussion

Message boards can possess a rich variety of behaviour, with several factors influencing discussion. We believe that the method outlined in this paper goes some way towards

Table 1: Parameter estimates for the network statistics of the eight profile model, ordered by in degree size.

Profile	In Degree	Out Degree	Weighted In Degree	Weighted Out Degree	Thread Initiation Rate	Reciprocity
1	1.62	1.55	1.74	1.65	0.30	1.38
2	8.19	7.99	10.04	9.85	0.59	73.10
3	22.54	23.20	34.12	35.39	0.78	461.30
4	38.26	39.39	915.84	902.55	3.68	18.47
5	50.15	52.95	84.25	89.51	6.54	863.94
6	68.33	70.83	163.11	171.52	2.34	203.25
7	105.34	111.05	291.98	302.46	26.11	1532.99
8	183.20	190.42	528.72	560.88	9.89	2579.67

Table 2: A table of the thirty largest profile combinations. While many boards users display low-level activity, the contrasting features of the smaller clusters provides insight into the behaviour of the more committed users.

Cluster	111111	222224	222221	222222	333332	333336	222244	121211	212111	333334
Size	9064	1109	585	315	257	182	105	83	81	77
	111114	212121	111141	333366	333333	555555	222121	212124	555553	555533
	76	67	65	51	49	49	38	28	27	26
	666663	443336	212221	221211	222242	212144	555566	332222	112211	666665
	25	23	22	20	20	19	18	17	16	16

identifying such factors in a quantitative manner. In particular, the use of the mixed membership famework has effectively modelled the overlapping characteristics of the user data. Future work may develop higher parameter versions of this model which incorporate, for example, higher order network statistics, or multiple viewpoints to facilitate analysis of a discussion board over time. Incorporation of the text users send to one another into the data set would mean that topic as well as role analysis could be jointly analysed (McCallum, Corrada-Emmanuel, and Wang 2005; Chang and Blei 2009).

We emphasise that the methods developed in this paper are intended to quantitatively facilitate what ultimately remains the qualitative task of role identification. While the analyst attempting to discover types of behaviour both common and unusual can utilise our approach in order to gain insight, they must be aware of several factors, most importantly the forum in which discussion is taking place, before a user's role can be identified. As ever, familiarity with a dataset is vital for analysis.

References

Abramowitz, M., and Stegun, I. A. 1965. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1 edition.

Agarwal, N.; Liu, H.; Tang, L.; and Yu, P. S. 2008. Identifying the influential bloggers in a community. WSDM '08, 207–218. New York, NY, USA: ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. In *ICWSM '10*, 215 –218.

Chang, J., and Blei, D. 2009. Relational topic models for document networks. In *AIStats '09*.

Erosheva, E. A.; Fienberg, S. E.; and Joutard, C. 2007. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics* 1(2):502–537.

Fisher, D.; Smith, M.; and Welser, H. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *HICSS '06*, volume 3, 59b.

Gleave, E.; Welser, H.; Lento, T.; and Smith, M. 2009. A conceptual and operational definition of 'social role' in online community. In *HICSS '09*, 1 –11.

Golder, S., and Donath, J. 2004. Social roles in electronic communities. Association of Internet Researchers.

Holland, P. W.; Laskey, K.; and Leinhardt, S. 1983. Stochastic blockmodels: First steps. *Social Networks* 5(2):109 – 137.

Lorrain, F., and White, H. 1971. Structural equivalence of individuals in social networks. *Journal of the Mathematical Sociology* 1(1):49–80.

McCallum, A.; Corrada-Emmanuel, A.; and Wang, X. 2005. Topic and role discovery in social networks. In *IJCAI*.

Rogers, S.; Girolami, M.; Campbell, C.; and Breitling, R. 2005. The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2:2005.

Wasserman, S., and Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge Univ Press.

Welser, H. T.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; and Smith, M. 2011. Finding social roles in Wikipedia. In *iConference '11*, 122–129. ACM.