Community Topic Usage in Social Networks

Ian D. Wood

Research School of Computer Science
Australian National University
Canberra, Australia

Insight Centre for Data Analytics
National University of Ireland, Galway
Galway, Ireland

ian.wood@anu.edu.au

ABSTRACT

When studying large social media data sets, it is useful to reduce the dimensionality of both the network (e.g. by finding communities) and user-generated data such as text (e.g. using topic models). Algorithms exist for both these tasks, however their combination has received little attention and proposed models to date are not scalable (e.g.: [4]). One approach to such combined modelling is to perform community and topic modelling independently and later combine the results. In the case of overlapping communities, this combination requires a method for attributing each users topic usage to the communities in which she participates. This paper presents a Bayesian model for attributing individual documents to communities which balances the users proportional community membership with community topic coherence. Community topic usage is modelled with a Dirichlet distribution with fixed concentration parameter, leading to a well defined conjugate prior. Thought the prior is computationally expensive, the already reduced dimensionality in both topics and communities make a tractable algorithm feasible, even for large data sets. The model is applied to a corpus of tweets and twitter follower relations collected on hash tags used by people with eating disorders [14].

Keywords

topic models; community detection; Bayesian inference; conjugate prior; Dirichlet distribution; author community membership

1. INTRODUCTION

Several studies have found that communities in the twitter follower network can act as a kind of forum on particular topics of discussion [9, 8, 7]. In this scenario, tweets intended for such a forum would reflect those topics, whereas tweets by the same users that are intended for other audiences would show distinct topical content. It is the aim of the work presented here to distinguish the intended audience (in terms of follower network communities) of each tweet and in this way estimate the topics used by those communities. One would expect that many twitter users would be members of/contribute to multiple communities, thus one would expect such communities to be overlapping [9].

Approaches to linking social media texts with network communities have been studied previously. Java et.al. [9] performed overlapping community detection on the full Twitter network and identified coherent themes in key terms used by some inferred communities, though they were not clear on how the key terms were identified and did not provide numerical measures of such coherence. There were about 94,000 twitter users in April 2007 when they performed their study, thus scale was less of an issue than today (in 2015 there are over 300 million Twitter users).

Duan et.al. [4] developed a full Bayesian model incorporating both a stochastic block model for community detection and hierarchical Dirichlet process for topic detection. In this model, all of an authors documents are assigned to just one community (hence they do not overlap) and it's scalability is questionable.

Li et.al. [10] present a different approach to combined community and topic detection by utilising extra thematic metadata — hash tags (twitter data) and publication venue (citation data). The twitter follower network was not utilised. In their model, communities (not documents) have topic mixtures and topics generate both words and hash tags/venues. The twitter data analysed was intended as a summary of hot topics over a 2 month period, in contrast to the data utilised here that intends to capture interactions within a restricted set of twitter communities over a longer period. In such a social data set, follower links are of great importance, as they represent the conduit over which interactions are possible.

Earlier, Li et.al. [11] applied a similar approach to what we propose here, combining the results of community detection and topic modelling and applying the resulting synthesis to social bookmarking data. The community model they ap-

TM'15, October 19, 2015, Melbourne, Australia.

ACM 978-1-4503-3784-7/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2809936.2809937.

plied, however, did not produce overlapping communities so a naive approach to inferring community topic proportions was effective.

In Section 2 I describe and develop the model, including the conjugate prior to the Dirichlet distribution. In Section 3 we present an algorithm based on Gibbs sampling for estimating the posterior. In Section 4 we develop two metrics for assessing model quality. In Section 5 we describe the data set and contributing topic and community detection models used as an example in this study. In Section 6 we present results showing that the model succeeds in it's aims. In Section 7 we summarise the contribution and discuss future work and implications.

2. DOCUMENT ASSIGNMENT MODEL

Assigning documents to their authors communities is done according to two premises: the proportion of an authors documents in a community should reflect the authors proportional community membership and the topic proportions of documents assigned to a community should be somewhat coherent. This is operationalised by the following generative model. A authors, C communities and N documents are modelled.

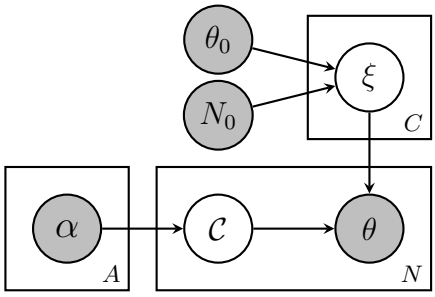

Figure 1: Generative Model

Document community assignments $\mathcal{C}$ are generated by a fixed multinomial whose probabilities are the document authors community membership proportions α. For each document d assigned to community c, topic proportions θ_d are drawn from a Dirichlet distribution (with parameters ξ_c) for that community. A conjugate prior for the ξ is provided, parametrised by N_0 and θ_0 (see Section 2.1 for the construction of the prior). The model is summarised in Figure 1. Grey nodes indicate observed or pre-set values.

The probability of assignment of document d with author a_d to community c, and the probability of a documents topic distribution θ_d are as follows:

$$P(d \in c) = \alpha_{a_d c} \tag{1}$$

$$P(\theta_d | d \in c, \xi_c) = B(\xi_c)^{-1} \prod_t (\theta_{dt})^{\xi_{ct}-1} \tag{2}$$

2.1 A Conjugate Prior For Dirichlet Distributions

The Dirichlet distribution is a member of the exponential family of distributions, and as such has a (conjugate) prior with a relatively simple, constant-dimensional Bayesian update. Given the equation for the T dimensional Dirichlet distribution with parameters ζ

$$P(\theta) = B(\zeta)^{-1} \prod_t \theta_t^{\zeta_t - 1} \tag{3}$$

$$B(\zeta) = \frac{\prod_t \Gamma(\zeta_t)}{\Gamma(\sum_t \zeta_t)} \tag{4}$$

where B is the beta function, it is easy to write down a candidate conjugate prior and corresponding posterior update after evidence $\{\theta_1 \dots \theta_N\}$:

$$P_\pi(\zeta) \propto B(\zeta)^{-N_0} \prod_t (\theta_{0t})^{\zeta_t - 1} \tag{5}$$

$$P(\zeta | \theta_1 \dots \theta_N) \propto B(\zeta)^{-(N_0+N)} \prod_t \left(\theta_{0t} \prod_n \theta_{nt} \right)^{\zeta_t - 1} \tag{6}$$

here, n ranges from 1 to N and t from 1 to T. The values for N_0 and θ_0 can be interpreted in terms of hypothetical prior observations: N_0 being the number of prior observations and θ_0 the element-wise product of those observations.

Note that due to the $\Gamma(\sum_t \zeta_t)$ term in $B(\zeta)$, this only defines a probability if $\sum_t \zeta_t$ is bounded. We could however multiply this candidate by an arbitrary function of ζ and it would remain a conjugate prior (ie: have convenient posterior form and update). For example we could choose to multiply by $\Gamma(\sum_t \zeta_t)^{-\sum_t \zeta_t}$ and the resulting function would have bounded integral (and could thus define a probability). For the purposes of this study, however, we chose instead to fix $\sum_t \zeta_t$.

For convenience we will express $\zeta = \Xi\xi$ with fixed concentration parameter $\Xi = \sum_t \zeta_t > 0$, a scalar, and $\sum_t \xi_t = 1$, $\xi_t \geq 0$. We can now write down the full probability of the model. Taking $\mathcal{C}_d$ to represent the allocated community for document d and N_c the number of documents allocated to community c, we have:

$$\begin{aligned}
&P(\mathcal{C}, \theta, \xi | \alpha, \theta_0, N_0) \\
&\quad = \prod_d P(d \in \mathcal{C}_d) P(\theta_d | d \in \mathcal{C}_d, \xi_{\mathcal{C}_d}) \prod_c P(\xi_c | N_0, \theta_0) \\
&\quad \propto \prod_d \alpha_{d\mathcal{C}_d} \left(B(\Xi\xi_{\mathcal{C}_d})^{-1} \prod_t (\theta_{dt})^{\Xi\xi_{\mathcal{C}_d t}-1} \right) \\
&\qquad\qquad \times \left(\prod_c B(\Xi\xi_c)^{-N_0} \prod_t (\theta_{0t})^{\Xi\xi_{ct}-1} \right) \\
&\quad = \prod_c B(\Xi\xi_c)^{-(N_0+N_c)} \left(\prod_{d \in c} \alpha_{dc} \right) \prod_t \left(\theta_{0t} \prod_{d \in c} \theta_{dt} \right)^{\Xi\xi_{ct}-1}
\end{aligned} \tag{7}$$

3. ESTIMATION

To obtain a maximum a posteriori (MAP) estimate for document-community associations and community topic distributions, we use a modified Gibbs sampling algorithm not dissimilar to that used in [6]. The method iterates between sampling from the posterior distribution of document-community associations and MAP estimation of ξ with those associations fixed.

To sample document community allocations, we need the conditional probability of a documents community membership given the current value of ξ. Omitting inconsequent conditional dependencies and terms independent of d and c, we obtain:

$$P(d \in c|\xi,\theta) \propto P(d \in c|\xi)P(\theta_d|\xi) \propto \alpha_{dc}B(\Xi\xi_c)^{-1}\prod_t(\theta_{dt})^{\Xi\xi_{ct}-1} \quad (8)$$

For the MAP estimation of ξ we need it's conditional probability given current document allocations. Again omitting inconsequent dependencies and terms independent of ξ_c and c, and writing θ_c for the set of topic proportions for documents in c, we obtain:

$$\begin{aligned} P(\xi_c|\mathcal{C},\theta) &\propto P(\theta_c|\mathcal{C},\xi_c)P(\xi_c) \\ &\propto \left(\prod_{d\in c} B(\Xi\xi_c)^{-1}\prod_t(\theta_{dt})^{\Xi\xi_{ct}-1}\right) \\ &\quad \times \left(B(\Xi\xi_c)^{-N_0}\prod_t(\theta_{0t})^{\Xi\xi_{ct}-1}\right) \\ &= B(\Xi\xi)^{-(N_0+N_c)}\prod_t\left(\theta_{0t}\prod_{d\in c}\theta_{dt}\right)^{\Xi\xi_{ct}-1} \end{aligned} \quad (9)$$

Estimates for ξ_c were obtained from Equation (9) via numerical optimisation. With fixed Ξ and due to the logarithmic convexity of the Gamma function for positive real numbers, this expression can be seen to be logarithmically concave, thus numerical optimisation of it's log can be expected to behave reasonably, as was found to be the case.

Neither Equation (8) nor (9) scale well, however due to the already reduced dimensionality of the input data through topic modelling and community detection algorithms, it has proved tractable on large data sets.

4. METRICS OF MODEL QUALITY

In this unsupervised setting, comparison to a ground truth is impossible. The numerical metrics below attempt therefore to assess the efficacy of the model in terms of the models goals. The metrics were applied both to estimated models and to naive document allocation via community membership proportions (α) alone. Results are presented in Table 1.

Community Topic Coherence.

To capture how effective the models had been at resolving coherent community topic proportions, the conditional entropy of community allocations $\mathcal{C}$ given community topic proportions $\mathcal{T}$ was employed.

$$H(\mathcal{C}|\mathcal{T}) = -\sum_c P(c)\sum_t P(t|c)\log_2 P(t|c) \quad (10)$$

This quantity captures the amount of extra information (measured in binary bits) needed to obtain the community document allocations given knowledge of community topic proportions. If the documents associated with a community are faithful to the topic proportions of that community, you would expect this to be low. On the other hand, if they have a diversity in their topic mixes, much extra information would be needed to identify them.

Taking N_a to be the number of documents from author a and recalling N represents the total number of documents, α_{ac} the affinity of author a for community c, and θ_{dt} the topic proportions of document d, probabilities for naive allocation can be written as follows:

$$\begin{aligned} P_{\text{naive}}(c) &= \sum_a P(a)P(c|a) \\ &= \sum_a \frac{N_a}{N}\alpha_{ac} \\ &= \frac{1}{N}\sum_a N_a\alpha_{ac} \\ P_{\text{naive}}(t|c) &= \frac{P(t,c)}{P(c)} \\ &= \frac{\sum_a P(a)P(t|a)P(c|a)}{\sum_a P(a)P(c|a)} \\ &= \frac{\sum_a \frac{N_a}{N}(\frac{1}{N_a}\sum_{d\in a}\theta_{dt})\alpha_{ac}}{\sum_a \frac{N_a}{N}\alpha_{ac}} \\ &= \frac{\sum_a \alpha_{ac}\sum_{d\in a}\theta_{dt}}{\sum_a \alpha_{ac}N_a} \end{aligned} \quad (11)$$

For the estimated models, we use MAP estimates of document allocations for community probability and expected values of the posterior community Dirichlet distributions, which are just their parameters ξ_c, for conditional topic probabilities.

$$\begin{aligned} P_{\text{estimated}}(c) &= \frac{N_c}{N} \\ P_{\text{estimated}}(t|c) &= \xi_{ct} \end{aligned} \quad (12)$$

To assess individual communities, we can also calculate the entropy $H(c|\mathcal{T})$ for some community c:

$$H(c|\mathcal{T}) = -\Big[P(c)\sum_t P(t|c)\log_2 P(t|c) + P(\neg c)\sum_t P(t|\neg c)\log_2 P(t|\neg c)\Big] \quad (13)$$

Again it is useful to compare entropies from a naive model and an estimated model. We already have formulae for $P(c)$ and $P(t|c)$ (Equations 8 and 9). For a naive model, we have:

$$\begin{aligned} P_{\text{naive}}(\neg c) &= \frac{1}{N}\sum_a N_a(1-\alpha_{ac}) \\ P_{\text{naive}}(t|\neg c) &= \frac{\sum_a(1-\alpha_{ac})\sum_{d\in a}\theta_{dt}}{\sum_a(1-\alpha_{ac})N_a} \end{aligned} \quad (14)$$

and for the estimated models, we have:

$$\begin{aligned} P_{\text{estimated}}(\neg c) &= \frac{N-N_c}{N} \\ P_{\text{estimated}}(t|\neg c) &= 1-\xi_c \end{aligned} \quad (15)$$

Faithfulness to Author Community Membership.

There can be a tension between respecting author community affinities and creating coherent community topic distributions. A model that produces excellent community topic distributions may require documents to be allocated in different proportions to their authors community affinities.

To assess this disparity, we use the Hellinger distance between estimated author community affinities calculated from document assignments and the actual affinities used as inputs to the model. Kullback-Leibler divergence was also considered, however this leads to uninformative infinite divergences if the estimate for a community is zero and the actual affinity non-zero.

$$\begin{aligned} &H(P_\alpha(c), P_{\text{estimated}}(c)) \\ &\quad= \frac{1}{\sqrt{2}}\sqrt{\sum_c \left(\sqrt{P_\alpha(c)} - \sqrt{P_{\text{estimated}}(c)}\right)^2} \\ &\quad= \frac{1}{\sqrt{2}}\sqrt{\sum_c \left(\sqrt{\sum_a \frac{\alpha_{ca} N_a}{N}} - \sqrt{\frac{N_c}{N}}\right)^2} \end{aligned} \tag{16}$$

Community membership of authors in the Twitter follower network is an indication of who they listen to. The model presented here makes the assumption that documents are divided between those communities in similar proportions to the number of links to those communities, but this may not be the case. The links represent the mix of sources of tweets that a user sees, whereas the documents assigned to a community represent tweets intended to be seen by that community. Proportions of active and passive communication may not always coincide. For example, other users followed for interest as sources of information are unlikely to be considered as targets for published tweets.

As such, we may not necessarily expect complete symmetry between listening (represented here by follower links and α) and speaking (represented by tweets and their allocation to communities), and low similarity may be acceptable. Note that many community affinities are zero, meaning no links exist to members of that community and no communication is possible. In these cases, the affinity is always respected (see Equation 8).

5. DATA SET

Hash tags have been identified as potential symbols of community membership [12, 3]. Drawing on this observation, tweets were collected on a selection of Twitter tags such as *#proana*, *#edproblems* and *#thinspiration* found to be used by the Twitter "pro-anorexia" and eating disorder community between December 2012 and December 2014[14]. During data collection, the lists of friends and followers of the author of each tweet were also collected.

Text Data and Topic Model.

Retweets were removed and Tweets were tokenised by standardising numerous text emoticon forms, isolating punctuation as individual word tokens and converting mixed case words to lower case (all caps words were retained). Url's, #tags, @mentions and apostrophised words (eg: "didn't") were left unchanged. Further pre-processing included removal of word tokens appearing less than 5 times and removal of tweets with less than 3 word tokens. This resulted in a corpus of 262,736 documents and a vocabulary of 18,713 words. A standard latent Dirichlet allocation [2] topic model with 20 topics was inferred for the resulting corpus. The LDA Dirichlet prior on topic/word probabilities was set to $\beta = 0.01$. Writing N for the number of words, D the number of documents in corpus, T the number of topics, the parameter for the LDA Dirichlet prior on document/topic probabilities was set $\alpha = 0.05N/DT$. This value allocates 5% of the probability mass for smoothing. A previous study found evidence that a topic model such as this can have some ability to resolve social psychological constructs such as identity salience [15].

Network Data and Community Model.

We consider only mutual follower links as they indicate some possibility of mutual interaction, a feature we expect of social communities. Initial analysis of the collected network data indicates that many follower links had not been polled since near the beginning of data collection, and in fact the distribution of "last polled times" is roughly linear in time. A rudimentary survival analysis (the median link age for links where both creation and removal events were observed[1]) indicated links on average lasted approximately 96 days. For the community analysis, links that had not been polled within 96 days of the last observation (December 2014) were discarded. Degree one nodes were removed as they are highly likely to be connected to other, unobserved, nodes and communities in the larger Twitter network. The resulting network has 66,744 nodes and 927,594 edges. Overlapping communities were inferred using a mixed membership stochastic block model [5, 1]. Visualisation of the network with community observations (Figure 2) reveals a relatively small number of fairly distinct, well separated communities with the remainder highly interconnected. For this reason we chose to make three models, one with the inferred 183 communities, and two with a smaller number of communities (50 and 20).

Combination.

When combining the two forms of data for this analysis, only users who appeared in both were retained. That is, users who had at least one tweet retained for the topic model as well as a (recently observed) link retained in the network data. This resulted in 15,515 users and 133,851 of their tweets.

6. RESULTS

Models were inferred for several values of Ξ and compared to naive document allocation via community membership proportions α alone using the metrics presented in Section 4. Initial experiments suggested a value of $\Xi \simeq 600$ would perform well and models were also estimated with $\Xi = 100$ and $\Xi = 30$ for comparison. Results are summarised in Table 1.

Overall Assessment.

As expected, the larger value of Ξ produced more resolved community topic proportions (lower entropy scores) and were less faithful to the community membership information inferred from the network data (Table 1). The increased distance to community membership information was however small compared to the improved resolution of topics for communities, thus higher values of Ξ should be preferred.

The implementation used for experiments presented here assigned batches of 100 documents between estimations of ξ_c for communities whose membership had changed. Estimations of ξ were done with scipy.optimize.minimize using the "Nelder-Mead" method [13], a hill climbing simplex method.

[1]weighted by the opportunity to observe links of that duration given the observation window

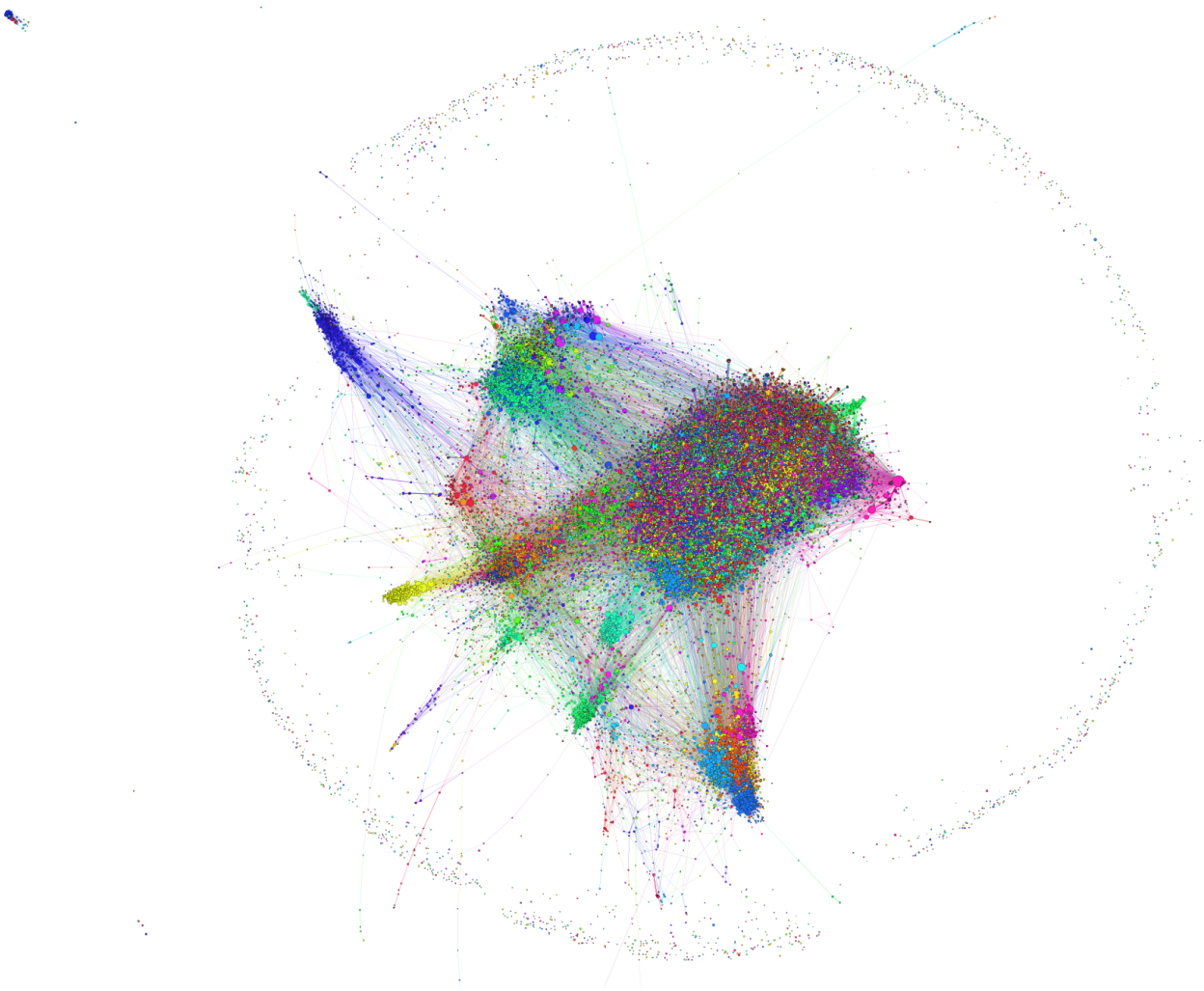

Figure 2: Network Visualisation with 183 Communities. Colours represent different communities, node size indicates bridgedness. Note few distinct, separated communities and many highly interconnected communities.

183 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.78	2.81	2.27	1.99
Hellinger	0	0.103	0.154	0.176

50 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.84	3.23	2.84	2.62
Hellinger	0	0.161	0.158	0.162

20 Communities

	Naive	$\Xi = 30$	$\Xi = 100$	$\Xi = 600$
Entropy	3.94	3.76	3.53	3.38
Hellinger	0	0.108	0.116	0.146

Table 1: Conditional Entropy $H(\mathcal{C}|\mathcal{T})$ (Equation 10) and Hellinger Distance $H(P(c|\alpha), P(c|\mathcal{C}_{\text{estimated}}))$ (Equation 16)

This method does not allow caching and control of the initial simplex, thus small perturbations of community membership require a similar number of function evaluations (in the order of 800) to uninformed starting points.

Substantial improvements in run times could be achieved with control of the starting simplex and simple heuristics for required precision at different stages during estimation. Execution times were not insubstantial (approximately 3 hours to converge for all configurations on a 16 core 2.3Ghz machine) but only around 30 full iterations (over all documents). We could expect at least an order of magnitude improvement with more intelligent and integrated numerical optimisation of ξ. It should also be noted that bounds checking of ξ was performed within the optimised function to prevent evaluation outside the simplex. Such checks would be better placed within the optimisation routine itself.

Interpretation and Discussion.

On inspection of the community topic allocations, it was found that approximately half the communities had more than half the probability mass concentrated on just one topic in all models. The individual community entropy scores (Equation 13) give a good indication of the level of concentration, the more concentrated having notably lower entropy scores. Figure 3 shows community allocations for 50 communities and $\Xi = 100$. Similar patterns were found for other models.

To better understand inferred relationships between topics and communities, a simple analysis of community relationships was performed. Working again with the 50 community model, community correlations (over documents, given document/community probabilities α from the network community model) were calculated followed by Principal Component Analysis (PCA) on the correlation matrix. The correlation data is intrinsically highly dimensional, with the maximum proportion of variance for a component being just 7.1%, more than half the components accounting for $> 1.9\%$ and all but one accounting for $> 1\%$. Hierarchical clustering using correlations as a similarity metric supported this observation, revealing just two clusters with internal correlations ≥ 0.2 (communities 19, 32, 46 and 14, 44) with the greatest correlation in the data at .385 between communities 32 and 46.

The clear associations between topics and communities testifies to the efficacy of the community detection and topic modelling algorithms and supports the hypothesis that communities in the mutual follower network often define fora for discussion or sharing along a particular theme. It also begs the question as to possible biases in the model — is it too good to be true? Further investigation of the model, perhaps using sythetic data sets with known properties may be appropriate to allay such suspicions.

7. CONCLUSIONS

This paper presents an approach to identifying Twitter communities and their topics of discussion. Existing efficient methods for community detection and topic modelling are leveraged and a novel Bayesian model and inference algorithm are developed to associate tweets with communities in which their authors participate.

A Dirichlet distribution is used to model community topic usage, and a conjugate prior for the Dirichlet distribution is developed. A modified Gibbs sampling procedure incorporating alternate sampling of document/community allocations and MAP estimation of community topic Dirichlet distributions is used to estimate the posterior. The MAP estimation step requires costly numerical optimisation, however due to the already reduced dimensionality of the problem (from text topic modelling and network community detection), this remains tractable for reasonably large data sets.

The model is applied to a collection of 262,736 tweets and 441,655 user follow relations collected from public tweets related to "pro-anorexia" and eating disorders. A substantial improvement of community topic coherence is demonstrated relative to a naive approach that utilises author community membership alone. Results show very distinct community topic usage for more than half the communities. This is a strong result, supporting the hypothesis that communities in the mutual follower network often serve as fora for particular themes, however it is also suggestive of possible inherent bias in the model which should be further investigated.

The principles used and design of the algorithms presented here are a step to understanding the relations between community structures and topic usage with the future aim of developing a joint model of community detection in author networks and topic modelling of document content.

Possible extensions of the model include estimation of the scaling parameter for community topic proportions and introducing a parameter to moderate the relative strength of author community membership and community topic coherence during inference.

8. ACKNOWLEDGMENTS

This work was funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT) and the European Union supported projects LIDER (ICT-2013.4.1-610782) and MixedEmotions (H2020-644632). Data collection and analysis were performed with the aid of the Australian National eResearch Collaboration Tools and Resources (NeCTAR) Research Cloud.

9. REFERENCES

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou,

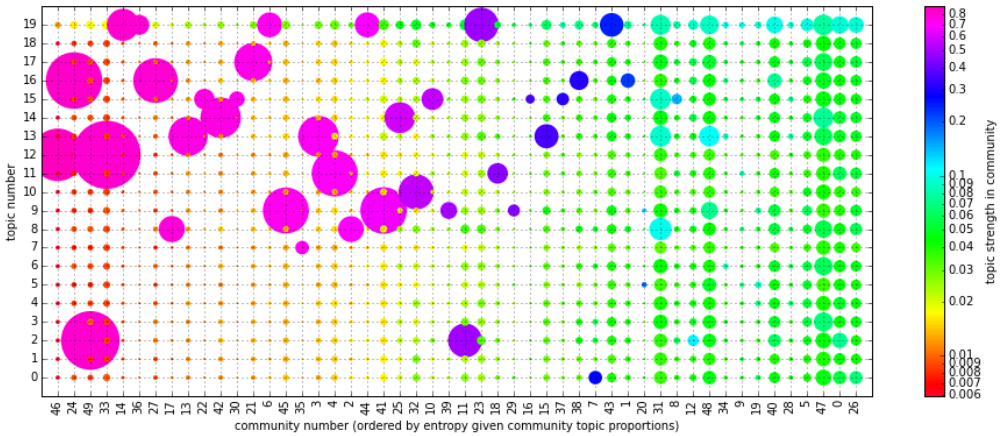

Figure 3: Community topic allocations with 50 communities, $\Xi = 100$. Data point area proportional to expected sum of document topic proportions for a community. Communities ordered by Equation 13.

editors, *Advances in Neural Information Processing Systems 21*, pages 33–40. Curran Associates, Inc., 2009.

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

[3] A. Bruns and J. E. Burgess. The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, University of Iceland, Reykjavik, 2011.

[4] D. Duan, Y. Li, R. Li, Z. Lu, and A. Wen. Mei: mutual enhanced infinite generative model for simultaneous community and topic detection. In *Discovery Science*, pages 91–106. Springer, 2011.

[5] P. K. Gopalan and D. M. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[7] I. Himelboim. Mapping twitter topic networks: From polarized crowds to community clusters. Blog post, Pew Research Center's Internet & American Life Project, 2014.

[8] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. SSRN Scholarly Paper ID 1313405, Social Science Research Network, Rochester, NY, 2008.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[10] D. Li, Y. Ding, X. Shuai, J. Bollen, J. Tang, S. Chen, J. Zhu, and G. Rocha. Adding community and dynamic to topic models. *Journal of Informetrics*, 6(2):237–253, 2012.

[11] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1565–1568, New York, NY, USA, 2010. ACM.

[12] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 2011.

[13] S. Singer and J. Nelder. Nelder-mead algorithm. *Scholarpedia*, 4(7):2928, 2009.

[14] I. Wood. A case study of collecting dynamic social data: The pro-ana twitter community. *Australian Journal of Intelligent Information Processing Systems*, 14(3), 2015.

[15] I. Wood. Using topic models to measure social psychological characteristics in online social media. In *Social Computing, Behavioral-Cultural Modeling, and Prediction*, volume 9021 of *Lecture Notes in Computer Science*, pages 308–313, Washington, DC, USA, 2015. Springer International Publishing.