Transparent User Models for Personalization

Khalid El-Arini*
Carnegie Mellon University
Pittsburgh, PA

Ulrich Paquet
Microsoft Research
Cambridge, UK

Ralf Herbrich*
Facebook, Inc.
Menlo Park, CA

Jurgen Van Gael*
Rangespan Ltd.
London, UK

Blaise Agüera y Arcas
Microsoft Corp.
Bellevue, WA

ABSTRACT

Personalization is a ubiquitous phenomenon in our daily online experience. While such technology is critical for helping us combat the overload of information we face, in many cases, we may not even realize that our results are being tailored to our personal tastes and preferences. Worse yet, when such a system makes a mistake, we have little recourse to correct it.

In this work, we propose a framework for addressing this problem by developing a new user-interpretable feature set upon which to base personalized recommendations. These features, which we call *badges*, represent fundamental traits of users (e.g., “vegetarian” or “Apple fanboy”) inferred by modeling the interplay between a user’s behavior and self-reported identity. Specifically, we consider the microblogging site Twitter, where users provide short descriptions of themselves in their profiles, as well as perform actions such as tweeting and retweeting. Our approach is based on the insight that we can define badges using high precision, low recall rules (e.g., “Twitter profile contains the phrase ‘Apple fanboy”’), and with enough data, generalize to other users by observing shared behavior. We develop a fully Bayesian, generative model that describes this interaction, while allowing us to avoid the pitfalls associated with having positive-only data.

Experiments on real Twitter data demonstrate the effectiveness of our model at capturing rich and interpretable user traits that can be used to provide transparency for personalization.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics

Keywords

personalization, Twitter, graphical models, transparency

1. INTRODUCTION

Whether we are reading news, searching the Web or connecting with our friends on a social network, personalization plays an important—if often hidden—role in shaping our experience. As the scale of online content grows, the ability to tailor information to the tastes and preferences of individual users is becoming critical for maintaining a positive user experience. For example, different users may prefer news articles from different blogs, diners with different tastes may trust different restaurant reviews, and so on.

However, along with the promise of personalization come many challenges, both technical and social, that hinder its potential:

- Users often do not know that their results are being personalized in the first place, and as such, may not understand why their Web experience is different from (and perhaps worse than) that of their friends.
- Even if they are aware that their results are personalized, users are rarely provided with information about how the particular site or service perceives them, and, as such, have little recourse to make corrections, if necessary. For example, in a famous critique of personalization as applied to television show recommendation in TiVo[1], Zaslow describes the drastic steps users feel they need to take in order to correct misperceptions that the system has of them, reaching the conclusion that “there’s just one way to change its ‘mind’: outfox it,” [16].
- Even when a system correctly models a user’s interests and tastes, it may not always be desirable to use such information. Whether because of privacy concerns or to avoid groupthink (cf. Pariser’s *The Filter Bubble* [10]), users may wish to selectively inhibit certain signals or attributes from being used for personalization.

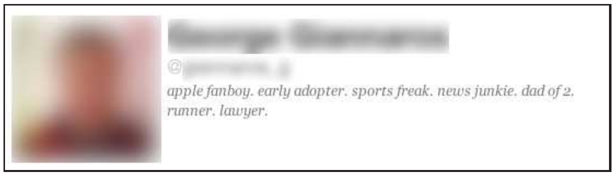
In this work, we seek to address these challenges by making personalization more *transparent*. In other words, users should know, (1) *when* personalization is happening, and, (2) *how* they are perceived by the system (with the ability to correct this perception as necessary).

We provide this transparency by representing each user as a set of interpretable, explainable attributes (e.g., “veg-

*Work done while at Microsoft Research, Cambridge, UK.

[1] `http://www.tivo.com`

KDD’12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$10.00.



[redacted]

apple fanboy. early adopter. sports freak. news junkie. dad of 2. runner. lawyer.

Figure 1: An example Twitter profile, showing an anonymized user's self-reported description. Here, if we are interested in the "Apple fanboy" badge, we can observe this user's actions on Twitter to help us figure out what it means to be an Apple fanboy.

etarian," "hipster," or "Apple fanboy") that we learn from user behavior. It is important that both the *meaning* of these attributes and *why they were assigned* to the user be readily apparent. Following the paradigm made popular by location-aware social networks such as Foursquare[2], we refer to these attributes as *badges*.

In particular, we consider the microblogging site Twitter[3], and we associate each badge with a characteristic *label* (e.g., "Apple fanboy") that a user might use to describe himself in his Twitter profile. For any user, we can observe his or her profile and determine whether or not it contains a particular label. For example, the user profile in Figure 1 contains the label "Apple fanboy," which we might associate with an Apple fanboy badge. It is important to note that there is a *probabilistic* relationship between labels and the badges they correspond to. For example, most users who adore Apple products will not explicitly identify themselves with "Apple fanboy" in their Twitter profiles[4]. Nevertheless, we wish to use the actions of those *self-identified* Apple fanboys to help us learn what it means to be one. We can then hope to predict which other users might also be Apple fanboys, even if they don't identify themselves as such.

Moreover, in this paper, we take the view that the set of possible badges (and their corresponding labels) are defined *a priori* in a supervised manner. Specifically, we assume we are *given* some set of possible badges (e.g., as in Table 1), and wish to *infer*: (1) their presence or absence for each user, and, (2) how they manifest themselves in terms of Twitter behavior.

In the remainder of this paper, we describe how we learn badges from user activity on Twitter, using a Bayesian framework to explicitly model uncertainty. We show experimental results on real Twitter users, and present both quantitative evidence and qualitative anecdotes demonstrating the effectiveness of our method.

2. MODELING BADGES

We describe each user as a set of *latent* badges that, collectively, explain the user's behavior. The fundamental problem we seek to solve is: *how do we infer the badges for each user based on observed actions and labels?*

For each user u, we observe two binary vectors:

1. The label vector $\boldsymbol{\lambda}^{(u)}$, with $\lambda_i^{(u)} = 1$ indicating that the Twitter profile of user u contains the label corresponding to badge i. As an example, if we let badge i correspond to the "runner" label, then $\lambda_i^{(u)} = 1$ if the Twitter profile of user u contains the word "runner," and $\lambda_i^{(u)} = 0$ otherwise.
2. The action vector $\boldsymbol{a}^{(u)}$, where $a_j^{(u)} = 1$ if user u is observed performing action j. In our Twitter domain, we take the set of possible actions to include all hashtags and retweets. For example, action j might correspond to tweeting with the hashtag `#runkeeper`, and $a_j^{(u)} = 1$ for users u that have such a tweet.

We model these observations as probabilistically arising from a latent set of badges $\boldsymbol{b}^{(u)}$, where $b_i^{(u)} \in \{0, 1\}$ indicates whether or not user u has badge i. In particular, we elect to define a generative—rather than discriminative—model; while the high precision labels may provide us with positive training examples, their low recall leads to no meaningful negative examples. Moreover, if a user chooses to decline a badge that we predict for him (e.g., he might not really be an Apple fanboy), this simply corresponds in our model to observing the latent variable $b_i^{(u)} = 0$. Additionally, we note that our model differs from traditional unsupervised latent variable models, such as topic models [2], in that the badge labels provide identifiability that we would not otherwise achieve. Thus, for example, if we define the label for badge i to correspond to those users with "runner" in their Twitter profile, then the actions explained by badge i will always correspond to (our view of) runners, which is a property we do not get with fully unsupervised topic models, such as latent Dirichlet allocation [3].

2.1 Generating labels

Given a particular user's badge assignments, the generative process for labels encodes our intuition that each label $\boldsymbol{\lambda}_i$ is a high precision, low recall indicator of the presence or absence of a badge. Specifically, "high precision" here means that it is very unlikely for someone without badge i (i.e., with $b_i^{(u)} = 0$) to use the corresponding label (i.e., $\lambda_i^{(u)} = 1$) in his profile, while "low recall" indicates that many users u with $b_i^{(u)} = 1$ nevertheless have $\lambda_i^{(u)} = 0$. For example, while most vegetarians on Twitter do not describe themselves as "vegetarian" in their Twitter profiles, it is much more rare (but not impossible) for non-vegetarians to have the word "vegetarian" in their profiles.

As such, we model label $\lambda_i^{(u)}$ as being *a priori* present with a true positive rate γ_i^{T} and false positive rate γ_i^{F} (with $\gamma_i^{\mathrm{F}} \ll \gamma_i^{\mathrm{T}} < 1$ and $\gamma_i^{\mathrm{F}} \approx 0$). Formally, we have:

$$p(\lambda_i^{(u)} = 1 \,|\, b_i^{(u)}, \gamma_i^{\mathrm{T}}, \gamma_i^{\mathrm{F}}) = \mathsf{Bernoulli}(b_i^{(u)}\gamma_i^{\mathrm{T}} + (1 - b_i^{(u)})\gamma_i^{\mathrm{F}}),$$

given the user's badge $b_i^{(u)}$. In other words, the presence of a badge does not necessarily imply its appearance in a user's profile, and it is precisely these badges that we aim to infer.

2.2 Generating actions

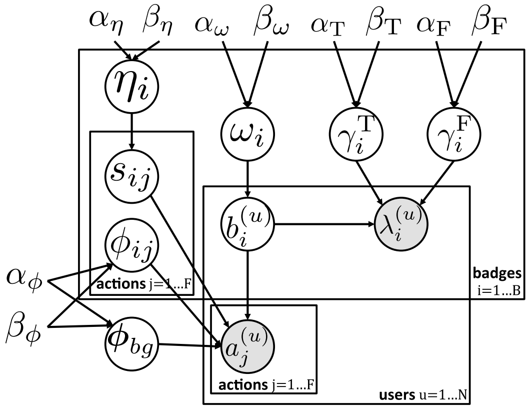
We assume that the observed actions $a_j^{(u)} \in \{0, 1\}$ of a user u can be explained by one or more of his latent badges $b_i^{(u)}$. In the Twitter domain, possible actions j might include a user *re-tweeting* some author, or using a particular *hashtag*.

For each possible badge i and action j, there is a probability $s_{ij}\phi_{ij}$ of associating them; it is decomposed into a context-specific rate $\phi_{ij} \in (0, 1)$ and a sparsity prior $s_{ij} \in \{0, 1\}$. The $\mathbf{s}_i$ variables for a badge i act as a mask, delineat-

[2] `http://www.foursquare.com`

[3] `http://www.twitter.com`

[4] This stands in contrast to Foursquare, where badges are deterministic (e.g., five check-ins at an airport *guarantee* the "jet-setter" badge).



```
for all badges $i = 1, \ldots, B$ do
  $\omega_i \sim \mathsf{Beta}(\alpha_\omega, \beta_\omega)$
  $\gamma_i^\mathrm{T} \sim \mathsf{Beta}(\alpha_\mathrm{T}, \beta_\mathrm{T})$
  $\gamma_i^\mathrm{F} \sim \mathsf{Beta}(\alpha_\mathrm{F}, \beta_\mathrm{F})$
  $\eta_i \sim \mathsf{Beta}(\alpha_\eta, \beta_\eta)$
  for all actions $j = 1, \ldots, F$ do
    $s_{ij} | \eta_i \sim \mathsf{Bernoulli}(\eta_i)$
    $\phi_{ij} \sim \mathsf{Beta}(\alpha_\phi, \beta_\phi)$
for all actions $j = 1, \ldots, F$ do
  $\phi_{\mathrm{bg},j} \sim \mathsf{Beta}(\alpha_\phi, \beta_\phi)$
for all users $u = 1, \ldots, N$ do
  for all badges $i = 1, \ldots, B$ do
    $b_i^{(u)} | \omega_i \sim \mathsf{Bernoulli}(\omega_i)$
    $\lambda_i^{(u)} | b_i^{(u)}, \gamma_i^\mathrm{F}, \gamma_i^\mathrm{T} \sim \mathsf{Bernoulli}\left(b_i^{(u)}\gamma_i^\mathrm{T} + (1 - b_i^{(u)})\gamma_i^\mathrm{F}\right)$
  for all actions $j = 1, \ldots, F$ do
    $a_j^{(u)} | \boldsymbol{b}^{(u)}, \boldsymbol{\phi}_{\bullet j}, \mathbf{s}_{\bullet j} \sim \mathsf{Bernoulli}\left(1 - (1 - \phi_{\mathrm{bg},j}) \prod_{i: b_i^{(u)}=1} (1 - \phi_{ij} s_{ij})\right)$
```

Figure 2: Plate diagram and generative model.

ing which actions can be explained by this particular badge, and their sparsity is controlled by a badge-specific prior η_i. Given that $s_{ij} = 1$, the variable ϕ_{ij} represents the probability that a user with badge i undertakes action j, in the absence of any other badges. For example, if a user only has the "runner" badge active, and $s_{\mathrm{runner},\texttt{\#runkeeper}} = 1$, meaning that the "runner" badge can explain tweeting `#runkeeper`, then our user will tweet `#runkeeper` with probability $\phi_{\mathrm{runner},\texttt{\#runkeeper}}$.

As a user may have more than one badge active that can explain a particular action, we combine their influence in a noisy-or fashion, indicating that a user performs an action j if at least one of his badges induce him to do so. Moreover, it is plausible that a user's behavior is influenced not just by his particular attributes, but by the environment at large, and thus we assume a background model $\phi_{\mathrm{bg},j}$, acting as a badge that every user shares, that has some probability of explaining every action.

Thus, formally, action $a_j^{(u)}$ is observed if it is explained by either the background or at least one of the badges of user u, which we can write as follows[5]:

$$p(a_j^{(u)} = 1 \,|\, \boldsymbol{b}^{(u)}, \boldsymbol{\phi}_{\bullet j}, \mathbf{s}_{\bullet j}) = \mathsf{Bernoulli}\left(1 - (1 - \phi_{\mathrm{bg},j}) \prod_{i: b_i^{(u)}=1} (1 - \phi_{ij} s_{ij})\right).$$

2.3 Prior probabilities

Keeping with a proper Bayesian approach, we specify prior distributions on our badge assignments $\boldsymbol{b}$, rates $\boldsymbol{\gamma}^\mathrm{T}$ and $\boldsymbol{\gamma}^\mathrm{F}$, sparsity masks $\mathbf{s}$ and action probabilities $\boldsymbol{\phi}$ encoding our modeling assumptions.

First, as some badges are more prevalent than others (e.g., there are likely more vegetarians on Twitter than machine learning enthusiasts), we assume that each badge assignment $b_i^{(u)}$ is drawn from a beta-distributed prior rate ω_i, shared across all users for each badge i.

Second, to encode that we expect the false positive rate to be considerably lower than the true positive rate, we place separate beta priors on $\boldsymbol{\gamma}^\mathrm{T}$ and $\boldsymbol{\gamma}^\mathrm{F}$, setting the hyperparameters accordingly.

Third, as we want each badge i to explain only a sparse set of actions, we place badge-specific beta-distributed prior rates η_i from which we sample s_{ij}, allowing different badges to have different degrees of sparsity.

Finally, we place vague beta priors on the action probabilities $\boldsymbol{\phi}$ seeking to learn these primarily from data.

A depiction of our graphical model and a summary of the full generative process can be found in Figure 2.

2.4 Badge inference

Given our model and the observations from each user, we wish to infer the latent badge assignments $\boldsymbol{b}$, as well as which actions are explained by each badge (and to what degree). As computing the exact posterior probabilities in a graphical model such as this is intractable, we employ Markov Chain Monte Carlo (MCMC) methodology and estimate the posterior probabilities on $\boldsymbol{b}$, $\mathbf{s}$ and $\boldsymbol{\phi}$ by deriving a Gibbs sampler with interleaved Metropolis-Hastings steps. In particular, we derive a *collapsed* Gibbs sampler, marginalizing out $\boldsymbol{\eta}$, $\boldsymbol{\omega}$, $\boldsymbol{\gamma}^\mathrm{T}$ and $\boldsymbol{\gamma}^\mathrm{F}$, leaving only the variables of interest. This results in the following sampler:

1. *Sample* $\boldsymbol{b}$. We sample each badge assignment $b_i^{(u)}$ for a particular user u from its conditional distribution, which we can write as proportional to the product of an action likelihood, a label likelihood and a prior:

$$p(b_i^{(u)} | \boldsymbol{a}^{(u)}, \boldsymbol{\lambda}_i, \boldsymbol{b}_i^{(-u)}, \boldsymbol{\phi}, \mathbf{s}, \boldsymbol{b}_{-i}^{(u)}) \propto p(\boldsymbol{a}^{(u)} | \boldsymbol{b}^{(u)}, \boldsymbol{\phi}, \mathbf{s}) \cdot p(\boldsymbol{\lambda}_i | \boldsymbol{b}_i) \cdot p(b_i^{(u)} | \boldsymbol{b}_i^{(-u)}). \quad (1)$$

2. *Sample* $\mathbf{s}$. We sample the binary variable s_{ij} from its conditional distribution, which we can write as follows:

$$p(s_{ij} | \boldsymbol{a}_j, \mathbf{s}_{(-ij)}, \boldsymbol{\phi}, \boldsymbol{b}) \propto p(s_{ij} | \mathbf{s}_{i(-j)}) \cdot p(\boldsymbol{a}_j | \boldsymbol{\phi}, \mathbf{s}, \boldsymbol{b}), \quad (2)$$

which is a product of a prior on s_{ij} and an action likelihood term. In practice, for statistical efficiency (and following Fox [5]), rather than sampling from the conditional distribution directly, we employ a Metropolis Hastings step with a deterministic proposal of flipping s_{ij} from some value s to its complement, $\bar{s}$ [6, 8].

[5] We note that, by assuming the influence of the badges to be independent of each other for a particular user, we can write this "at least one" clause as the complement of a product of complements. While this assumption may be violated in practice, we posit that the computational savings we achieve by this simplification will outweigh the induced bias.

Table 1: The 31 badges we defined for our experiments, as specified by their corresponding labels.

1	vegetarian	17	entrepreneur
2	Apple fanboy	18	golfer
3	cyclist	19	wine lover
4	gamer	20	book worm
5	runner	21	coffee
6	hacker	22	Harry Potter
7	feminist	23	Ruby on Rails
8	photographer	24	Manchester United
9	teacher	25	Hello Kitty
10	artist	26	anime
11	foodie	27	Warcraft
12	hipster	28	jetsetter
13	NASCAR	29	Taylor Swift
14	redneck	30	Lady Gaga
15	country music fan	31	jQuery
16	yoga		

3. *Sample* $\boldsymbol{\phi}_i$. To sample ϕ_{ij}, we first write its conditional distribution as a product of a prior and an action likelihood term:

$$p(\phi_{ij}|\boldsymbol{a}_j, \boldsymbol{\phi}_{-(ij)}, \mathbf{s}, \boldsymbol{b}) \propto p(\phi_{ij}) \cdot p(\boldsymbol{a}_j|\boldsymbol{\phi}, \mathbf{s}, \boldsymbol{b}). \quad (3)$$

   We use a Metropolis Hastings step here to obtain our sample, with a beta-distributed proposal distribution:

$$q(\phi'_{ij}|\phi_{ij} = \phi) = \mathsf{Beta}(\phi'_{ij}; \phi\nu, (1-\phi)\nu), \quad (4)$$

   parameterized with mean ϕ and effective sample size ν, meaning that each proposal is centered around the ϕ of the previous step.

4. *Sample* $\boldsymbol{\phi}_{\mathrm{bg}}$. We sample the background action probability, $\phi_{\mathrm{bg},j}$, in the same manner that we sample the per badge action probability, using a Metropolis Hastings step with the same proposal distribution.

Further details of our sampling algorithm, including complete derivations of all conditional distributions, can be found in the supplemental material [4].

3. EXPERIMENTAL RESULTS

3.1 Data

We evaluate our model on a data set of approximately seven million active Twitter users by monitoring Twitter's "firehose" stream in early August 2011, recording users with non-empty profiles. We scanned through these seven million users–which at the time of collection represented approximately 3.5 percent of all Twitter users–and manually defined a set of 31 badges by specifying a label for each one, based on the occurrence of a particular phrase or word in each user's Twitter profile. For example, we define a "vegetarian" badge by specifying:

$\lambda^{(u)}_{\text{vegetarian}} = 1$ *if user* u *has the word "vegetarian" in her Twitter profile.*

Table 1 contains a full listing of our 31 badges. We note that while these badges were defined as a proof of concept by the authors, a real personalization system would include a principled method for defining a large quantity of badges, as discussed further in Section 5.

Of the seven million users in our data set, we identified 376,916 that had at least one of the 31 labels in their profiles. We took this subset of users and monitored the firehose for one week, from 5 August 2011 to 12 August 2011, collecting every tweet and retweet by these users. This resulted in a set of approximately two million tweets. From these tweets, we extracted all unique hashtags (e.g., #`runkeeper`) and retweeted users (e.g., `@MacRumors`), defining a vocabulary of actions. We culled this vocabulary to remove any actions belonging to just a single user, leaving us with a final vocabulary of 32,030 actions (broken down into 18,003 hashtags and 14,027 retweets), performed by 75,880 different users. The most common action over this time period was the hashtag #`londonriots`, referring to the riots that took place in the British capital during the week of our data collection. Moreover, Figure 3(a) shows how many users in the data set described themselves using each of the 31 labels, with the most common being "artist."

Finally, we note that our model is not dependent on Twitter, as both the labels and the actions could be defined to take advantage of any other user signals one has access to, including location, shopping patterns, clicks, query logs and so on[6]. Twitter is, however, a convenient open platform for experimentation.

3.2 Evaluation

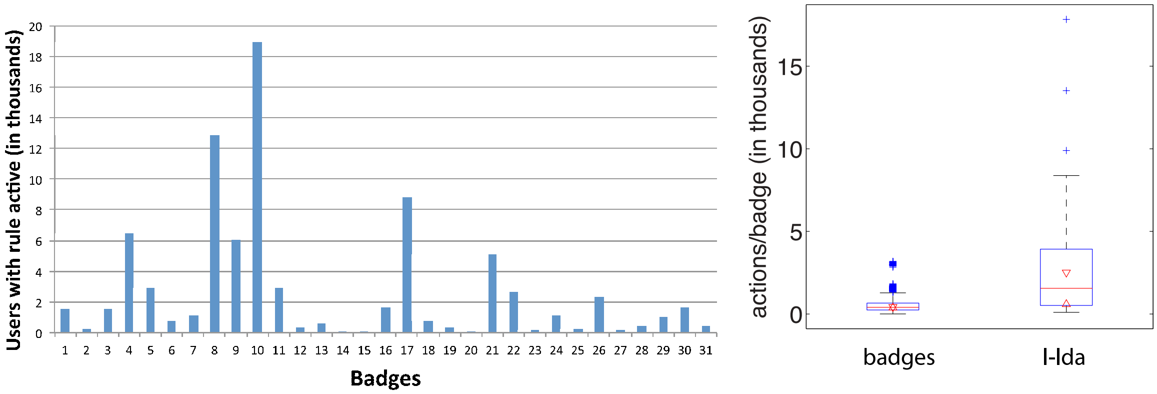
We ran our sampler on the data set described above, estimating posterior probabilities of badge assignments ($\boldsymbol{b}$) and badge definitions ($\boldsymbol{\phi}$ and $\mathbf{s}$) under our modeling assumptions. For each iteration, our sampler has time complexity $O\left(B\left(F+N\right)\right)$. Our implementation in the F# functional programming language achieves a runtime of approximately 3.5 minutes per sample, which is the time it takes to make a single, complete pass over more than 3.3 million random variables. Our hyperparameter settings and initialization condition are detailed in the supplemental material [4].

In an effort to compare our methodology to a state-of-the-art alternative, we wanted a model that: (1) can represent multiple labels per user; (2) provides a mechanism for interpreting the definitions of each badge; and, (3) can probabilistically infer badge assignments, especially in cases where the corresponding label is not present for the particular user.

We found the most suitable comparison technique to be the labeled latent Dirichlet allocation (labeled LDA) model of Ramage et al. [13]. This model makes a slight, but important, modification to the original LDA model, by assuming that each document is labeled with one or more tags, with each tag associated with one and only one topic. Thus, e.g., a document labeled with tags 1, 2 and 5 is assumed to have been generated from topics 1, 2 and 5, and no others. Like our model, labeled LDA provides a level of identifiability not obtained in traditional unsupervised approaches.

In order to adapt labeled LDA to our setting, we first associate a tag (and thereby a topic) with each badge, as well as an additional tag corresponding to a background topic. In our particular example, this leads to 32 unique tags. We then run labeled LDA twice: once for learning badge defi-

[6] For instance, we can imagine generalizing "labels" to instead represent any high-precision, low-recall *rule* that we intend on associating with a badge, not necessarily based on the content of a user's Twitter profile.



(a) Number of users with each label present in their Twitter profile

(b) Sparsity

Figure 3: (a) A bar chart indicating how many users in our Twitter data set have each of the 31 labels (corresponding to each of our badges.) (b) A box plot showing that the badge definitions we learn are significantly sparser than the topics learned by labeled LDA. The horizontal line in the middle of each box represents the median number of actions per badge, and the triangles delineate confidence intervals, giving us a 5% significance threshold.

nitions, and once for inferring badge probabilities for each user.[7] This two-stage approach contrasts with our model–which performs both functions simultaneously–and is necessary here because the labeled LDA model does not allow us to specify uncertainty in the label assignments.

In the first run of labeled LDA, we assign each user tags corresponding to the labels present in his or her user profile, as well as the background tag. For example, a user with the word "runner" in his Twitter profile would have to be modeled by only two topics: the one corresponding to the "runner" label, and the background topic. This first run learns a topic corresponding to each of the badges, giving the probability that each badge explains each action in our vocabulary. However, as each topic is a multinomial distribution over actions, its probabilities must sum to one, leading to qualitative differences with the badges learned from our model. First, for a given badge i in our model, the probability of each action ϕ_{ij} lies in the set $(0, 1)$, and are conditionally independent of each other given their prior. This allows several actions to have high probability of being explained by the same badge, if that is what can best model the user data. Second, by explicitly modeling sparsity using the $\mathbf{s}$ variables, a badge is not forced to explain actions it is only weakly associated with, simply for the sake of getting its distribution to sum to one. This is a characteristic not only of labeled LDA, but of all such topic models. Figure 3(b) shows the difference in the sparsity of our badge definitions when compared to labeled LDA.

After learning the topics with the first run of labeled LDA, we keep them fixed and infer the badge assignments, this time giving all 32 tags to each user, allowing for badges to be inferred beyond the ones corresponding to observed labels. However, as before, because labeled LDA provides no explicit model of sparsity, and is modeling a multinomial distribution, every user will, in expectation, be assigned to one badge, but this probability mass will be spread over all 32 topics, even if they are all unlikely.

We take the badges learned and inferred by our model and compare it to those from labeled LDA, evaluating both the interpretability of the badge definitions and the correctness of the badge assignments.

3.2.1 Interpretability of Badge Definitions

One important desideratum from our problem description is that, whichever model we use, if we are to bring transparency to the personalization process, we must provide users with meaningful and interpretable answers when they ask, "Why did I get badge X?" A convenient way to visualize badge definitions is via word clouds, with the size of an action proportional to its weight in the badge, concisely summarizing what it means to have a particular badge. Specifically, in our model, the "weight" of action j in badge i refers to the quantity $s_{ij}\phi_{ij}$, while in labeled LDA, the weight is the probability of action j coming from topic i.

Figure 4 shows six examples of badges learned from running our model on the Twitter data set described above.[8] These badges do an excellent job of describing actions that are consistent with their definitions, and are precisely the types of explanations we would hope to expose to the user. For instance, the "runner" badge in Figure 4(d) explains the action #`runkeeper`, which is a hashtag automatically tweeted when using a particular smartphone application[9] that helps manage and track a user's workouts.

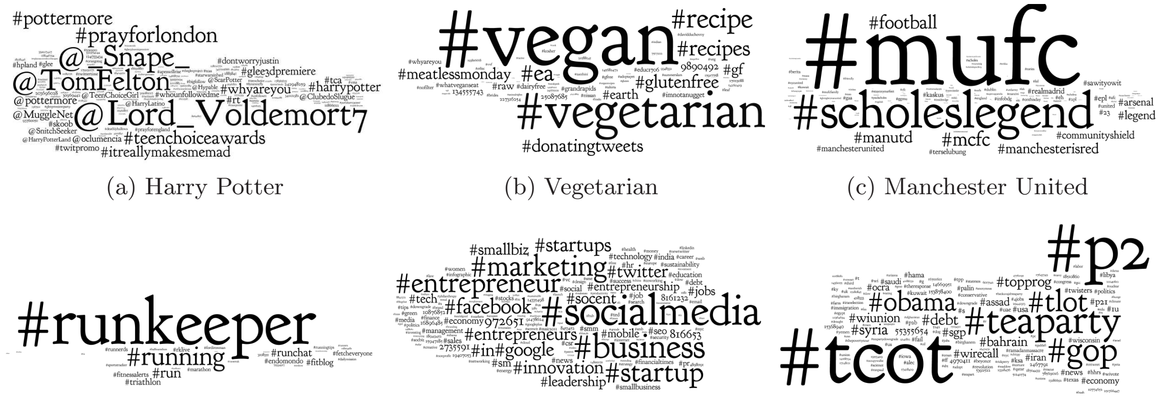
However, looking at Figure 4(f), which we learn by generalizing from the actions of self-described "rednecks," we find that while some actions are expected for such a badge (e.g., #`teaparty` and #`tcot`[10]), others are more surprising, (e.g., #`p2`, a popular hashtag among progressives). In fact, looking at other hashtags here such as #`obama`, #`debt` and #`syria`, we see that we actually learn a more general badge, referring to American politics and global affairs, rather than

[7] In both cases, we use the implementation provided by Ramage and Rosen in the Stanford Topic Modeling Toolbox, using the default hyperparameter settings and the CVB0 inference algorithm [14].

[8] A full visualization of all 31 badges learned from our model can be found with the supplemental material [4].

[9] `http://www.runkeeper.com`

[10] "Top Conservatives on Twitter"



(a) Harry Potter

(b) Vegetarian

(c) Manchester United

(d) Runner

(e) Entrepreneur

(f) Redneck

Figure 4: Word clouds (generated via `wordle.net`) representing six (of 31) badges learned by our model. Here, the size of a word is proportional to the action probability ϕ_{ij}.

one narrowly focused on rednecks. A plausible explanation for this phenomenon can be found in Figure 3(a), where we see that "redneck," associated with label 14, appears in the Twitter profiles of very few users–perhaps too few to effectively learn what it means to be a "redneck" on Twitter.

Figure 5 shows two other badges corresponding to labels present in very few users. The first example, Figure 5(a), is a more extreme form of overgeneralization than the "redneck" badge, as we see the idea of a wine lover translating to actions representing enjoyable (and often expensive) interests and activities, such as #`swarowski`, #`travel` and #`jewelry`. Figure 5(b), on the other hand, shows the extreme situation where the original label–in this case for "Ruby on Rails"–is present in so few users that the badge can be completely overwhelmed by a more popular topic. Here, we see that this badge has been taken by actions relating to the London riots, which was the most prevalent news item in our data.

When we look at the topics learned by labeled LDA, we find that they also represent interpretable badge descriptions. However, as we described earlier in this section, we do not find the same sparsity that we achieve using our model, because topic modeling approaches assume each topic is a distribution over the entire vocabulary. This contrast is made clear in Figure 6, where we see an extremely sparse badge representation for Apple fanboys, as learned by our model, compared to a much denser distribution over actions, learned by labeled LDA. The badge we learn focuses on a few informative actions, such as following `@MacRumors`, whereas the topic learned using labeled LDA includes, in addition to many Apple-related actions, many actions that are just tangentially related (e.g., #`runkeeper`).

3.2.2 Correctness of Badge Assignments

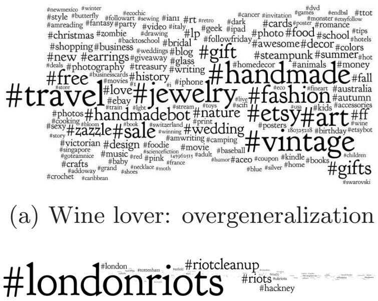
The fundamental goal of this work is not just to produce interpretable badges, but to accurately assign badges to users based on their actions. After all, badges are only useful for personalization if we infer them correctly. We expect our model to significantly outperform labeled LDA here, as we explicitly model the uncertainty relating badges and their corresponding labels, which labeled LDA does not.

In order to quantitatively measure our performance in this area, and compare it to that of labeled LDA, we retrained both models on the Twitter data set, this time holding out a random tenth of present labels, which we treat as ground truth labels that we seek to recover. Specifically, of all badge-user pairs (i, u) corresponding to present labels $\lambda_i^{(u)} = 1$ (of which there are 83,020), we select 10% uniformly at random, and hold them out. We then take the perturbed data set and run both labeled LDA (two stages, as before) and our model, leading to estimated posterior probabilities of badge assignments, $b_i^{(u)}$.

In order to compare the two models as fairly as possible, we take each user and rank his or her badges from most probable to least probable, and see where the held out points (i, u) appear. When ranking badges for a user u using our model, we rank them by descending posterior probability of $b_i^{(u)} = 1$. When ranking based on the labeled LDA model, we rank the topics for each user by decreasing topic proportion. If label i was held out for user u, then the better of the two models will rank badge i closer to the top.[11] Figure 7(a) demonstrates that our model significantly outperforms labeled LDA on this metric, with the held out badges appearing, on average, approximately four positions higher in the ranking. Moreover, we hypothesize that the more active a user is (i.e., the more actions we observe for her in our data set), the better we will do in predicting the held out badge, as we will have more information to base our inference on. This prediction is confirmed in Figure 7(b), where we see a six position difference in the ranking separating the most active from the least active users.

Moving beyond this quantitative comparison, we observe several qualitative properties of our inferred badges that provide anecdotal support for our model's effectiveness. First, as we model each user's badge assignments as a binary vector, we can use our samples to estimate the posterior marginal probability on *pairs* of badges appearing together in the same user, as predicted by our model. These results, shown in the annotated matrix in Figure 8, indicate that the hot spots of badge co-occurrence correspond to pairs of badges that we would expect to see together. In particular, the top four pairs of badges, ranked by decreasing posterior probability, are:

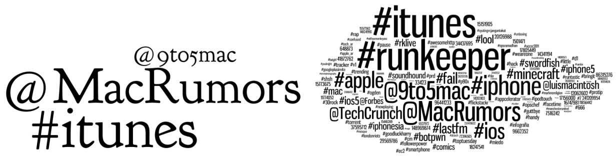
[11] We note that this is a fairer comparison than directly measuring the posterior badge assignment probability, since the labeled LDA probabilities are constrained to sum to one, giving us an unfair advantage. By ranking the badges, we avoid this problem.



(a) Wine lover: overgeneralization

(b) Ruby on Rails: overwhelmed

Figure 5: Two cases showing current limitations of our model that arise when we have labels present in only a small number of people.



(a) badges (b) labeled LDA

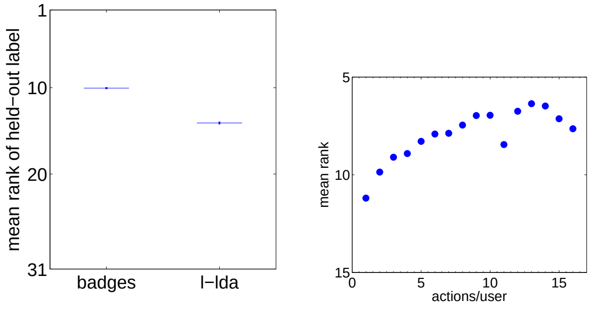
Figure 6: Two word clouds for the "Apple fanboy" badge, contrasting the sparsity of our learned badge (a) versus the corresponding labeled LDA topic (b).

1. Entrepreneur and jQuery
2. Feminist and "London riots" (originally the "Ruby on Rails" badge)
3. Feminist and "American politics" (the generalized "redneck" badge)
4. Photographer and "London riots."

The matrix shows many other correspondences, for instance, tying together pop icons Taylor Swift and Lady Gaga.

Another view of our inferred badge assignments can be obtained by selecting a population of Twitter users from our data set, and visualizing their collective badge profile in aggregate. Here, the size of a badge label in each word cloud is proportional to its mean posterior probability across the entire population of interest. For example, we can compute the mean probability that someone with the word "conservative" in his Twitter profile is assigned the "entrepreneur" badge. Figure 9(a) shows such a word cloud for precisely this population of conservative Twitter users from our data set. We see a heavy focus on the redneck badge among this population, since it's plausible that those with the word "conservative" in their Twitter profiles are apt to tweet about American politics. Interestingly, this is true even though conservatives are more likely to describe themselves as "teachers" or "entrepreneurs", as shown in Figure 9(b), potentially indicating that an interest in politics plays a larger role in influencing Twitter behavior than, say, being a teacher.

Finally, we can compare the inferred badge profiles of two different populations, and look at the differences between them. Naturally, to contrast with the "conservative" population of Twitter users, we also gather all users in our data set with the word "liberal" in their Twitter profile. We can than take our estimate of the mean posterior probability for each badge in each of the two populations, and visualize the difference between them, which we display in Figures 9(c)



(a) (b)

Figure 7: Results showing that, on a held-out set of labels, (a) our algorithm is better able to recover (approximately) ground truth badges than labeled LDA, and, (b) more active users get better predictions. (Error bars indicating standard error are too small to be visible.)

(conservatives - liberals) and 9(d) (liberals - conservatives). For example, if we look at the word "feminist" in Figure 9(d), its size is proportional to our estimate of the mean posterior difference, $p(b_{\texttt{feminist}}^{(\text{liberals})}) - p(b_{\texttt{feminist}}^{(\text{conservatives})})$. As expected, we see, e.g., that the "feminist" badge is the one that is most likely to occur in a liberal and not a conservative.

4. RELATIONSHIP TO PRIOR WORK

Since Zaslow's article on TiVo was published in 2002 [16], many forms of personalization have appeared or intensified on the Web, and they operate with varying degrees of transparency:

- The most transparent and interpretable personalization today is arguably done by Amazon.com. A user gets a selection of "recommended for you" items, with each being associated with a "because you purchased" explanation. If the recommendation is questionable, a user is allowed to correct the system by selecting a "because you purchased" item and indicating "don't use for recommendations." Amazon's feedback is similar to a user revealing the true value of an inferred badge in our model, but more specific. Furthermore, their item-to-item personalization holds the advantage that only user feedback is revealed in any explanations. The reliance on a well-represented catalogue is not feasible in many scenarios, though, and this is most notable where content is user-generated.
- Pandora[12] provides music recommendation based on likes and dislikes of each user, grounded in a "deeply detailed, hand-built musical taxonomy" known as the Music Genome Project[13]. Pandora's personalization service is quite transparent, with each song recommendation accompanied by an explanation for why it was selected, making clear the connection between user feedback and personalization. However, again, a hand-curated feature set like Pandora's is problematic to maintain for more dynamically generated content.
- Google provides users with a Privacy Center[14], from

[12] http://www.pandora.com

[13] http://blog.pandora.com/press/pandora-company-overview.html

[14] http://www.google.com/intl/en/privacy/

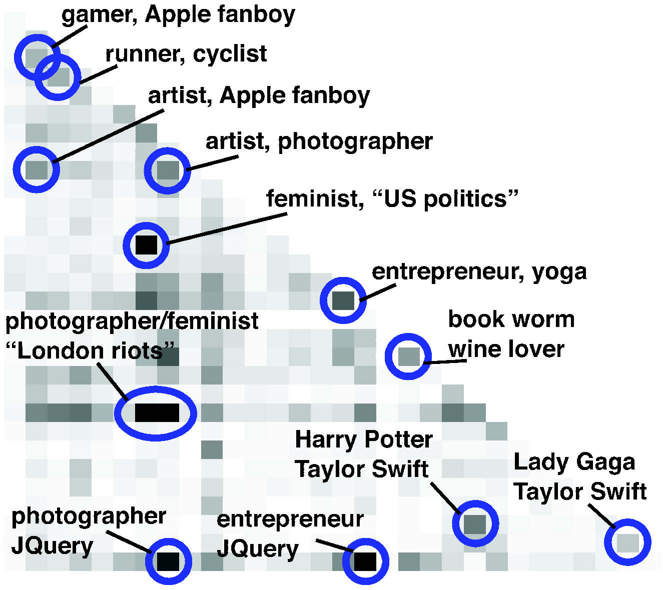
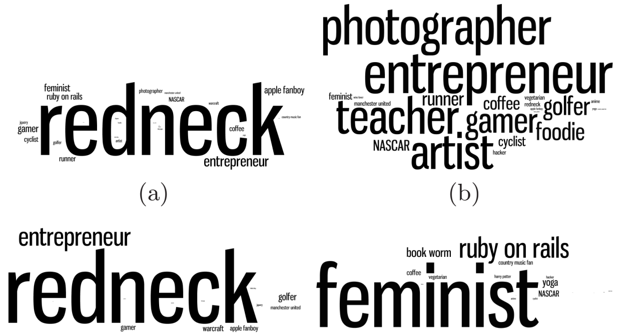


Figure 8: A matrix depicting the posterior marginal probability of badge co-occurrence, as estimated by our model. Darker squares represent higher probabilities. The badges are ordered in the same manner as presented in Table 1.

which they can opt out of having their Web and search history tracked, which affects personalization of the search results. Additionally, Google has recently allowed its users to see the attributes that it has inferred about them for ad prediction (e.g., "Demographics - Age - 65+"), giving them the option to decline any incorrect or undesired attributes[15]. While in the spirit of what we propose in this paper, this particular dashboard is focused on addressing the quality of personalized advertising, which is not the primary reason users interact with Google. At the time of writing, there does not seem to be any such window for viewing how search results are personalized.

- Bing search[16] is personalized based on search history when signed in with a Microsoft Live account; users see "Search history has changed the ranking of these results. Learn more." on the bottom of their search results. Clicking on "Learn more" provides an opportunity to toggle on or off the personalization, as well as information about how to remove some or all of the search from search history, but it is non-transparent how this affects personalization.
- Social networks such as Facebook and Google+ are (quite openly) repositories of structured personal information, and thus the primary transparency issue with these sites is not what information they infer about their users, but rather how it is used for personalization. For example, at the time of writing, Facebook's News feed is divided into a personalized list of "Top stories" and a timely list of "Recent stories," but there is no way for a user to see why a particular status update appeared in one feed and not the other.



(a) (b)

(c) (d)

Figure 9: These four word clouds depict the badges learned for two populations of people: those with the word "conservative" in their Twitter profiles (254 such users in our data set), and those with the word "liberal" in their profiles (327 of them). (a) Shows the mean badge probabilities for self-described conservatives. The size of a badge name is proportional to the mean probability of that badge as estimated by our model. (b) Shows the proportion of "conservatives" that have each of the 31 labels present in their profiles. (c) These badges are more likely for self-described conservatives than self-described liberals, and the size of each badge label is proportional to the absolute difference $(p(b_i^{(\text{conservatives})}) - p(b_i^{(\text{liberals})}))$. (d) Badges more likely for liberals than conservatives. (Recall from Figure 5(b) that the "Ruby on Rails" badge is instantiated as a "London riots" badge.)

While this is certainly not an exhaustive list, and personalization capabilities are routinely being added and removed from these sites and others, the websites mentioned represent a large portion of user interaction on the Web (e.g., Google and Bing search amounted to 85 percent of all web searches from the United States in September 2011[17], while over 500 million users log on to Facebook per day at the time of writing[18]), demonstrating the importance of studying methods for making personalization more transparent.

From the perspective of methodology, personalization by inferring latent user features has come a long way. While collaborative filtering through factorizing a user-item matrix, and variants thereof, is extremely successful as a backbone of recommender systems [7], the latent features don't have any interpretable meaning. Interpretability often means discretization. In this vein, Porteous et al. extended matrix factorization with discrete class allocations [11], and although we are not aware of its existence, it is entirely feasible to enforce labels to certain class allocations.

When the domain of interest includes user-generated content, like blogs and tweets, latent topic models have frequently provided a successful modeling framework. Unfortunately, as is the case with matrix factorization techniques, the topics learned in such unsupervised models do not lend

[15] http://googleblog.blogspot.com/2011/10/increasing-transparency-and-choice-with.html

[16] http://www.bing.com

[17] http://www.comscore.com/Press_Events/Press_Releases/2011/10/comScore_Releases_September_2011_U.S._Search_Engine_Rankings

[18] http://newsroom.fb.com

themselves to interpretability, as they are not identifiable, a problem we avoid by associating each badge with a label. Topic models that incorporate supervision at the topic level, such as labeled LDA, explained in Section 3.2, and the more recent work by Andrzejewski et al. [1], provide a mechanism for such identifiability. For instance, labeled LDA was recently used to model another axis of personalization on Twitter, mapping posts as informative, personal status updates, or inter-user social communication [12].

5. DISCUSSION

In this paper we have presented a Bayesian inference algorithm to learn both a descriptive model and predictor for *badges* based on user activity on a micro-blogging site (Twitter). In our work, a badge is seeded by a label—more generally, a high precision, low recall rule—based on self-reported user information, while our predictive model for badges explicitly relies on the presence or absence of user actions. Both these modeling decisions contribute greatly to the transparency of the prediction and we believe transparency will be a critical building block for personalization systems that users will find acceptable and suitable.

We have shown empirically that our model outperforms state-of-the-art models such as labeled LDA in terms of predictive performance, while producing interpretable descriptions of badges.

There are a number of open questions and challenges that need to be addressed in future work.

- **Scale**: The current inference algorithm is a combination of exact inference (*collapsing* some of the conditional priors; see Section 2.4) and Gibbs sampling with interleaved Metropolis-Hastings steps. The latter is applied for all B badges and F actions. This can become problematic if the number of actions grows unbounded, e.g., if an action is the presence of a word in a status update. In order to scale this approach to real-time streams, a single-pass approximate inference algorithm needs to be developed, with special attention paid to inference algorithms conducted in a distributed setting (cf. [9, 15]).
- **Dynamics**: Currently, our model assumes a fixed but unknown dependency of user actions and badges. In a more realistic setting, this dependency will vary over time. In future work, we will study a time-dependent model, allowing for both “topic drift” as well as the online addition and deletion of badges and actions.
- **Causality**: The proposed model is a purely “correlational” model which bases its inferences about latent badges on observed user activity. This can sometimes lead to *spurious* correlations which we also observed in our experiments, for example, with redneck badges. We have chosen this framework because our analysis is based on logs of activity only. Going forward, we will study causal models of badges by assuming we can influence the choice of actions that a user has (for example, suggest hashtags).
- **Crowd-sourcing**: As badges are defined simply by indicating a label, and in a realistic setting, a large number of them will be needed to model the full spectrum of user behavior, it is natural to consider a crowd-sourced solution for large-scale badge definition. Such an approach would maintain human supervision of badge definition (which we feel is necessary to ensure that badges are meaningful) while scaling our model to be more expressive.

Despite these current limitations and future challenges, our work introduces a promising framework that we believe will be successful in bringing transparency to a myriad of personalized services on the Web.

6. ACKNOWLEDGMENTS

First author was partially supported by ONR grants MURI N000141010934, N000140810752 and PECASE N000141010672.

7. REFERENCES

[1] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proc. IJCAI*, 2011.

[2] D. M. Blei and J. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: classification clustering, and applications*. Chapman and Hall, 2009.

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

[4] K. El-Arini, U. Paquet, R. Herbrich, J. Van Gael, and B. Agüera y Arcas. Transparent user models for personalization: Supplemental material. `http://www.cs.cmu.edu/~kbe/badges`.

[5] E. B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

[6] A. Frigessi, P. Di Stefano, C. Hwang, and S. Sheu. Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *Journal of the Royal Statistical Society, Series B*, pages 205–219, 1993.

[7] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[8] J. Liu. Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83(3):681–682, 1996.

[9] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Distributed GraphLab: A framework for machine learning and data mining in the cloud. In *PVLDB*, 2012.

[10] E. Pariser. *The Filter Bubble*. Viking, 2011.

[11] I. Porteous, A. Asuncion, and M. Welling. Bayesian matrix factorization with side information and Dirichlet process mixtures. In *Proc. AAAI*, 2010.

[12] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proc. ICWSM*, 2010.

[13] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[14] D. Ramage and E. Rosen. Stanford topic modeling toolbox. `http://nlp.stanford.edu/software/tmt/`.

[15] A. J. Smola and S. Narayanamurthy. An architecture for parallel topic models. *PVLDB*, 3(1):703–710, 2010.

[16] J. Zaslow. If TiVo thinks you are gay, here's how to set it straight. *The Wall Street Journal*, Nov. 26 2002.