Finding Patterns in Behavioral Observations by Automatically Labeling Forms of Wikiwork in Barnstars

David W. McDonald
The Information School
University of Washington

dwmc@uw.edu

Sara Javanmardi
School of Informatics
& Computer Science
University of California Irvine

sjavanma@ics.uci.edu

Mark Zachry
Human Centered Design &
Engineering
University of Washington

zachry@uw.edu

ABSTRACT

Our everyday observations about the behaviors of others around us shape how we decide to act or interact. In social media the ability to observe and interpret others' behavior is limited. This work describes one approach to leverage everyday behavioral observations to develop tools that could improve understanding and sense making capabilities of contributors, managers and researchers of social media systems. One example of behavioral observation is Wikipedia Barnstars. Barnstars are a type of award recognizing the activities of Wikipedia editors. We mine the entire English Wikipedia to extract barnstar observations. We develop a multi-label classifier based on a random forest technique to recognize and label distinct forms of observed and acknowledged activity. We evaluate the classifier through several means including use of separate training and testing datasets and the by application of the classifier to previously unlabeled data. We use the classifier to identify Wikipedia editors who have been observed with some predominant types of behavior and explore whether those patterns of behavior are evident and how observers seem to be making the observations. We discuss how these types of activity observations can be used to develop tools and potentially improve understanding and analysis in wikis and other online communities.

Categories and Subject Descriptors

H.5.3 Group and Organization Interfaces – Computer-supported cooperative work

General Terms

Design, Human Factors.

Keywords

Wikipedia, multi-label learning, behavioral patterns.

1. INTRODUCTION

> *"That person is driving too fast!"*
>
> *"Sally is a numbers person."*
>
> *"Travis is friendly and works well in groups."*

We often make casual observations about the behaviors and actions of those around us. The psychology or sociology of the everyday is one way that we interpret and navigate the varying social and behavioral circumstances that surround us. At times we are very good at this and other times we don't do so well. In some cases individual observations are sufficient, but in other cases we need many observations before we can understand and interpret another's behavior.

Faced with very large and growing online communities many people struggle to understand and interpret the behaviors of those around them so that they can act accordingly. As a result, participants in online communities often act with little, or at best, attenuated information. Further, tools to support better understanding of large behavioral datasets generated by participants in online communities are difficult to design and implement. Community managers, researchers, and social analysts are often faced with the need to hand code, hand label, or otherwise create subsets of expansive behavioral datasets for one-off analysis. Few tools currently support using the self-reflective nature of the community members to enable analysis.

This research begins to address how we can leverage community based behavioral observations; observations that people make of those around them. In a large online community with a wide range of behaviors, some participants will observe, interpret and label the behaviors of others in the community. A key question is whether those observations are reliable and whether those observations can be used in some reasonable way. In this research we explore one way to use behavioral observations that are made by participants in Wikipedia. We use barnstars, a community created mechanism for identifying and acknowledging activity of others, to develop a machine-learning tool. We apply our machine classifier to a large set of barnstars to explore whether barnstars can reveal patterns of user activity. By leveraging the observations of individuals within the community, we can begin to move beyond simple enumeration of a behavior (i.e., Bob has received 10 barnstars) and potentially characterize how others see behaviors (e.g., Bob likes to edit History articles and is considered helpful to others). Further, by considering repeated observations of behavior, it would then be possible to use reliably classified observations as a mechanism for sampling or selecting useful populations of individuals from the community for further study or analysis.

In the following we describe barnstars, how the barnstar dataset was collected and how we coded these barnstars as activity observations. We describe the development of a set of multi-label classifiers for our grounded activity codes and compare their performance with the average Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). We apply the best performing multi-label classifier, a random forest of 1000 trees, to explore whether possible patterns of behavior as labeled by the classifier exist in individuals' collections of received barnstar observations. Individuals from the community who receive a

predominant type of behavioral observation could be good candidates for further analysis or as users of other types of recommendation systems. We apply our classifier to previously unlabeled observations to identify editors who have been observed with some predominant types of behavior. We explore those patterns of behavior and how observers seem to be making the observations as one validation of the classifier performance. We conclude by outlining some prior work that frames activity observations made within Wikipedia and with a brief discussion of future work that could apply the classifier to other types of activity analysis.

2. BARNSTARS AS ACTIVITY OBSERVATIONS

In its simplest form, a barnstar is an image accompanied by a short personalized statement of appreciation for some work by another Wikipedia editor. Two example barnstars are shown in Figure 1. Wikipedia barnstars were invented for the purpose of allowing individuals to recognize the work of others.[1] Anyone can create, copy, customize and give out these tokens to any other editor.
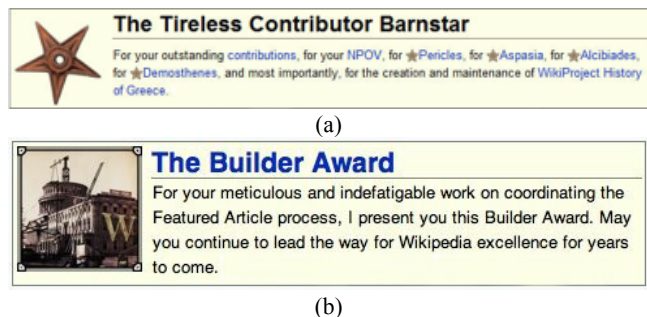
(a)

(b)

Figure 1. Example anonymized barnstars recognizing (a) adherence to NPOV policy, article work, and support of a Wikiproject community, and (b) leadership and participation in a formal process.

Givers typically post barnstars to the recipient's user or user talk page. Barnstars carry relatively high value to some recipients given their prominence on user pages. Some users move their barnstars to a "gallery" of achievements. While barnstars usually acknowledge some form of work, they can also serve to salve social slights, recognize overlooked work, encourage new editors, foster competition, or even to antagonize a recipient. Barnstars can be framed as behavioral observations because anyone can give a barnstar to anyone else for any reason and in most cases they are given for actions taken.

While we focus on Wikipedia barnstars, other communities have adopted barnstars as one form of user recognition.[2] Further, barnstars can be considered as a specific form of recognition similar to more widely used badges and achievement tokens. These are being adopted in a range of content communities where they serve to convey status and motivate further contribution and participation to the community.

2.1 Creating the Labeled Dataset

We created our initial labeled training set as part of an earlier study [8]. Here we reiterate important details of the creation of

that labeled training set and detail the creation of a testing data set.

We extracted barnstars from the November 2006 English Wikipedia database dump by creating a hand-tuned parser to generously identify candidate barnstars. The parsing extracted 14,573 barnstars given to 4880 unique users. A simple analysis of barnstar givers and receivers revealed that roughly one third of the population had only given barnstars. Another one third had only received barnstars. The last one third had both given and received at least one barnstar. This suggests that the use of barnstars does not constitute a closed community of mutual appreciation. If the community were largely one of mutual appreciation then the far majority of participants would be both givers and receivers of barnstars.

A codebook was developed through an initial open coding of a random sample of 200 barnstars. During an open coding process coders attempt to identify categories that cover as many items from the sample set as possible. In this stage categories are iteratively proliferated and condensed in an attempt to best fit the sample. The initial codebook was validated and refined based on a second random sample of 200 barnstars, exclusive of the prior 200 item sample. The attempt to systematically code the second random sample resulted in some refinements to the codes.

The codebook was then used to iteratively code a random sample of 2400 barnstars, excluding the prior 400 barnstars. The barnstars were divided randomly into six bins. Pairs of coders from our research group were then assigned to independently code the barnstars in each bin. After the initial independent coding, one coder reviewed the codes and noted all discrepancies.

The text of a barnstar often suggests multiple legitimate codes, either because a particular phrasing calls out multiple activites or because the barnstar contains multiple independent statements. We chose to apply multiple codes rather than force-fitting one dominant code. We use some examples below to illustrate this.[3] The first example is an award to an editor who edits from an IP address (an anonymous user). This barnstar acknowledges three distinct types of work, which all fall in a top level category called *Collaborative Actions and Disposition*. This barnstar acknowledges that the recipient took some action (CDA), by explaining what is acceptable to Wikipedia (CEXP), and keeping a cool head (CDD) in what could have resulted in a post-war.

> I award you this barnstar as recognition of the fine work you did in trying to explain to <an anonymous user> about why his posts on genetics aren't acceptable on Wikipedia - you managed to deal with what could have ended up as a huge post-war with a cool head. Thank you! {codes: CEXP; CDA; CDD}

In a second example two different top level categories are invoked; *Editing Work*, our category covering all things related to editing and *Social and Community Support* which covers a range of Wikipedia community support activities. In the Editing Work category, the barnstar acknowledges the creation of new articles (EINI) and that the editing contributions improve the encyclopedic content making a real difference (EMAJ). In the

1. See http://en.wikipedia.org/wiki/Wikipedia:Barnstar.

2. See http://www.wikihow.com/wikiHow:Barnstar

3. We have made a reasonable attempt to anonymize our examples. We have removed usernames and award dates. However, we recognize that the uniqueness of some text makes these awards easy to find.

Dimension of Observed Activity	Training Set		Test Set	
	Codes	%	Codes	%
Editing Work	852	28.8	180	29.1
Social and Community Support	763	24.9	150	24.2
Border Patrol	342	11.2	81	13.1
Administrative Actions and Formal Processes	284	9.3	54	8.7
Collaborative Actions and Disposition	244	8.0	41	6.6
Meta-Content Work	128	4.2	23	3.7
Undifferentiated Work	447	14.6	90	14.5

other category the participation in Wikipedia democratic processes is part of illustrating leadership in Wikipedia functions (SL).

> "The Working Man's Barnstar" For your many and varied efforts in creating and improving union, revolutionary and labour history related articles, as well as Hungarian and Australia-related articles and participation in wikipedia democratic processes, I take pleasure in awarding you the Working Man's Barnstar. Your contributions are well written and encyclopedic, making a real difference to Wikipedia. Regards {codes: EMAJ; EINI; SL}

Since multiple codes may be applied to any barnstar-awarding instance, any additional or missing code by any coder was considered a discrepancy. Coders iteratively discussed discrepancies until there was consensus for each pair of coders.

Of the 2400 multi-labeled barnstars, 274 (about 11%) were determined to be clear parsing errors and were removed from the set. The remaining 2,126 barnstars were used as the training set for all of the following experiments.

We constructed a test set by randomly selecting and coding 586 additional barnstars, excluding all previously coded barnstars. A pair of coders independently coded these barnstars. Code discrepancies where again noted and iteratively discussed until agreement was reached by the pair of coders. Of the 586 barnstars a total of 108 were determined to be parsing errors and were removed from the set (about 18%), yielding 478 coded barnstars for our test set.

2.2 Distribution of Observation Codes

The category scheme has seven top-level categories of activity. Six categories, *Editing Work*, *Social and Community Support*, *Border Patrol*, *Administrative Actions and Formal Processes*, *Collaborative Actions and Disposition*, and *Meta-Content Work*, recognize specific types of work acknowledged in barnstars. The seventh category, *Undifferentiated Work*, includes barnstars that clearly acknowledge work, but the specific type of work was unclear or undifferentiated. Table 1 lists the distribution of the codes for training and test sets.

Second level categories capture the detailed observations in barnstars. Our prior work [8] provides a further description of those categories. The examples above used some of the specific

codes when describing the barnstars. For the later machine classification problem, we decided that it would be too difficult to classify down to the second level categories because some of them have too few examples to be effectively learned. Therefore we focus on the top-level categories because they will be the targets for the classifier that we develop and describe in subsequent sections.

Editing Work (Training: 28.8%, Test: 29.1%). The category of editing work is the largest single category of observed activity. Editing work includes activities that are most commonly associated with creating and editing an encyclopedia. But this category also includes specialized multimedia content creation work that makes Wikipedia a rich end-user experience. This includes contributions of photos, diagrams, graphic design, and specialized audio. Activities in this category also include the work of applying templates and forms to pages. This includes use of category tags, the application of templates, such as notices of the page status or the insertion of informational boxes (info boxes). While editing is the largest single category across both training and test sets it is less than one-third of the activity observed and acknowledge by Wikipedians.

Social and Community Support (Training: 24.9%, Test: 24.2%). The second most common type of observed activity is the work necessary to support members and keep the community functioning. This includes welcoming newcomers, initiating or leading new projects, rewarding individuals who give out barnstars, and general social support. This category acknowledges the activities of Wikipedians who create and sustain Wikiprojects and efforts to lead others through some set of needed tasks. The leadership activities in this category are distinct from the work of Wikipedia admins and the admin activities supported through the platform admin tools.

Border Patrol (Training: 11.2%, Test: 13.1%). Border patrol includes a range of activities to manage and control vandalism. Vandalism not only happens on main article pages, but in user, category and template pages. Vandalism can be rather subtle such as slightly rewording text to make it inaccurate, or blatant such as inserting advertising messages. This category also recognizes the activities to maintain a sense of significance in what is covered. Wikipedia relies on the concept of "notability" for deciding whether a topic should be covered. Border patrol includes the activity to identify and remove non-notable pages. Lastly, Wikipedia must work to remain clean from copyright violations. The activities to identify and remove copyright violations, whether intentional or not, is an important border patrol activity.

Administrative Actions and Formal Processes (Training: 9.3%, Test: 8.7). This category includes the activities of Wikipedia admins, but also participation in formal decision-making processes by regular users. Examples of formal processes that fall in this category include Editor Review, Featured Article and Good Article review, Request for Adminship review, Request for Comment, and Request for Arbitration. These are important non-exclusive activities in which all members of the community can participate. This category also includes a range of activity specific to administrators such as the ability to Check User, or privilege granting such as making a regular user into an admin.

Collaborative Actions and Disposition (Training: 8.0%, Test 6.6%). Collaborative Actions and Disposition is differentiated from Social and Community Support Actions by the direct implication of collaborative activity, such as conflict mediation on talk pages. Some aspects of the collaborative activity may be

Table 2. Examples of ngram features for the seven activity categories

Dimension of Activity	ngram Features
Administrative	admin, sysop, new mop, username block, supervision, mediation, my rfa
Border Patrol	revert, vfd, rfcu, copyright, wp:cp, patrol
Collaborative Action	consensus, survey, humor, reconciliation, rationality, npov, [[wp:ppol	policy]]
Editing	make this article, numerous, contributions, typo, minor edit, categorization, wikifying, restructuring, reference
Meta-Content	tagging speedy deletes, infobox, logo, article assessment, format, css class, user categories
Social and Community	commitment, persistence, up to date, for founding, esperanza
Undifferentiated	anniversary, cake, birthday, promotion, count of

Table 3. Area Under the Curve for the independent binary classifications. Bold indicates best performing technique for a given category.

Dimension of Activity	Logistic Regression	Naïve Bayes	Random Forest (1K trees)	KNN (k=10)
Administrative	0.833	**0.949**	0.942	0.903
Border Patrol	0.922	0.941	0.952	**0.956**
Collaborative Action	**0.750**	0.722	0.743	0.725
Editing	0.878	0.875	0.879	**0.884**
Meta-Content	0.835	0.842	**0.883**	0.800
Social and Community	0.802	0.796	0.797	**0.805**
Undifferentiated	0.847	0.848	0.844	**0.854**
Avg. (AUC)	0.838	0.853	**0.862**	0.847

implied. For example, the giver of the barnstar may mention how the recipient maintained a "cool head" when dealing with an unnamed person who was being difficult. Observations in this category include activities to facilitate adherence to Wikipedia policies, often through a careful dialog or explanation.

Meta-Content Work (Training: 4.2%, Test: 3.7%). Meta content work includes the observations of the work necessary to develop tools, create templates, create and manage category schemes, and contribute to the work of clarifying or creating formal processes. Meta work tends to be rare because it can require specialized skills. Meta work is a form of articulation work [11] which tends to be a rarer form of activity.

Undifferentiated Work (Training: 14.6%, Test: 14.5%). Some observations are just not very specific about the activity. There are a number of generic statements that thank a person for their "help" or "work" on a specific article. While editing is very common there are many activities that relate to articles that may not explicitly fall in to the editing category. Further, there are slang phrases in the Wikipedia community that point to generic work, such as references to "Janitorial Services" and "mop and bucket," reflecting a general ethic of cleaning and maintaining various aspects of Wikipedia. Because these observations are not detailed they are placed in this category.

While coding and describing the types of observed activity is important, our goal is not just to understand what types of activity are performed. We would like to be able to use these everyday observations for other purposes, such as selecting a set of users for further analysis or for making recommendations to a specific user about work that might match their interests. Hand coding activity observations (like barnstars) just would not scale to a growing community. Our initial parse of barnstars yielded over 14,000, and our labor-intensive hand coding has at most covered 3300.

A tool that could read barnstars and generate a reliable multi-label classification would simplify using these activity observations for other purposes. In the next section we describe our approach to creating a tool that can take observations, like barnstars, and label the observations.

3. LEARNING TO RECOGNIZE AND LABEL OBSERVED ACTIVITIES

A machine learning approach to this problem requires three things: (a) a model that maps each barnstar observation to a vector of features; (b) a classifier that maps these feature vectors onto {0, 1}, and (c) a corpus of pre-classified barnstars, where 1 denotes the barnstar belongs to a category and 0 denotes that it does not.

The feature set was constructed by learning popular ngrams (n<=4) from the text of the barnstar observations for each activity code category. We also extracted ngrams from the image file names that often accompany a barnstar observation. For example, "barnstar_of_reversion2.png" is a common image name for barnstars with Border Patrol related observations and "wmbarnstar.png" is a common image for barnstars in the Editing Work category.

In total, we extracted 585 ngram features from the training set for the seven categories. Table 2 shows some common ngrams features for each activity category. For example, "support me for adminship" or "new mop" are features for detecting barnstars in Administrative Actions and Formal Processes category.

We combine these raw ngram features into 55 aggregate features based on the likelihood of seeing them in each of the subcategories that comprise the top-level activity categories. These aggregate features are used for the current multi-label classification.

Existing methods for multi-label classification follow two approaches: *problem transformation* (PT) methods, and *algorithm adaptation* methods. Problem transformation methods transform a multi-label classification problem into several single-label classification problems. Adaptation methods extend a specific learning algorithm in order to handle multi-label data. In doing so adaptation methods change the definition of the loss function to account for possible correlations between different labels [14]. In the following we test both PT and algorithm adaptation approaches on our data.

Table 4. Area Under the Curve for PT4 transformation Classification with Mulan and MLkNN. Bold indicates best performing technique for a given category.

Dimension of Activity	Logistic Regression	Naïve Bayes	Random Forest	MLkNN (k=10)
Administrative	0.755	**0.913**	0.843	0.806
Border Patrol	0.878	0.901	0.830	**0.903**
Collaborative Action	0.724	**0.726**	0.710	0.704
Editing	0.822	0.804	0.813	**0.824**
Meta-Content	0.661	**0.765**	0.727	0.685
Social and Community	0.731	0.740	**0.777**	0.747
Undifferentiated	0.771	**0.769**	0.767	**0.769**
Avg. (AUC)	0.763	**0.802**	0.781	0.777

For our first attempt we transformed our multi-label classification problem into seven independent binary classifications, one for each top-level category. This method is known as PT1 problem transformation. We then used the training set to learn a classification model, and the test set to measure its performance. Since both training and test sets are very imbalanced with respect to the activity codes, common performance metrics such as accuracy are not useful measures. Instead we use receiver operating characteristic (ROC) as our primary measure of classification performance as reported by the Weka data mining package [6]. Table 3 shows the classification performance for the seven categories based on different classifiers.

While none of the classifiers outperforms the others across all seven categories, random forest based on 1000 trees results in the best classification performance with respect to the average AUC (Area Under Curve) values.

One interesting observation is that all of the classifiers do well for some categories, such as Border Patrol or Administrative, but not so well for other categories, such as Collaborative Actions and Disposition or Social and Community Support. As we pointed out above, in these categories there are implied actions or traits that are not always explicitly mentioned in the barnstar observation. Thus users and classifiers must infer some of what is not explicitly mentioned. This is a difficult thing to do accurately.

In other work, we explored how different features of barnstars, like specific title text, use of policy references or article references, influence the accuracy of classification [10]. Features such as Wikipedia policies, guidelines, templates, and links are valuable for improving performance. For example, an explicit mention of conduct policies such as Wikipedia:Civility could be a signal for an observation in Social and Community Support. Similarly, mentions of dispute resolution policies help detect observations in Collaborative Action and Disposition.

In a second approach, we used multi-label classification methods from the Mulan Java package [13]. We tried several other problem transformation methods such as PT3 and PT4, in conjunction with different classifiers. PT4 is the most common PT

method. This PT learns |L| binary classifiers Hl: X → {l, ¬l}, one for each different label l in L [14]. The original dataset is transformed into |L| datasets Dl that contain all examples of the original data set, labeled as l if the original labels contained l and as ¬l otherwise. It is the same solution for a single-label multi-class problem using a binary classifier.

We also tried several algorithm adaptation methods such as MLkNN, MMPLearner, and IBLR_ML [3, 18]. The MLkNN adaptation resulted in the best classification results. MLkNN uses the KNN algorithm independently for each label. It finds the k nearest examples to the test instance and considers those that are labeled at least with l as positive and the rest as negative. What mainly differentiates this method from the application of the original KNN algorithm to the transformed problem using PT4 is the use of prior probabilities. Table 4 shows the AUC values of MLkNN (k=10) for the seven categories.

Comparing Table 3 and Table 4 shows that independent binary classifiers outperform their multi-label versions. Label cardinality and label density measure how heavily multi-labeled the data set is [14]. Label cardinality of D is the average number of labels of the observations in D:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|$$

Label density of D is the average number of labels of the observations in D divided by |L|.

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|}$$

Where D is the number of observations in the dataset, |Yi| is the total number of labels for observation i, and |L| is the total number of labels in the dataset.

In our case LC (training set) and LD (training set) are 1.25 and 0.178, respectively. The minimum possible value of LC is 1 and the maximum is 7; and these values are 1 and 0.14 for LD. The low values of label cardinality and label density show that the data is not heavily multi-labeled. As a result the relations between categories would not be significant. But low cardinality will not guarantee that independent binary classifiers always outperform relational classification techniques. There may be some features that allow the classifier to learn correlations between some classes better. This would provide a boost to performance in those cases where multiple labels are present.

In our case the low LC and LD simply suggests why performance by the independent classification approach is higher compared to the relational classification approach. The relational classification approach is not gaining benefit from any correlated labeling across barnstars and, in our case, it looks like the lack of multi-label correlations has associated costs in terms of labeling performance. Parameterizing those costs is an open problem for multi-label learning that we will not address here.

Recall, with the problem we outlined, having an automatic multi-labeling tool is only part of a solution. The goal is to use that automatic labeling for other objectives, such as identifying an interesting set of individuals from the community for further analysis or study. But before we can do that we need to explore whether there are possible patterns of observations being made by participants in the community about the activities of other members.

Table 5. Number of users with potentially predominant behavior patterns in each category. Average percentage of barnstar observations for candidates in each category

Dimension of Observed Activity	Label	Avg. %	Cand.
Editing	E	67.9	25
Border Patrol	B	73.2	13
Social and Community	S	61.5	54
Administrative	A	66.4	75
Collaborative Actions	C	52.0	1
Meta-Content	M	76.8	4
Undifferentiated	U	60.0	10
			182

4. VERIFYING OBSERVATIONAL BEHAVIOR PATTERNS

We explored whether automatic labeling of barnstars could reveal patterns of activity observations. Starting with the complete set of barnstars (all 14,573), we selected recipients who had 9 or more barnstars. This cutoff is somewhat arbitrary, but provided a reasonable number of candidates for this exploration (259 recipients out of 4880, receiving 4327 barnstars). We applied our multi-label classifier to all barnstars received by individuals in this subset. Keep in mind that these are largely barnstar observations that have not been previously seen and coded by hand. The primary selection for this sample was recipient, where the recipient had a minimum of 9 barnstars. All prior samples were random selections of barnstars ignoring both the recipient and the individual making the barnstar observation. Another way to state the distinction is that prior analyses were observation or barnstar centric and this analysis is recipient centric.

Applying the multi-label classifier to all barnstars received by individuals in this subset, if the same label was applied to more than half of the barnstars for a given recipient then that recipient is considered to have a predominate observed behavior in that category. Table 5 provides the number of recipients who both had more than 9 barnstar observations and where more than half were labeled in the indicated category of observed behavior.

Finding that a recipient has a predominate number of barnstar observations in a given activity class is a first step toward understanding whether that individual exhibits an actual activity pattern that falls within the identified class of activity. However, we are basing this on human observations and they may be biased. For example, the barnstar recipient may have performed one really important action and observers are noting, though barnstars, that they recognize the significance of that one act. In that case, the recipient would have many observations that are automatically labeled as the same class of activity, but their actual behaviors might not match. That is, the observation and the subsequent barnstars may not be independent.

We took a random sample of 39 (21.4%) recipients from the 182 that our automatic labeler found with predominately observed activities. We reviewed the text and the classification of all barnstar observations for each of these recipients in the random sample. By reviewing the observations we hoped to see whether the pattern of observed behaviors as labeled by the automatic classifier reasonably aligned with the pattern present in the collective observations of the individuals awarding the barnstars. The random sample of 39 recipients resulted in 544 barnstars. Our review identified 65 barnstars that were not barnstars (11.9%). This rate is commensurate with that identified in the original

coding of the training set [11] and below the rate identified in the random sample selected to generate the test set (described above). We calculated the label cardinality and label density for the 544 labeled barnstars at 1.74 and 0.249 respectively. Both label cardinality and density are slightly higher in the automatically labeled set than in the training set.

4.1 Independence of Observations

One potential bias of barnstar observations is possible duplication of observation. For example, if a user has 10 barnstars and 6 of them are for editing, we would like to know whether those were all given for the exact same editing event or if they were independent observations for different editing events. Simply relying on the date of the barnstar award is not sufficient because the observers might be seeing the exact same significant event, but observing and awarding barnstars at different times.

Through our review we found **no** evidence that different observers were awarding one individual multiple barnstars for the exact same observed event. For example when seeing multiple barnstars all labeled for Editing Work, frequently the awardees mentioned different articles or different general topics. There are cases where Social and Community Support commitments to the same group or WikiProject are mentioned, but they often included different phrasing of what was being observed, such as a leadership trait or a commitment to expanding articles related to the project.

We illustrate this with a small sample of four barnstars from a user (u3818) whose predominate behavior was categorized as "S" (Social and Community Support) with 9 out of 17 observations labeled as mentioning "S" type activities. In each case the recipient is acknowledged for a different type of social support to either an individual or to the Wikipedian community more widely.

> "the surreal barnstar" for putting the unblock reason "f@ck you unblock me now" actually into the {{tl|unblock reviewed}} template, i award you this barnstar. i lol'ed. {labels: *A*, M, S}

> "the da vinci barnstar" thank you for your assistance at [[wp:ani]] and at [[wt:srnc]]. it takes a bold move for someone to finally put their foot down at a user like <username> in an effort to reach a consensus, keep editors from bailing out of wikipedia, and to bring more civility and peace to this whole naming mess. kudos and thank you for your service! {labels: C, *M*, S}

> "the random acts of kindness barnstar" to <recipient>, for being a kind administrator, for taking the time to assist other editors, and for doing the little things that make others "time in wikipedia super" (to quote foresth2). {labels: A, S}

> "the defender of the wiki barnstar" as a supervising admin, <recipient> helped resolve one of the most contentious, mind-numbing, extensive disputes in wikipedia history over the naming of state highways. for this commendable and exhausting work, he deserves the eternal gratitude and wiki-love of every wikipedian. {labels: A, *B*, C, S}

It is worth mentioning some other aspects of these examples. Like the examples in Section 2.1 we have made a reasonable attempt at anonymizing the text of the barnstar observation. Similar to our examples in Section 2.1, we have included the "labels" that were derived by our automatic classifier. We distinguish "labels" generated by the classifier from "codes" that were hand applied

based on our codebook. All four examples above have been labeled with two or more types of activity. But not all of the labels would be correct; we know that from the Average AUC performance reported above. In cases where our validation review found a label to be incorrect, we have listed it, but indicated such by underlining the code.

However, barnstar duplication was present in the set. When making a barnstar observation some people save intermediary versions of the award. In Wikipedia the revision logs will contain all intermediary versions. Our parser did not detect this, and neither did a random selection of barnstars that were coded to generate train and test sets. However, by considering all barnstars awarded to a specific individual this duplication could be observed. We found 88 duplicates (16.1%) in our subset with 27 of the 39 candidates having at least one duplicate.

The effect of duplicates is uneven across the recipients. Ten of the candidates only had only one or two duplicates. Seven of the candidates had five or more duplicates, accounting for approximately half of all duplicates (43 of 88). That is, a small number of candidate users account for a disproportionate share of these duplicates. The user with the largest number of duplicates (10) also had the largest total number of barnstar awards in the sample (38). For this user (u1025) the duplicates did not impact the pattern of observed behaviors. At the other end of the spectrum, two candidates having five duplicates had been awarded 9 and 10 barnstars respectively, meaning half or more of their barnstars were duplicates (u147924, u4695). In the case of these users, the remaining observations were too few to reveal any meaningful pattern.

4.2 Activity Patterns in Barnstar Observations

Activity patterns were visible for many users in our sample based on the barnstars they received. We consider several examples to illustrate patterns identified through the automatic labeling. Above we noted that there were some duplicates that resulted from our parsing. Since there are ways to remove a significant number of duplicates, in the following we report the results in absence of duplicates. Recall our goal in this validation is to see whether there are patterns present in an individual's collection of barnstar observations as a function of the automatic labeling. In the following we present six individuals who have the most observations in our set because they represent a spectrum of what was discovered.

Candidate u1025 received the largest number of unique barnstar observations in the sample set (25). Of those, 19 were labeled as "A" type observations (Administrative Actions and Formal Processes) and our pattern identification criteria placed this user in the "A" category. A "conservative" hand validation of the labeling identified that 10 of the "A" labels are potentially incorrect[4] - resulting in 9 "A" labels out of the 25 observations. While the corrected labeling would not have passed our 50%

threshold, looking more deeply at the observations for this user revealed another key characteristic. The 9 accurate "A" labels were solo labels; meaning that only the "A" label was applied. Further, 9 other observations made of this user were labeled "M" with only one of those having any other label assigned. That is, this user has strong bi-modal activities in "A" and "M", but would not have crossed our simple 50% threshold in either case.

Candidate u1382 received the second highest number of unique barnstar observations in our sample (23). The automatic labeler and our pattern criteria classified this user as having a Border Patrol pattern with 21 labeled "B". Reviewing the barnstars for this user revealed that they were indeed solid observations of border patrol, but also revealed something else. It turns out that this candidate is not actually a person, but an anti-vandalism Bot to which people had awarded barnstars. The remaining two barnstar observations were acknowledgement for the bot's creator.

The next two candidates ordered by most observations were candidate u1232 having 9 of 22 labeled with "S" (Social and Community Support), and u3818, having 9 of 17 labeled with "A" respectively. In the case of u1232 the predominate activity would not have crossed our 50% threshold after our 'conservative' validation. Considering the other activities for u1232, the next most frequent observed activity was "B" for Border Patrol in 5 of the 22 barnstars.

Examining the barnstar observations for u3818 the classifier labeled 9 of these as having "A" characteristics. Our quick hand validation revealed only one of these as incorrect. But our hand validation also revealed that for this candidate there were 9 observations labeled "S" which were all accurate. For candidate u3818 there were two patterns nearly as strong, and both consisting of more than half the total barnstar observations. Sample barnstars from this user were highlighted in the prior subsection.

The fifth and sixth users both had 15 observations. User u5699 had 13 observations labeled "A" for Administrative Actions and Formal Processes. Considering these observations closely, more than half are for participation in Request for Adminiship review (RfA). The others are for a range of administrative actions relating to a range of discussions relating to decision making activities (e.g., wp:cfd, wp:drv, wp:mfd). User u1211 had 9 of 15 observations labeled "S"—most of them for noticing and welcoming new users.

4.3 Patterns in Collective Observation

Taking a step back, our big picture question for this small validation was whether by applying our machine classifier, we could see a pattern of behavior in one individual similar to the collective observations present in barnstars for that same individual. Through our exploration and "conservative" hand validation we have found that, in our sample, few observations are being made of the same person for the exact same behavioral event. This suggests that individuals making behavior observations through barnstars are doing so somewhat independently. This improves the likelihood that techniques such as ours can be used to characterize patterns of behavior for an individual based on observations by other members in the community.

Our current approach relies on a somewhat arbitrary cutoff requiring the recipient to have received 9 or more barnstar observations and with more than 50% of those barnstars being labeled in one activity. This simple threshold of requiring more

---

4. By "conservative" we mean we biased to deciding against the automatic labeling if the barnstar was not clear. This seems reasonable since label density and cardinality are high for the automatic labeling. This has a side effect of making the labeling seem somewhat less accurate than what one would expect when considering Average AUC. Additionally, it is reasonable to expect somewhat lower real-world performance relative to lab validation for this type of application.

than half of observations to fall in a single activity category found a rough cut of individuals who might have observed patterns of activity. However, this threshold seems too strict in a real system. For example we found cases where individuals had strong bi-modal activities. In those cases there is not one single activity observed and labeled that crosses the 50% threshold, but two activities come very close. This suggests that a more dynamic threshold might be useful. For example, one could use a threshold related to the label density of each individual user. If the label density is higher, then one could require a higher number of observations be labeled with a specific type of activity before that activity would be considered a possible predominate activity. This approach could allow for the identification of more than one possible predominate activity for a given user.

Our approach is not without some caveats. Observational datasets like the one we have leveraged will likely be heavily skewed in the numbers and types of observations that people make. In the case of the work activity in barnstars this has some roots in the way work is organized [8, 11]. This skew can be a challenge for machine classification techniques. Further we only focused on the observations in barnstars and have not pushed this all the way back to the actual set of activity traces in the Wikipedia dataset.

The next step in validating patterns of behavior would be to compare the pattern of observed behaviors to the edit history for the set of individuals we have identified as likely to have a specific type of activity trace pattern. But before that is practical we would need to do a hand validation, like we have described here for a larger sample of our identified candidates (see Table 5); the current random sample of 39 people is too small to be useful because of noise and variance present in most user behavior. Validating a larger number of people from our candidate set and attempting to link their predominately observed behaviors to their actual activity trace is ongoing work.

5. RELATED WORK

Prior work on leveraging observational data from general participants in an online community is limited. In most cases making observations and using them empirically or analytically has been the domain of social scientists. But the prior literature does include ways that observations and observational data have been used to understand behavior in Wikipedia.

Early studies of Wikipedia examined article editing as an activity contributed by participants. Bryant et al. [2] noted some less obvious activities and that learning how to participate in those activities was important to becoming a member of the Wikipedia community. This means that individuals must observe in order to participate. Our own prior work specifically considered the types of observations that community members make of other members [8]. We used barnstars to describe the various dimensions of work observed and acknowledged by Wikipedians and showed how some types of work are rare and less well observed.

In more recent work, Geiger & Ribes [5] studied the decision making process for banning a vandal. Their study raises some nice issues about the work to remove vandalism and how that is performed, illustrating how Wikipedians consider the activities of others. In the case of the decision to ban a vandal, the effort is to understand whether the activities are intentionally designed to corrupt content or wreak havoc on the community itself. This is important because it requires a participant in the community to observe an activity and gauge some level of intent - which is an interpretation of the observed activity.

Work by Luther et al. [9] focused on success factors in an online creative community. This work explores success factors in the creation of collaborative videos. These include the planning and structure of the creative project, the reputation and experience of the project leader and the communications among the individual contributors. While the researchers use individuals' comments and insights to build their analytical framework, the community they studied does not have a mechanism for the participants to arbitrarily note their observations of other contributors. This research approach is common in the study of online community. Our approach differs distinctly from this, in that the patterns of activity we uncover are based on community members observing and noting the behaviors of other members of the community. Both approaches generate grounded results, but just change whose observations become the basis for generating patterns; those of the community or those of the researcher.

Except for these studies, the results from prior literature do not focus on the activity observations and resulting interpretations that community members make about others.

More often the prior research has focused on a specific technique or one specific form of work. For example, Burke & Kraut [1] studied the Request for Adminship (RfA) process. In this process Wikipedians consider the activities and work contributions of a person desiring to be promoted to administrator status. They found that some criteria articulated as important to the decision do not weigh heavily in a statistical regression model of RfA promotion. In this case, the way Wikipedians observed and interpreted behaviors of others was not an explicit part of their model. Instead the relevant variables in the regression model were aspects of edit history that could be mined or counted. These could then be factored relative to the success of each person who has undergone an RfA review. While the model is important, we don't really know if individuals who consider others for RfA actually use the factors in the model or if those factors are tightly correlated to other observations that are being made.

Krieger, et al. [7] focused on the design of a tool to support the meta-work of helping community members understand what work is in need of attention. The tool took the form of a task or 'to-do' list. While this involved a categorization of the types of activity to facilitate structuring a task list, the focus was not on how an individual understands and interprets the activities of another in the community, such as understanding what previous people have contributed and what activities are still necessary to complete some collective project. The focus instead is on the development of the task list tool.

In a study of editor leadership in the work of Wikipedia projects, Ung & Dalle [15] characterize differentiate editor behavior based on time-based patterns of editing activity. They observe different activity patterns associated with coordination of identified groups versus coordinating the contributions of individual, generalist editors. Their findings are based solely on the analysis of coordination activities rather than accounting for the perspectives of those who are coordinated or are beneficiaries of the coordinated effort.

Another approach to understanding activities in Wikipedia relies on visualizations. Viegas et al. [16] focused on editing work, visualizing how article content changes over time as a result of individual editing activity. Suh, et al. [12] developed Wikidashboard which focused on editing and visualizing patterns of editing. These are examples of a genre where article editing visualizations help users understand or infer trust in the content.

But neither of these focused in on how users in the community observe and interpret the editing activity of another user. The tools focus on displaying editing frequency.

Also using a visualization approach, but taking a slightly broader view of activity, Wattenberg et al. [17] motivated their visualization with the idea that patterns of activity might be visualized. Using data of Wikipedia administrators they created 'chromograms.' Through the visualizations they found heterogeneity in the administrators' activities. They classified activity into two categories: systemic tasks (e.g. list-based tasks like sorting stubs) and reactive tasks (e.g. watching for vandalism or welcoming new users). The visualizations are compelling, but do not provide an explanation of the activity.

A slightly different view of understanding activity is present in the work by Cosley et al. [4]. The SuggestBot system builds a profile for each user based on the articles that they have edited. The user's edit history can be used for a content based similarity, page link similarity or co-edit similarity with another user to make recommendations about what tasks might be done next. This approach is largely focused on the activities related to article editing. This work does not really consider how one user understands the contributions of another, but it clearly is attempting to account for a pattern of user activity.

In general, the prior findings are not focused on how individuals observe or how they interpret the activities of others in the community. This is not meant to be a criticism of the prior work because systems like Wikipedia are generating truly massive archives of activity data that grow daily, and the research community is still coming to grips with how to effectively mine and utilize this type of massive activity trace. Indeed, a careful reading of these papers illustrates that Wikipedians are observing the behaviors of others and are acting upon what they see or what they come to believe about others' activity. But any one person can only observe a small portion of this massive activity dataset, which is why there is a need for the type of tool we are attempting to develop.

6. CONCLUSION & FUTURE WORK

People naturally see different things in the people they observe. A person's activity is going to be messy, contingent, and open to interpretation–that's just how we are as people. The development of wikis that facilitate wide ranging scales of collaborative activity invite varied forms of participation. Understanding these forms of participation is important to the design of new systems and for understanding both successes and failures for open collaboration.

In our approach we use one type of activity observation made by participants in a community about other participants in the community. These observations are similar to observations that researchers might make if they were conducting an observational study of activity. Researchers would attempt to see something of the breadth of activity, and understand if there were prevalent patterns. The trade off here being that while the researchers might look at a smaller sample of participants and triangulate, these types of participant-created observations are more numerous and come from a more diverse population of observers. A high level characterization of our approach is that we rely on a form of citizen science that could be called Citizen Social-Behavioral Science.

We took behavioral observations in the form of barnstars and a previous grounded coding as a way of recognizing the activities that individuals in the community valued. In short, it seems unlikely that individuals would bother spending their time developing awards or tokens for behaviors that they did not value in some way. And the community has other mechanisms for dealing with undesirable activities, which could be a separate focus for another project. We applied machine learning techniques to create a multi-label classifier that is reasonably accurate at labeling barnstar observations with possibly multiple types of activity.

The goal of this effort was to see whether behavioral observations made by the community could reveal some patterns or predominate activities of others in the community. With the machine labeling, we conducted a modest hand validation that revealed two key things. First, that few individuals are receiving barnstars for the exact same action. This suggests that the individuals observing behavior and awarding barnstars are doing so somewhat independently. This makes a collection of barnstars awarded to one individual potentially more valuable as a characterization of what that individual contributes to the community. Second, the validation was able to uncover patterns in barnstar observations. That is, it is possible to find individuals who have been observed and awarded for doing the same types of activities on numerous occasions.

The value of our approach does not stop at simply recognizing a pattern for a single individual. The ability to rely on participant based observations of social activity could be more useful when identifying a subset of individuals for further analysis. Many studies of wikis focus on basic editing activity because there are few ways to get a handle on more complex social behaviors. A more general application of our approach could be used to identify sets of users who informally mediate social disputes, who are effective at welcoming or mentoring other users, who are good at explaining the community norms and policies - all of which are important activities but which are currently very hard to see through what is mostly an edit count. Community managers, researchers, and social analysts, faced with expansive behavioral datasets could use an approach like the one we have described to analyze and understand a wider range of community defined activities. In this way they can move beyond simple frequency counts of an activity to a community based interpretation of observed activity.

Future work follows two distinct threads. In the first thread, we plan to explore ways to improve the multi-label classification. The current classifier uses a small hand-tuned feature set. We plan to explore a wider range of features and mechanisms for automatic or semi-automatic selection of those features. This could further generalize our approach to classification of behavioral observations. Further, we plan to improve our current pattern identification scheme by implementing a dynamic threshold mechanism that would account for the significance of a given labeling based on the label density for each individual.

In a second thread we would like to take pattern identification further. That is, given a set of users who are identified as having a particular predominate activity, say Social and Community Support, we would like to go back to those individuals' specific edit histories and see if we can identify significant sets of edits that illustrate such patterns. This seem possible for activity categories like Border Patrol or aspects of Editing work, but is a bit more challenging when attempting to identify patterns that are more social, collaborative or which might be in a range of meta work. The range and diversity of behavior for any one individual

will introduce some amount of noise in a wiki activity trace that could obscure some important patterns.

Our everyday observations about the behaviors of others around us shape how we decide to act or interact. In large social media systems our ability to observe and interpret others' behavior is limited. Our work lays an important foundation for the development of tools that can both help individuals understand the others around them, in a social media system, and help researchers, analysts or community managers further assess the broader patterns of activities in their social media systems.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Burke, M. and Kraut, B. 2008. Mopping up: modeling wikipedia promotion decisions. *Proceedings of CSCW'08*, 27-36.

[2] Bryant, S. L., Forte, A. and Bruckman. A. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. *Proceedings of GROUP'05*. 1-10.

[3] Cheng, W. and Hullermeier, E. 2009. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*. 76 (2-3):211-225.

[4] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. 2007. SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. *Proceedings of IUI*.

[5] Geiger, R. S., and Ribes, D. 2010 The work of Sustaining Order in Wikipedia: The Banning of a Vandal. *Proceedings of CSCW'10*. 117-126.

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11 (1).

[7] Krieger, M., Stark, E. M., and Klemmer, S. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. *Proceedings of CHI'09*. 1485-1494.

[8] Kriplean, T., Beschastnikh I., and McDonald, D. W. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. *Proceedings of CSCW'08*. 47-56.

[9] Luther, K., Caine, K., Ziegler, K., and Bruckman, A. 2010. Why It Works (When It Works): Success Factors in Online Creative Collaboration. *Proceedings of GROUP'10*. 1-10.

[10] Sajnani, H., Javanmardi, S., McDonald, D. W., Lopes, C. V. 2011. Multi-Label Classification of Short Text: A Study on Wikipedia Barnstars. In the AAAI-11 Workshop on Analyzing Microtext.

[11] Strauss, A. 1985, Work and the Division of Labor. *The Sociological Quarterly*. 26, 1. 1-19.

[12] Suh, B., Chi, E., Kittur, A., and Pendleton, B. 2008. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. *Proceeding of CHI'08*. 1037-1040.

[13] Tsoumakas, G., Vilcek, J., Spyromitros, E., Vlahavas, I. (2010) Mulan: A Java Library for Multi-Label Learning. *Journal of Machine Learning Research* 1, 1-48.

[14] Tsoumakas, G. and Katakis, I. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3 (3) 1-13, 2007.

[15] Ung, H. and Dalle. J. 2010. Project management in the Wikipedia community. *Proceedings of the 6th International Symposium on Wikis and Open Collaboration (WikiSym '10)*.

[16] Viegas, F. B., Wattenberg, M., and Dave, K. 2004. Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of CHI'04*.

[17] Wattenberg, M., Viégas, F. B., and Hollenbach, K. 2007. Visualizing Activity on Wikipedia with Chromograms. *Proceedings of INTERACT 2007*. LNCS 4663, Part II. 272-287.

[18] Zhang, M. L. and Zhou, Z. H. 2005. A k-Nearest Neighbor Based Algorithm for Multi-label Classification. *Proceedings of the 1st IEEE International Conference on Granular Computing*.