Automated inference of point of view from user interactions in collective intelligence venues

Sanmay Das SANMAY@SEAS.WUSTL.EDU
Allen Lavoie ALLENLAVOIE@WUSTL.EDU
Washington University in St. Louis

Abstract

Empirical evaluation of trust and manipulation in large-scale collective intelligence processes is challenging. The datasets involved are too large for thorough manual study, and current automated options are limited. We introduce a statistical framework which classifies *point of view* based on user interactions. The framework works on Web-scale datasets and is applicable to a wide variety of collective intelligence processes. It enables principled study of such issues as manipulation, trustworthiness of information, and potential bias. We demonstrate the model's effectiveness in determining point of view on both synthetic data and a dataset of Wikipedia user interactions. We build a combined model of topics and points-of-view on the entire history of English Wikipedia, and show how it can be used to find potentially biased articles and visualize user interactions at a high level.

1. Introduction

The Web has enabled an unprecedented democratization of information. We increasingly rely on decentralized sources such as blogs, social news, and wikis to stay informed. While this transition has many benefits, it also creates opportunities for individuals and groups to shape available information and thereby influence public perception. As a result, reliability has been a primary concern since the early days of knowledge-sharing platforms such as Wikipedia.

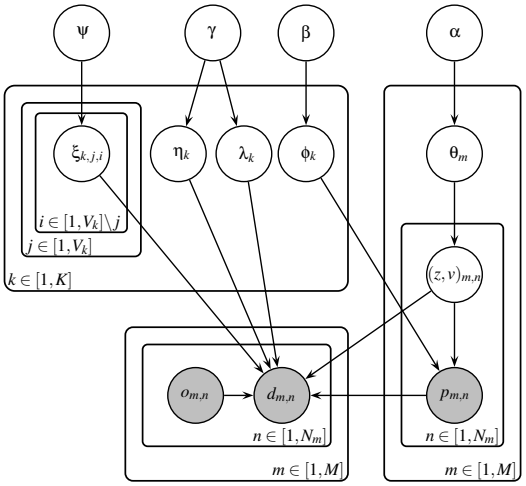
While there has been some work on systematically identifying manipulation and bias in purely quantitative collective intelligence venues like rating systems (Mobasher et al., 2007), work on venues with free-form information (like Wikipedia) has been less thorough and more anecdotal. In order to move towards a principled, quantitative methodology for evaluating bias and trustworthiness in such venues, in this paper we introduce a generative model of users' points of view. Our method is based on Latent Dirichlet Allocation (LDA), a popular topic model (Blei et al., 2003). By observing the pages a user chooses to edit and her interactions with other users on those pages, we are able to infer both topics and points of view simultaneously. While we focus on Wikipedia as a case study, our technique is general and can be used to study issues of bias and trust in many collective intelligence processes.

Two issues make identifying point of view a particularly hard—and, in our view, understudied—problem: subjectivity and scale. Point of view is notoriously difficult to quantify, even for humans considering single documents. There is little concrete information on which to base inferences about bias, and none of it is structured. The problem is exacerbated by an adversarial effect, whereby authors attempt to appear objective (Das et al., 2013). While the size and scale of the web are what makes sites like Wikipedia influential, this scale makes identifying point of view more difficult and human supervision problematic. The problem calls out for an efficient and accurate automated approach.

We pose the novel problem of identifying points of view in a large collective intelligence environment (e.g. Wikipedia). While point of view is fundamentally entangled with human communication, we posit that it also has observable characteristics in collective intelligence which do not involve natural language. In this paper, we show how to use data on user interactions to quantitatively study bias and point of view both in aggregate and on the level of individual users and documents.

Contributions We introduce a generative model of topics and points of view based on user interactions, and give an efficient inference algorithm based on Gibbs sampling. We study the performance of the model in inferring topics and points of view from synthetic data, finding that both are recoverable from the model's observed variables. Using a complete Wikipedia dataset, we perform model selection

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.



(a) Graphical depiction of the model using plate notation, where plates (boxes) represent repeated variables. Nodes in the first row are beta or Dirichlet distributions, nodes in the second row are categorical or Bernoulli. The shaded nodes are observed.

	Beta, Bernoulli distributions
ψ, $\xi_{k,j,i}$	Revert probability between POVs $i \neq j$
γ, η_k, λ_k	Different topic (η_k), same POV (λ_k) revert
	Dirichlet, categorical distributions
β, ϕ_k	Pages associated with each topic
α, θ_m	Topic and POV preferences for a user
$(z, v)_{m,n}$	Topic and point of view (categorical)
	Observed variables
$p_{m,n}$	The page an edit is on (categorical)
$o_{m,n}$	Parent edit (ordering; not modeled)
$d_{m,n}$	Whether edit disagrees with parent
	Counts, parameters
M	Number of users
N_m	Number of edits by user m.
K	Number of topics
V_k	Number of points of view for topic k
P	Number of pages

(b) Notation: random variables, distributions, and model parameters.

Figure 1. High level overview of the model.

and validate the approach by finding pairs of users with antagonistic relationships. Our approach, jointly modeling topics and points of view, significantly outperforms a social roles model, a model that fixes topics before considering points of view, and a non-Bayesian graph-based approach. Finally, we study the topics and points of view inferred from the entire history of English Wikipedia. This allows us to visualize shifts in point of view, revealing the evolution of the encyclopedia and its users over time, and provide insights into how Wikipedia functions. The model also provides a wealth of information about the process by which individual pages were created: as one example, we find pages on otherwise controversial topics which nonetheless are dominated by a single point of view.

1.1. Related work

Discovery of community structure in collective intelligence is a well studied problem. Kittur et al. (2007) use reverts to create a small-scale clustering of users while studying conflict and coordination on Wikipedia. Bogdanov et al. (2010) find communities on Wikipedia by comparing users based on multi-topic agreements and disagreements and then performing clustering, using LDA to inform the topic of text added or removed from a page. Pathak et al. (2008) propose a generative model for community extraction, modeling communication content. Sachan et al. (2011) find communities based on user-to-user links in a social network using a generative model, also considering interaction types. We are the first to exploit an explicit synergy between user interests (topics) and interactions (governed by points of view within a topic) in community discovery.

Another line of literature uses topical structure to model human communications. Rosen-Zvi et al. (2004) model documents from multiple authors, where authors have a distribution over topics, and words in a document are generated by first choosing an author, then a topic from that author's distribution, and finally choosing a word from that topic. McCallum et al. (2007) model directed messages, where both the sender and recipient are significant. We show that points of view within a topic can be used to effectively model the sentiment of communications.

Several models have been proposed to extract points of view or related concepts from natural language. For example, Paul & Girju (2010) study a multi-faceted topic model which can be used to find viewpoints in text. Lin & He (2009) model words in movie reviews as having both sentiment and topical components. Fang et al. (2012) find contrasting opinions in collections of text written from different perspectives: press releases from U.S. politicians and articles from major Chinese, Indian, and U.S. news sources. Their model exploits the often disparate language used when framing an issue from different perspectives (e.g. "life" and "choice" when debating abortion). However, this line of work (1) relies on clean sources of ideologically-relevant material, and (2) operates on a much smaller scale than the large collective intelligence processes we target. We address the former issue in part by modeling point of view in a user-centric way, which provides the statistical strength necessary to differentiate between hundreds of points of view across different topics (where previous work

assumes two or three total). The latter issue is addressed by using higher level observations: focusing on what users do rather than the specifics of what they say.

To summarize, we address three main challenges which prevent previous work on point of view modeling from applying to collective intelligence. (1) Very low signal to noise ratio: most Wikipedia revisions are not related to point of view. (2) Data on a completely different scale, with approximately four orders of magnitude more documents and words than previous work. (3) A lack of overarching ideologies across topics, which previous work takes advantage of in more specialized corpora. We present a novel model of point of view in collective intelligence based on user interactions, along with an efficient inference algorithm, which together address these challenges.

2. Model

We begin with a topic model much like Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is often used to model a set of documents, where documents are assumed to be generated by first selecting a topic from a document-specific distribution, and then selecting a single word from a topic-specific distribution, repeating this process for every word in each document. Instead of words we have pages, and instead of documents we have users: each user makes a collection of edits to different pages. As in LDA, each user (document) has a preference (probability distribution) over topics, and each topic has a distribution over pages (words). Each time a user makes an edit, they draw a topic from their personal topic distribution, and they then select a page to edit from that topic's distribution over pages. So far the only observable is the set of pages that each user chooses to edit, and edits are exchangeable within that set. This is exactly equivalent to LDA, where only the words in each document are observed.

Now suppose that each topic has a small number of points of view (POVs) a user editing on it can take, and each edit has one of these POVs associated with it in addition to its topic. Then each user has a distribution of preferences over (topic, POV) pairs rather than over topics alone. The chosen page for each edit still depends only on the topic of that edit, but POV determines interactions with other users on that page. We model this as a two-stage process: first every user chooses a topic, POV, and page for each of their edits, then interactions between users take place. These interactions are simple: for each edit, the user decides to disagree with its parent edit or not. Each edit's parent is observed but is not modeled: edits, except those at the beginning of a page, have an externally specified parent denoted $o_{m,n}$ in the model. There are three cases: (1) the parent edit is on a different topic; (2) the parent edit is on the same topic and the same POV; (3) the parent edit is on the same topic and a different POV. Disagreements in cases (1) and (2) might be mundane and unrelated to POV: style or formatting, for example. The probability of a disagreement in these cases is determined by the topic of the latter editor, and we would expect such disagreements to be unlikely. In case (3), however, the probability of a disagreement is determined by the relationship between the two POVs involved: some might have a very antagonistic relationship, others less so, but we would expect more disagreements here than in cases (1) and (2) as a fraction of the opportunities for disagreement. We refer to disagreements in case (3) as POV disagreements, and to others as non-POV disagreements. The model does *not* interpret every Wikipedia revert as being a disagreement relevant to POV. Although we expect more reverts in case (3) as a fraction of opportunities for disagreement than we do in (1) or (2), the preponderance of cases (1) and (2) means that we would expect them to produce a significant fraction of all reverts.

for topic $k = 1 \to K$ **do**
  $\eta_k, \lambda_k \sim \text{Beta}(\gamma)$ // Non-POV revert probabilities
  $\phi_k \sim \text{Dirichlet}(\beta)$ // Page distribution
  for POV $j = 1 \to V_k$ **do**
    for POV $i = 1 \to V_k$ excluding j **do**
      $\xi_{k,j,i} \sim \text{Beta}(\psi)$ // Probability POV j reverts i
for user $m = 1 \to M$ **do**
  $\theta_m \sim \text{Dirichlet}(\alpha)$ // Topic and POV preferences
  for edit $n = 1 \to N_m$ **do**
    $(z, v)_{m,n} \sim \text{Categorical}(\theta_m)$ // Edit's topic and POV
    $p_{m,n} \sim \text{Categorical}(\phi_z)$ // Edit's page
for user $m = 1 \to M$ **do**
  for edit $n = 1 \to N_m$ **do**
    if $z_{o_{m,n}} = z_{m,n}$ **then**
      if $v_{o_{m,n}} = v_{m,n}$ **then**
        $d_{m,n} \sim \text{Bernoulli}(\lambda_k)$ // Disagree? Same POV
      else
        $d_{m,n} \sim \text{Bernoulli}(\xi_{k,v_{m,n},\text{POV}(o_{m,n})})$ // Diff. POV
    else
      $d_{m,n} \sim \text{Bernoulli}(\eta_k)$ // Different topic

Figure 2. Generative model pseudo-code. $x \sim D(y)$ indicates a random variable x drawn from distribution D parameterized by y.

Figure 1(a) depicts this model graphically, and Figure 1(b) summarizes the notation. Figure 2 provides a rigorous pseudo-code description of the model.

Simplifying assumptions We use symmetric Dirichlet distributions, with a single parameter (α and β for pages and (topic, POV) pairs, respectively). We set $\beta = 0.1$ and $\alpha = 5/(VK)$ where K is the number of topics and V is the number of POVs per topic: we expect users to be focused on a small number of (topic, POV) pairs (and as we add more topics or POVs, we expect them to become increasingly specialized). These choices are similar to those of

Griffiths & Steyvers (2004) for the equivalent LDA parameters. For the remainder of the paper, $V = V_k$: every topic has the same number V of points of view. For the beta distributions, we set $\psi_\alpha = 0.8$, $\psi_\beta = 0.2$ and $\gamma_\alpha = 5$, $\gamma_\beta = 95$: for example, $\eta_k \sim \text{Beta}(\alpha = 5, \beta = 95)$. This encodes the belief that disagreement probabilities will be low for non-POV interactions, and may be higher for POV interactions.

For an edit B which has an opportunity to disagree with edit A, we refer to A as B's parent. If an edit C has an opportunity to disagree with B, then C is B's child. All references are between edits on the same page. We assume that each edit references (has the opportunity to disagree with) at most one other edit (its parent), and is itself referenced by at most one edit (its child). As in Kittur et al. (2007), an edit's parent is the immediately preceding edit, and a disagreement (if any) is only with that edit. This disregards complexities which can arise when an edit reverts multiple prior edits, or when a single edit makes a complex contribution and subsequent edits disagree with different parts of it, but simplifies while being correct in most situations.

2.1. Inference

Griffiths & Steyvers (2004) introduce a collapsed Gibbs sampler for inferring LDA's latent variables, integrating out the real-valued categorical distributions associated with documents and topics. A single Gibbs iteration samples each latent variable once according to its full conditional distribution (conditioning on the values of all other latent variables). For LDA, this means that each topic assignment is re-sampled taking into account the most recent topic assignments for all other words (edits in our case). The algorithm eventually converges to a stationary distribution, where topic assignments are drawn from their posterior distributions given the observed data.

We use a parallel approximation to collapsed Gibbs sampling for inferring the topic and POV of each edit. The collapsed Gibbs sampler is similar to that of LDA, with both topic and POV repeatedly re-sampled rather than the topic only. We use a tightly-coupled parallel approximation to Gibbs sampling—similar to the GPU inference of Yan et al. (2009) and Approximate Distributed LDA (Newman et al., 2008)—where each thread starting a (topic, POV) re-sample takes into account all previously recorded assignments, then is itself recorded before going on to the next edit (maintaining consistency). This tight coupling increases communication between threads (using shared memory extensively), but comes as close as possible to true Gibbs sampling (where sampling is serial).

Re-sampling is according to the full conditional probability of a (topic, POV) pair for a single edit, conditioning on the assignments of all other edits. As in LDA, this depends on the probability of the user selecting a given (topic, POV) pair, and the probability of selecting the observed page given that choice. Additionally, it depends on the probability of a disagreement between the edit and its parent (if any), and its child (if any). This full conditional distribution can be written as:

$$p((z,v)_{m,n} = (k,j) \mid (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})$$
$$\propto \left(n^{(m)}_{-n,(k,j)} + \alpha\right) \frac{n^{(p_{(m,n)})}_{-(m,n),(k,j)} + \beta}{n^{(\mathbf{p})}_{-(m,n),(k,j)} + P\beta}$$
$$p(d_{m,n} \mid (z,v)_{m,n} = (k,j), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})$$
$$p(d_{\text{child}(m,n)} \mid (z,v)_{m,n} = (k,j), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})$$

Where $n^{(A)}_{-B,C}$ denotes a count across object(s) A excluding the assignment of B on topic C. Bold variables denote a vector of all the associated values (see Figure 1(b)), except $(\mathbf{z},\mathbf{v})_{-(m,n)}$, which excludes the n^{th} edit by user m (the edit currently being re-sampled). For edits lacking a parent or a child, the corresponding disagreement probability is omitted. We omit the normalizing constant on the factor $n^{(m)}_{-n,(k,j)} + \alpha$ (the probability of the user selecting this (topic, POV)), as the normalizing constant does not depend on the (topic, POV) being considered. The probability of observing a disagreement depends on the topic and POV assignments of the edit r and its parent o_r:

$$p(d_r \mid (z,v)_r = (k,j), (z,v)_{o_r} = (k',j'), (\mathbf{z},\mathbf{v})_{-(m,n)}, \mathbf{p}, \mathbf{d}, \mathbf{o})$$
$$= \begin{cases} k = k', j = j' & \frac{n^{(z=z', v=v', d)}_{-(m,n),k} + \gamma_\alpha}{n^{(z=z', v=v')}_{-(m,n),k} + \gamma_\alpha + \gamma_\beta} \\ k \neq k' & \frac{n^{(z\neq z', d)}_{-(m,n),k} + \gamma_\alpha}{n^{(z\neq z')}_{-(m,n),k} + \gamma_\alpha + \gamma_\beta} \\ k = k', j \neq j' & \frac{n^{(z=z', v=j, v'=j', d)}_{-(m,n),k} + \psi_\alpha}{n^{(z=z', v=j, v'=j')}_{-(m,n),k} + \psi_\alpha + \psi_\beta} \end{cases}$$

In the above counts, we ignore disagreements between the edit under consideration and its parent *and* child (as those depend on the edit's previous (topic, POV) assignment). The variable $n^{(A)}_{-B,C}$ again denotes easily computable counts based on the topic and POV assignments of other edits and whether those edits disagree or not.

Inference then consists of repeatedly drawing new (topic, POV) pairs with probability proportional to the full conditional distribution specified above. An iteration of Gibbs sampling consists of re-sampling the topic and POV of each edit once. To initialize, we randomize each assignment.

Computation We use 64 threads in parallel on a single machine (64 cores) for inference. Sampling with 200 topics and 4 POVs takes approximately 6000 CPU hours for 200 burn-in iterations and 400 additional samples (we save every fifth). However, this sampling need only be done

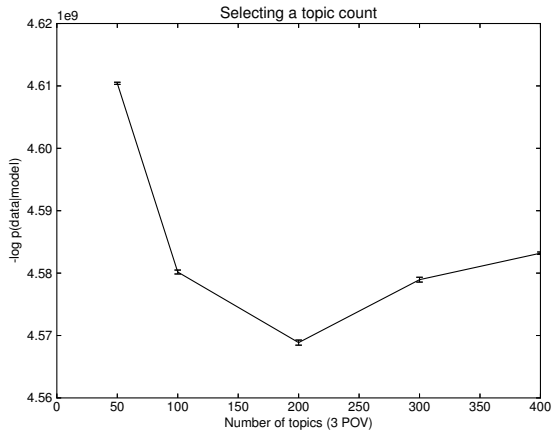


Figure 3. Negative log likelihood with the number of points of view fixed at 3. Error bars: twenty times standard error.

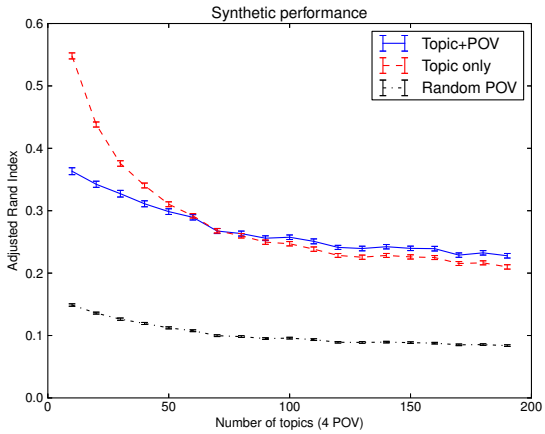


Figure 4. Synthetic performance, clustering edits according to their (topic, POV). Also shows performance when POV is ignored (Topic only), and when POV is randomized within a topic (Random POV). Error bars show the standard error of the mean.

once: the 80 saved assignments and one high-probability assignment of topics and POVs to revisions are all that is required to produce the results we present (aside from synthetic data and model selection), and these samples can be reused to compute new page and user statistics on the fly. The time complexity of each Gibbs sampling iteration is $O(KVN)$ where $N = \sum_m N_m$ is the total number of edits.

3. Data and model selection

Dataset We use the complete edit history of English Wikipedia as of November 2012, with 31583222 users, 9806233 pages, and 341026287 total edits. For anonymous users, we treat all edits from the same IP address as belonging to one user. Reverts are modeled as disagreements, either when the hash of a page matches the hash of a previous version of that page, or when "revert" or "rv" is mentioned in the edit comment. Wikipedia edits have a parent defined where applicable, which we honor except in rare cases where more than one edit has the same parent or the parent is on a different page (in the case of merged/split pages); in these cases, we treat the edit as not having a parent. We only use edits to pages in namespace 0 (the article namespace), ignoring talk and administrative pages.

Selecting a topic and POV count We perform model selection by estimating $p(\mathbf{p},\mathbf{d}|K,V,\gamma,\psi,\alpha,\beta)$: the probability of the observed variables given the model, with the topic and POV assignments integrated out. We use an estimator due to Murray & Salakhutdinov (2009), which exploits forward and backward transition operators of the Markov chain and is easily implemented on top of a Gibbs sampler. For a comparison, see Wallach et al. (2009).

Fixing the number of POVs V at 3, we find that the data is most likely under a model with $K = 200$ topics; see Figure 3. The number of topics and the number of POVs per topic are not interchangeable: fixing the product $KV = 600$ and trying $V \in \{1,\ldots,6\}$, $V = 3$ and $K = 200$ again assigns the highest probability to the data. We then optimize the number of POV V, fixing $K = 200$, and find that $V = 4$ maximizes the probability of the data, although the difference is smaller than for the number of topics. We now fix $V = 4$ and $K = 200$ unless otherwise noted.

While we do assume a fixed number of points of view, there is flexibility built into the model. Points of view are not necessarily antagonistic (the relevant prior, ψ, is very weak). If there are only two points of view A and B on a topic, the model is free to create points of view A_1, A_2, B_1, and B_2 such that A_1 and A_2 have a positive relationship to each other and a negative relationship to the B points of view. We can model any small number of "true" points of view in this way (below 5 if $V = 4$). We now turn to validation of the model assumptions, using real user relationships to test the inferred relationships between points of view.

4. Experimental validation

4.1. Synthetic experiments

Topic modeling can be viewed as a clustering problem, where words are assigned to topical clusters. In our case, we wish to assign edits to (topic, POV) clusters. In order to test the success of this method, we first generate the data directly *from* the model, and then perform inference and check the "correctness" of the inference. A perfectly "correct" clustering is unlikely – topics overlap, and there is limited additional information when there are multiple edits by the same user on the same page (although our

model does leverage additional information in the form of disagreements). However, given the success of LDA in the past decade, we can compare our method with the baseline of simply using a topic model, ignoring points of view.

In order to do so, we generate data with 5000 users, 1547 pages (roughly keeping the ratio of users:pages the same as the Wikipedia data), and an upper truncated Pareto distribution for the number of edits per user (Ortega, 2009) with an upper truncation of 147696 and a slope of 0.8, roughly matching the Wikipedia distribution. The parameters $\alpha = 5/(VK)$, $\beta = 0.1$, $\psi_\alpha = 0.8$, $\psi_\beta = 0.2$, $\gamma_\alpha = 5$, $\gamma_\beta = 95$, $V = 4$, and K are the same for generation and inference. We evaluate a high probability assignment of topics and POVs found through iterative maximization after 100 iterations of Gibbs sampling, comparing the resulting clustering to the true assignments from the synthetic data.

Figure 4 varies the number of topics K, measuring clustering performance using the adjusted Rand index, which is corrected for chance (its values are between -1 and 1, with 0 being the expectation of a random assignment and 1 being a perfectly correct assignment). The algorithm is able to effectively infer points of view in addition to topics, on par with how well LDA infers topics. Small values of K make the topic assignment problem easier (consider a trivial example with a single topic), while a fixed number of POVs per topic keep the (topic, POV) assignment problem challenging even when considering very few topics.

4.2. Rule violation reports, reverts, and baselines

We turn now to validating our model on real data. We collect a dataset of *rule violation reports*, where one Wikipedia user reports that another has violated Wikipedia's Three Revert Rule (3RR): a user may not perform more than three reverts on a single page in a 24-hour period (subject to an administrator's interpretation and conditions on what constitutes a revert for 3RR purposes). The act of reporting another user implies a significant disagreement: reporting a user who shares your point of view, while a noble concept, is understandably unlikely in practice.

This gives us 7179 unique pairs of antagonistic users, where one has reported the other for a 3RR violation. Along with negative examples, they form comparisons:

RRP Randomly permuted reporting pairs provide negative examples, generating 7179 random pairings. Disputes are on specific topics, so random pairs with significant disagreements should be unlikely.

NR RRP with pairs of users who have reverted each other removed, leaving 986 reporting pairs.

WP With page information. Positive examples are where a reporting pair has edited consecutively on the same page, negative examples are from consecutive random edits by users who have never reverted each other (respectively 19683 positive and negative examples).

SP WP restricted to the set of pages that have both positive and negative examples, to eliminate any effects from choosing more controversial pages (4252 positive, 4536 negative examples).

For the datasets without page information (RRP and NR), we consider a thought experiment, placing edits by a pair of users next to each other on the same page, each edit serving as the parent with probability 0.5. We can then compute the expected probability of a POV disagreement over the possible assignments of topic and POV to the two edits, taking into account the topic and POV preferences of the users and the relationships between each POV. This POV disagreement probability measures the level of antagonism between users as inferred by the model. Viewing RRP as a ranking task, area under the ROC curve (AUC) is 0.85: a randomly selected true report pair will have a higher disagreement probability than a randomly selected non-report pair 85% of the time (0.5 is random guessing). Removing pairs who have reverted each other at least once (NR), the model is still quite discriminative, with an AUC of 0.72 on this more difficult task. How much of this performance is due to topical—rather than point of view—disparities between users in the permuted pairs? We address this question using the datasets WP and SP, which restrict examples to the same pages, and hence mostly to the same topic, and find that the model still performs well (Table 1). Computing the model's probability of any revert, rather than the probability of a POV revert specifically, yields significantly worse performance on all of these datasets: the model is not predicting reverts, it is predicting POV disagreements.

Why does the model work well, and are there alternative, simpler models that may be as powerful? We consider two alternatives. One hypothesis is that there are roughly four social roles on Wikipedia, and that users can be described just as well by these four groups as by many groups split across topics (i.e. disputes are not topical). In order to test this hypothesis, we consider a baseline model with a single topic and four points of view, using the same methodology as for the full model when ranking pairs of users. This model nets an AUC of 0.57 on the full permuted user interactions dataset, but loses its discriminative power when we remove pairs who reverted each other (AUC 0.48). It does much better when non-reporting pairs are chosen to have edited consecutively on the same page (WP, 0.69), and better still when the reporting and non-reporting pairs are on the same set of pages (SP, 0.75). Social roles seem to play a role in animosity, but fail in cross-topic comparisons.

A second idea is that disputes are *entirely* topical. Is there any benefit to having a full model of topics and points

of view over first determining topics and then clustering within topics to find points of view? We evaluate a two-level model which does the latter: first fixing 200 topics (standard LDA, but a single high probability assignment), then clustering revisions within those topics into four points of view. This hierarchical baseline consistently performs significantly worse than the simultaneous model. Table 1 summarizes these model comparisons.

Table 1. Model comparisons (AUC). Pairwise differences within each dataset are significant ($p < 0.01$) except for starred* pairs, computed via empirical overlap across 10^4 bootstrapped datasets. Bootstrapped 95% confidence intervals are $\pm.01$ for RRP, $\pm.02$ for NR, $\pm.01$ for WP, and $\pm.02$ for SP and Reverts.

Model	RRP	NR	WP	SP	Reverts
Social roles	.57	.48	.69	.75	**.88**
Hierarchical	.80	.68	.71	.72	.80
Simultaneous	.85*	**.72**	**.82**	**.80**	.85
ApproxMaxCut	.85*	.57	.59	.62	.51*
RevedSamePage	**.88**	.62	-	-	.50*

Table 1 includes results for revert prediction: given held-out pairs of users with page information (794 reverts in 6835 edits, 654 pages), determine whether the latter editor reverts the former. The social roles model does very well on this task. This is likely because most reverts are not related to POV disputes, but instead are typical "maintenance" tasks on Wikipedia; simple revert prediction is not our goal. The value of the simultaneous model is in domain adaptation: trained on reverts, it not only predicts those, but also more reliable indicators of user relationships, as demonstrated by the other datasets. Even though the majority of reverts are maintenance tasks, other reverts do contain information about deeper topical disputes, which can be harnessed by considering topics and points of view.

Also in Table 1 are results for two simple non-Bayesian baselines. ApproxMaxCut first builds graphs of users, with an edge if either user has reverted the other at least once on a given page. It then partitions users on each page into two groups via an approximate maximum cut, computed by selecting the best of 50 greedily optimized random partitions. For the datasets with page information (WP and SP), it predicts a positive label if the users are on opposite sides of that page's cut and negative otherwise. For the datasets without page information (RRP and NR), its predicted scores are a zero-one average across pages the users have edited in common (zero if they are on the same side of a cut, one otherwise). The performance is generally poor, with the exception of RRP. For RRP, memorizing pairs of users who have any relationship at all is profitable, since the permuted pairs are unlikely to have any pages in common while the reporting pairs almost certainly do. To illustrate this, RevedSamePage predicts a positive label when two users have reverts on at least one page in common, and a negative label otherwise. The Bayesian models do not memorize, instead summarizing relationships with a small number of topics and points of view, yet still excel on RRP.

Table 2. Selected topics, with the top pages by number of edits on that topic (ignoring POV). From a high probability assignment.

Topic 61	Topic 68	Topic 23
Killer whale	Anarchism	2006 Lebanon War
Tiger	Race and IQ	Muhammad
Lion	Capitalism	Gaza War
White shark	Libertarianism	Islam
Cougar	Iraq War	Israel
Giraffe	Socialism	Lebanon

Table 3. Active pages (more than 100 editors) which—as of November 2012—had more than 60% of their edits on a single, controversial POV of a controversial topic.

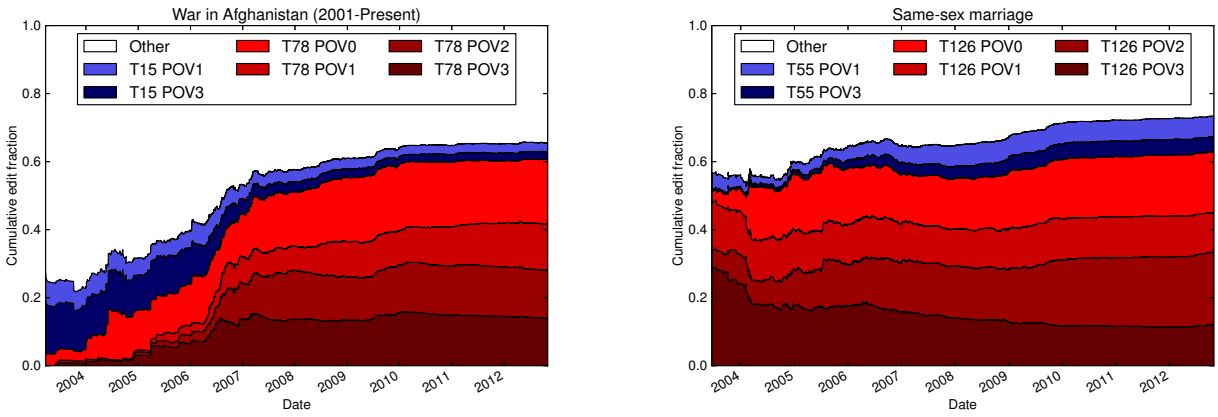
Page title	POV%
Private finance initiative	64%
World War II casualties	67%
John Prendergast (activist)	65%
1948 Palestinian exodus from Lydda and Ramle	70%
Chilean presidential election, 2005–2006	69%

5. Model insights

What can points of view tell us about Wikipedia? Having validated the model, we now explore some of its insights.

Unlike in traditional topic modeling, where documents are mixtures of topics, we model *users* as such mixtures. This leads to topics where pages are grouped semantically—as we would expect from a traditional topic model on documents and words—only if user behavior is well explained by those topics. Table 2 includes examples of such topics that contain pages which are deeply semantically similar. However, our model also reveals topics which have more to do with user behavior than with the subject of edited pages: for example, one topic deals exclusively with vandalism and those who remove it from the encyclopedia. A user may then be a mixture of not only several topics but also several kinds of topics (e.g. animals and anti-vandalism).

Changes over time Labeling each revision with a point of view allows us to visualize page dynamics. Has the nature of a conflict changed over time? Were the current points of view always well represented? Figure 5 shows active topics and POVs over time on two popular pages. Figure 5(a) shows a shift in the topics used to explain editing and edit conflicts: early Wikipedians were often—by necessity, considering the number of editors—generalists. With growth, editors became increasingly specialized. This



(a) Specialization over time. Topic 15 encompasses many disputes—terrorism, politics, and articles about Wikipedia itself—and is used to explain many early edit conflicts. As Wikipedia matured, users specialized more: topic 78 can be described as "contemporary wars", and better explains later conflicts on this page. Topic 78, POV 0 is composed of casual editors (17 on-POV edits/user), while POV 3 consists of "power editors" (269 edits/user).

(b) Topic 126 covers issues related to human gender and sexuality, with POV 0 generally taking a more socially conservative stance. POV proportions on this page are relatively stable, after an initial increase in opposition (POV 0) as the encyclopedia became more notable. Topic 55 explains the interactions between vandals and those who remove vandalism, and shows up on many popular pages.

Figure 5. Cumulative fraction of edits on the top 6 topics and POVs for two popular pages (War in Afghanistan and Same-Sex Marriage).

shift is reflected in the topics represented on the page, and in the points of view used to explain the changing conflict.

Figure 5(b) shows a traditional topic—dealing with the page's subject matter—coexisting with a behavioral topic explaining the interactions between vandals and anti-vandals. Points of view show a similar duality: POVs in Figure 5(a) deal as much with types of users (casual vs. heavy editors) as with page content, whereas those in Figure 5(b) are more focused on subject matter disputes.

As an aside, some point of view disputes are not apparent from natural language, e.g. the "modern wars" topic includes a dispute over WWII casualty numbers. Many disputes over figures have this property, and vandalism is another case where actions are more informative than words.

Page and user statistics Modeling POV provides a rich source of information about pages and users. Consider the problem of finding pages which could benefit from contributions by editors with a different POV: the model allows us to not only find these pages, but also to find users on different POVs who could be interested in the topic. For example, Table 3 shows the five most controversial pages that had more than 60% of their edits come from a single, controversial POV of a controversial topic. Here we define controversial topics as those with rare same-POV reverts ($< 3\%$) and more common different-topic reverts ($\geq 6\%$), and controversial POVs as those that have a high probability of reverting or being reverted by a different POV on the same topic ($\geq 30\%$). The model provides flexibility in querying for specific patterns over topics and POVs.

6. Discussion

As we become increasingly reliant on collective social processes to aggregate information, understanding these processes is critical. In the presence of incentives for manipulation, having information sources wear their biases "on their sleeves" has enormous value both for users who must evaluate information from multiple sources, and for information sources themselves attempting to maintain credibility. We propose a scalable model which takes a first step toward uncovering bias in collective intelligence processes, and which can help aggregation venues police themselves.

This kind of modeling applies to a wide variety of collective intelligence and aggregation venues, and has the potential to make online information sharing more transparent. By augmenting human judgment with machine inferences from large datasets, we can ease the transition from traditional centralized information aggregation models, allowing more reliable and more useful information sharing.

Acknowledgments

This work was supported in part by a US National Science Foundation (NSF) CAREER award (IIS-1414452), and in part by NSF grant IIS-1124827.

References

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003. ISSN 1532-4435.

Bogdanov, Petko, Larusso, Nicholas D., and Singh, Ambuj. Towards community discovery in signed collaborative interaction networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pp. 288–295, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4257-7. doi: 10.1109/ICDMW.2010.174.

Das, Sanmay, Lavoie, Allen, and Magdon-Ismail, Malik. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. In *Proceedings of the Twenty-Second ACM Conference on Information and Knowledge Management*, CIKM '13, pp. 1097–1106, 2013.

Fang, Yi, Si, Luo, Somasundaram, Naveen, and Yu, Zhengtao. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, pp. 63–72. ACM, 2012.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April 2004. ISSN 0027-8424. doi: 10.1073/pnas.0307752101.

Kittur, Aniket, Suh, Bongwon, Pendleton, Bryan A., and Chi, Ed H. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 453–462, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240698.

Lin, Chenghua and He, Yulan. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 375–384, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646003.

McCallum, Andrew, Wang, Xuerui, and Corrada-Emmanuel, Andrés. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1): 249–272, October 2007. ISSN 1076-9757.

Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology*, 7(4):23, 2007.

Murray, Iain and Salakhutdinov, Ruslan. Evaluating probabilities under high-dimensional latent variable models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21*, pp. 1137–1144. 2009.

Newman, David, Asuncion, Arthur, Smyth, Padhraic, and Welling, Max. Distributed inference for latent Dirichlet allocation. In Platt, J.C., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1081–1088. MIT Press, Cambridge, MA, 2008.

Ortega, Felipe. Wikipedia: A quantitative analysis. Doctoral thesis, Universidad Rey Juan Carlos, Móstoles, Spain, April 2009.

Pathak, Nishith, Delong, Colin, Banerjee, Arindam, and Erickson, Kendrick. Social Topic Models for Community Extraction. In *The 2nd SNA-KDD Workshop '08 (SNA-KDD'08)*, August 2008.

Paul, Michael and Girju, Roxana. A two-dimensional topic-aspect model for discovering multi-faceted topics. *AAAI Conference on Artificial Intelligence*, 2010.

Rosen-Zvi, Michal, Griffiths, Thomas, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 487–494, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 0-9749039-0-6.

Sachan, Mrinmaya, Contractor, Danish, Faruquie, Tanveer, and Subramaniam, Venkata. Probabilistic model for discovering topic based communities in social networks. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, CIKM '11, pp. 2349–2352, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063963.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1105–1112, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553515.

Yan, Feng, Xu, Ningyi, and Qi, Yuan. Parallel inference for latent Dirichlet allocation on graphics processing units. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 2134–2142. 2009.