FINDING SOCIAL ROLES IN WIKIPEDIA

Howard T. Welser
Ohio University
welser@ohio.edu

Dan Cosley
Cornell University
danco@cs.cornell.edu

Gueorgi Kossinets
Cornell University
gkossinets@gmail.com

Austin Lin
Cornell University; Microsoft Inc
Austin.lin@gmail.com

Fedor Dokshin
Cornell University
fad7@cornell.edu

Geri Gay
Cornell University
gkg1@cornell.edu

Marc Smith
Connected Action
marc@connectedaction.net

ABSTRACT
This paper investigates some of the social roles people play in the online community of Wikipedia. We start from qualitative comments posted on community oriented pages, wiki project memberships, and user talk pages in order to identify a sample of editors who represent four key roles: substantive experts, technical editors, vandal fighters, and social networkers. Patterns in edit histories and egocentric network visualizations suggest potential "structural signatures" that could be used as quantitative indicators of role adoption. Using simple metrics based on edit histories we compare two samples of Wikipedians: a collection of long term dedicated editors, and a cohort of editors from a one month window of new arrivals. According to these metrics, we find that the proportions of editor types in the new cohort are similar those observed in the sample of dedicated contributors. The number of new editors playing helpful roles in a single month's cohort nearly equal the number found in the dedicated sample. This suggests that informal socialization has the potential provide sufficient role related labor despite growth and change in Wikipedia. These results are preliminary, and we describe several ways that the method can be improved, including the expansion and refinement of role signatures and identification of other important social roles.

General Terms
Human factors, theory

Keywords
Social roles, Wikipedia, structural signatures, social networks, online community

1. INTRODUCTION
Wikipedia has forever changed how we use, find and think about information. Both pundits like Stephen Colbert and researchers [23], [13] have been pre-occupied with the question of whether Wikipedia is of sufficient quality and whether its pages constitute legitimate references [23]--which some people argue will never be the case as long as it relies on non-expert volunteers of unknown identity [6]. In this paper, instead of prognosticating about the potential of the Wikipedia project, we focus on understanding how Wikipedia has achieved the success that it has, as a reasonably good resource for many topics that is often the first link suggested in search engine results. How has the "pretty good" and incredibly extensive resource been achieved? And how has this been possible given the absence of the resources and controls of conventional firms and bureaucracies [27]?

Large-scale, distributed, collaborative projects like Wikipedia are changing how we think about the nature of work. Research suggests that success of Wikipedia has stemmed from three key sources: infrastructural and social features that help people find and define their roles in the organization [4], [18], technical innovations that allow substantial economies of scale in the performance of many of those roles [8], and social mechanisms that support coordination and conflict resolution [26]. This paper concentrates on problem of finding and defining roles in a large, distributed organization and looks to identify the informal roles roles that affect the quality and coordination of participants' contributions.

Following Gleave et al. [12], we use qualitative methods to identify an initial set of potential roles, and identify potential quantitative signatures of those roles [7], [29]. Although roles are continually evolving and being recognized in Wikipedia [18], our qualitative analysis highlights four such roles: technical editors, who correct small errors related to style or formatting of articles; vandal fighters, who revert vandalism and sanction norm violators; substantive experts, who improve the quality of the content of the articles; and finally, social networkers, who support community aspects of Wikipedia and contribute little to the content and form of articles directly.

We then use simple metrics based on edit histories to examine two samples of Wikipedians--a collection of long term dedicated editors, and a cohort of editors from a 6 month window of new arrivals--to explore how users adopt and adapt to these roles. Technical editors and vandal fighters concentrate their edits on content pages while devoting relatively few edits to the discussion pages for those content pages. In contrast, substantive experts show greater investment in discussion, both related to the articles and directly with other editors. Social networkers also concentrate on discussion and user page edits, but make very few

edits to content pages. Egocentric network visualization provide a second set of suggestive patterns. First, social networkers and substantive experts tend to develop denser community structures with more active alters, and engage in more reciprocated ties, while vandal fighters and technical editors are likely to have larger proportions outward links to local isolates. Finally, we find that the proportions of potential editor types in the new cohort are similar to the rates observed in sample of dedicated contributors, suggesting that the informal socialization into helpful roles in Wikipedia was generating enough new role players to sustain and grow the population and proportion of helpful contributors. These results are preliminary, and we suggest several ways that the method can be improved, including the expansion and refinement of role signatures and how these methods could be extended to study role ecology in online communities.

2. FINDING SOCIAL ROLES

2.1 Roles in Interaction and Wikipedia

Across social settings we can identify people who are playing social roles: advisors, parents, brokers, editors, managers, or vandals. The concept of "social role" has long been used in social science describe the intersection of behavioral, meaningful, and structural attributes that emerge regularly in particular settings and institutions [20], [5].

Social roles have mainly been studied online in the context of text based discussion spaces, where a variety of roles have been identified, including local experts, answer people, conversationalists, fans, discussion artists, flame warriors, trolls, and even lurkers [11], [25], [16], [29]. Insight into these social roles has been gained through ethnographic study of the content of interaction and through the use of behavioral and structural cues [29], [12].

Wikipedia differs from discussion spaces in that the primary activity of the community is the construction of an artifact. In this way, Wikipedia is similar to open source software development, and in both domains, researchers have studied questions about why people participate [4], [27], the quality of the resulting artifacts [1], [23], and how coordination relates to the quality and structure of the work [3] [17].

Relatively little research, however, has gone into how these groups define and manage specific roles in coordinating their work. In the case of open source, project roles (owner, developer, bug reporter, technical support) are clearly defined, which likely makes coordination easier (though non-trivial), while communication and coordination patterns often align with the structure of the software itself [3].. In contrast, Wikipedia has few clearly defined roles; those that exist, such as administrator and bureaucrat, are primarily used to grant extra powers such as the ability to block troublesome users from editing and protecting controversial pages from vandalism. Kriplean et al. [18] show that informal awards (Barnstars) are used to encourage and reward different types of valued work, and suggest that these Barnstars may be used to identify existing or emerging types of work that may correspond to different roles in Wikipedia.

Though formal roles are few, Wikipedians recognize a number of informal roles as well, including fighting vandalism, welcoming new users, managing the featured article process [26] and writing tools to help the community [8]. These informal roles provide an open structure that supports legitimate peripheral participation [19], the process by which new users learn to contribute by observing, and eventually emulating, the behavior of established Wikipedia editors [4]. Studying these informal roles may help us how community processes and monitoring can support the coordination problem of collective action, as well as providing tools for reasoning about the current and future health of the community (e.g., if there are not enough answer people in a discussion group, or enough vandal fighters in Wikipedia, the value of the community to others may suffer).

2.2 Operationalization of Roles

Social roles can be conceptualized at several different levels of abstraction. The challenge for researchers is to identify roles that affect the course of social action. Gleave et al. [12] contend that the best way to identify roles that matter is to begin at the level of meaningful social action, and to work both downward towards identifying the key related behavioral regularities and distinctive positions in social networks (signatures of social roles), and upwards to abstract theoretical categories that allow us to tie these particular roles to general research objectives that transcend any particular study or social context.

Focusing on lower-level behavioral regularities or distinctive positions is flawed because it is overly inductive: through extensive analysis of behavioral and social data, we may detect an enormous number of patterns, but there is no a priori reason to think that the ones that initially stand out will be of any social significance. Starting from abstract categories, like 'altruist', commits the error of over-deduction. Just because we can assign a label to an activity does not necessarily mean that those behaviors are motivated by altruism (evolutionary biology presents a number of such “altruistic” cases [21]).

Instead, Gleave et al [12] argue for the value of multi-level analysis of social systems, describing five levels of analysis where role related patterns can be conceptualized. Consider the role of a Wikipedia editor who concentrates on contributing new content, one that we describe as substantive expert. First, a ***behavioral regularity*** for a substantive expert would include establishing new pages, expanding brief entries (called stubs), or refining the content in a range of related pages. Because changing content often requires discussion with other contributors, we might expect some ***network attributes*** of substantive experts to include relatively large community structures with similar alters. Wikipedians can also ***self-identify*** as substantive experts on their user profile pages, indicating their areas of editing interest or naming relevant qualifications. A qualitative investigation into the first three levels can help researchers develop a clear ***role definition*** that maps onto important dimensions of interaction in the social setting. Finally, that role definition can be related to ***abstract categories and classifications*** of social types, like altruists, or cooperators.

iConference'11, February 8–11, 2011, Seattle, Washington, USA.

Connecting higher-level theories to observable behavioral regularities and distinctive network positions through "medium level" phenomena that are socially meaningful to participants can be informative by connecting behavior, process, and theory. For example, Crandall et al. [9] observe a sharp rise in "similarity between two Wikipedia editors" (abstract category) as measured by overlap in edited pages (a behavioral regularity) at about the time they first communicate with each other (creating a network attribute) through a large-scale quantitiative analysis of editing behavior. By itself, this analysis sheds no light on *why* this happens. They went on to examine, qualitatively, a series of cases of "first contact" between editors. This analysis showed that initial communications generally happen because the two editors are editing the same article at about the same time, and one decides to make the coordination more explicit, e.g., to encourage further work or to resolve a disagreement, becoming more like a team working together (role definition).

2.3 Structural Signatures

This paper uses systematic patterns in contribution to identify the signatures of particular roles in Wikipedia. The strategy of using structural signatures of social attributes of actors has been applied to a variety of settings. Researchers at Bell Labs identified fraudulent telephone accounts by leveraging patterns in volume, timing, and the identity of in-bound callers [7]. In a similar fashion, consistent driving records are used as proxy for good credit risks where credit data are unavailable [14]. In more direct measurement situation, Ebay sellers can be identified as reliable or trustworthy through the accumulation of many highly evaluated transactions, a type of emergent, and hard to fake reputation [10]. Our strategy draws on these works in general. In particular we follow earlier studies of social roles in Usenet that used visualizations of cumulative patterns in message contribution as well as attributes of local social networks [11] and to identify social roles like that of answer person [29].

3. METHODS

3.1 Identification and choice of roles

Building an encyclopedia requires a great deal of work across a broad variety of tasks. Kriplean et al. [18] identify no less than forty two types of work the community values, grouped into major categories such as editing work, community support, "border patrol" (i.e., maintaining standards), administration, managing collaboration, and contributing meta-content such as tools and templates. These kinds of work can be very specialized, such as managing the process by which articles are "featured", or promoted as high-quality examples [26], or maintaining and watching closely over specific articles to maintain both quality and personal investment [24]. Further, the distribution of work and activity levels across editors of a specific article can impact its quality [17].

In the following sections we seek to identify, then explore the distribution of, some of the most visible roles in Wikipedia, several of which are related to the high-level categories from [18]. We ground our roles both in those categories, which are drawn from Wikipedia users' explicit recognition of others' activity, and in the visible activity of page editing. Wikipedia pages belong to different "namespaces" which group the pages based on what role they play: the articles themselves (the "Main" namespace), pages for discussing article creation ("Talk"), members' pages and pages for communicating with other members ("User" and "User Talk"), and so on. These kinds of activity roughly correspond to four main roles that we focus on: substantive experts, technical editors, social networkers and counter vandalism editors. We recognize that these roles are exhaustive[1], or even necessarily the most common--as we will see, social networkers comprise a small fraction of Wikipedians. Instead, we chose these roles because they are relevant to both the social interaction common on Wikipedia, and they play important roles in the construction of the encyclopedia. The next section provides a brief discussion of range of these socially meaningful roles in Wikipedia and relates them to the organizational challenge of coordinating contribution in the absence of explicit top down management.

Substantive experts. Substantive experts contribute by providing substantive content to article pages. The may display extensive knowledge in a topic, and some cite real world credentials on their user pages to bolster their credibility. They contribute substantially to pages within a particular subject area and resolve article-related disputes on article talk pages in their areas of expertise. Though their credentials may or may not come into play, substantive editors are often people who invest time in fact checking and article talk to discuss details of articles.

Technical editors. There are dozens of areas in Wikipedia where small errors can crop up: spelling, grammar, hyperlink format, out of date facts, links to other language editions of Wikipedia, and so on. Likewise, there are activities such as categorization and building templates that help to organize and standardize Wikipedia. The term technical editor refers to all contributors who engage largely in these sorts of incremental improvements and maintenance of Wikipedia's content.

Counter vandalism. Counter vandalism editors find vandalized articles, correct them, and sanction vandals. Users tend to self-identify as part of several groups such as the Counter Vandalism Unit; a large percentage of these users identify with the notion of fighting vandalism will revert occasional vandalism. The percentage of those users who actively participate in significant anti vandalism efforts is smaller. Though there is a range in numbers of anti-vandalism contributions for editors, many editors are devoted strictly to anti vandalism tactics. Because these editors find vandalism either on the "Recent Changes" page, anti-vandalism bots or by tracking specific users, counter vandalism editors will have a higher percentage of article edits with little or no relation between article topics.

Social networkers. Lacking well-defined rules and boundaries, Wikipedia offers users many possibilities for interacting with one another. Those contributors who make frequent use of Wikipedia's networking and communication potential will be referred to as social networkers. Social networkers build strong ties with other users through channels other than article collaboration. They utilize User Talk extensively, make "Wikifriends," and create elaborate profiles that showcase their Wikipedia personalities. Their User Pages often contain many Userboxes, small snippets of self-identifying information

[1] Likewise, people may play several roles simultaneously, and these roles may change over time. In our analysis this shows up as noise in behavioral signatures of roles, but is itself an important phenomenon worth studying.

including interests, group membership, and personal characteristics. Social networkers often participate in projects that can be seen as community-building. These include "The Birthday Committee," a variety of projects associated with "Wikipedia Culture," the "Welcoming Committee" for new users, and parts of the now defunct "Esperanza" project whose goal was to strengthen the Wikipedia community.

3.2 Analysis Strategy

Following work identifying roles in online discussion [29] our analysis proceeds through two stages. The first is an exploratory stage where we use data visualization, descriptive statistics, and content-based fact checking to learn the structural signatures of different social roles in Wikipedia. The investigations begin with broad qualitative explorations that identify individuals performing interesting roles, after which maps of the structural positions of populations are used to identify network patterns that differentiate users. The next step is to analyze the context of participation and the content of behaviors of the actors whose interactions formed those social network structures. This iterative process moves between content and structure to refine our understanding of social roles and validate the relationship between structural attributes and behaviors. The remainder of the methods section includes a description of our population samples, our data, and how we grouped the Wikipedia namespaces into meaningful units related to the roles we identified.

3.3 Samples

Directed (N=40) The directed sample includes hand-picked contributors who, based on the types and content of their edits, as well as their user pages, seem to perform the roles of substantive expert, technical editor, counter vandalism, or social networker. We use this sample to document the patterns and exceptions we find within each type and to develop insight into metrics for distinguishing these roles from other types of editors. For technical editors and counter vandalism we were able to use lists of participants in related Wiki projects (pages that exist to help people interested in specific topics or issues around Wikipedia find each other) to identify possible role players. Substantive experts and social networkers were primarily identified through their presentation of self on their user page and by reading through their edit histories.

Dedicated Editor (N=1954) This sample consists of a population of dedicated, long-term Wikipedia contributors. To generate this sample, we selected all editors whose first edit was on or before January of 2004, and who made at least one edit during January of 2005. This population allows us to explore how experienced Wikipedia users distribute their work across the project.

Cohort (N=5839) The cohort sample consists of all editors who created accounts and made at least one edit during the month of January 2005. This sample allows us to measure the proportion of all users, not just committed users that fall into each of the observed roles, and to see if the adoption of roles has changed over time.

3.4 Data

Our data are drawn from records of edits, organized by editor, distributed in the Fall 2006 data dump, available for free download from MediaWiki (http://meta.wikimedia.org/wiki/Data_dumps). Data extends from the beginning of the English Wikipedia to October of 2006.

Table 1. Namespace designations

Category	Namespace	Comments
Content	[0,6]	Articles and images.
Content Talk	[1,7]	Discussion pages
User	[2]	Personal pages related to login identity.
User Talk	[3]	Primary user to user communication mode.
Wikipedia	[4,5]	Community pages, help desk, village pump; related talk pages.
Infra-structure	[8,9,10,11,12,13,14,15, 100,101]	Pages that provide infrastructure for other tasks in Wikipedia; template, categories and portals.

We used this file to generate the Dedicated and Cohort samples, as well as to construct monthly activity records for all of our users. For each month, for each user, we group their activity into one of six categories based on the namespaces in which they edited. Table 1 presents a list of the Wikipedia namespaces (see http://en.wikipedia.org/wiki/Wikipedia:Namespace) and our categorization of them.

Table 2. Edit totals and percentages for sample datasets

	Coh.	Ded.	Dir.	Dir. Sub	Dir. Tech	Dir. CV	Dir. SN
Months Active	4.5	32	16	20	20	12	10
Total Edits	251	5k	7k	7k	12k	6k	2k
Edit Rate	55	159	464	374	613	517	249
Content	68%	73%	59%	53%	65%	68%	21%
C- Talk	9%	8%	7%	15%	2%	4%	6%
User	5%	3%	6%	5%	4%	5%	22%
User Talk	6%	5%	8%	11%	4%	8%	26%
Wiki pedia	9%	9%	14%	13%	15%	11%	23%
Infra Struct.	3%	2%	6%	3%	9%	4%	1%

As we discuss in detail below, the proportion of edits dedicated to a particular namespace as well as the distribution of those edits across time and across pages can reveal signatures of different social roles. We use histograms to compare the distribution of edits between namespaces; Table 2 presents the overall breakdown of activity across namespaces for each of the samples and for users in the directed sample.

We also use the data to construct local network visualizations, where users are nodes and a directed tie exists from A to B if user A edited user B's "User Talk" page, to explore how network

structures and structures of relationships can serve as indicators of users' roles.

4. ANALYSIS

4.1 Role types by distribution of edits

How are different roles revealed by how people divided their edits across namespace categories? We begin to address this question by comparing the average distributions of edits across the six activity categories across the four roles in our directed sample, shown in Figure 1. Because content edits often account for more than 50% of users' activity, we make a first distinction between content edits and all other edits (the pie chart in each graph) in order to improve readability of the histogram, which shows the percent of the non-content edits that fall into each of the other five categories: content talk, user, user talk, wikipedia, and infrastructure.

These distributions were calculated from averages for each category of qualitatively identified role holders, thus, they are possible general trends in edit distribution from a set of editors who self-identified through membership in projects or through declaring their interests, activity, and self-identified roles on their user pages. Because the samples are small, and because people can play multiple roles, they are should be taken as preliminary insights into possible connections between roles and edit distributions. Still, they are suggestive.

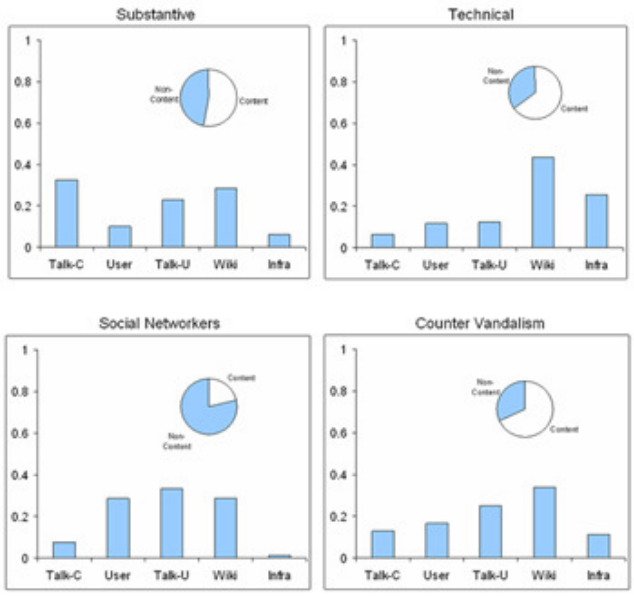

Figure 1. Edit distributions by role types in directed sample

The average **substantive expert** makes only about 50% of edits to content space. This rate is about 10% lower than what we observed for anti-vandals or technical editors. However, substantive experts have a much higher rate of posting to content talk than other roles. This suggests generally, that when substantive experts contribute to content pages, their contributions are likely to be more costly (take more time and thought, and are more likely to require explanation, justification and discussion on the content talk pages and sometimes with individual users on the user talk pages). Their distinctive investment patterns are defined by the combination of a lower content edit rate with elevated content talk, followed by edits to Wikipedia namespace pages.

Technical editors make numerous small changes to content pages, frequently specializing in a particular type of problem (spelling, grammar, faulty links, improper copyright information, etc.) Thus there are several subtypes of specific edit patterns within this class of editors. However, within this class there is a shared tendency to invest primarily in content edits, with contributions to Wikipedia pages a strong second in edit rates. These editors are primarily making the specific content changes associated with the issue(s) they like to address and holding discussions about community and infrastructure related to reinforcing those tasks (hence the Wikipedia name space edits).

Counter vandalism editors have edit profiles very similar to other types of technical editors. They have moderately high rates of content edits, followed by investment in Wikipedia and user pages. Where technical editors concentrate their non-content edits primarily in Wikipedia namespace, counter vandalism editors had surprisingly high rates of edits to the User and User Talk namespaces. Closer inspection of edit histories for counter vandalism editors showed that large portions of their User page edits were made to *other* people's User pages, which was surprising at first. However, this behavior is explained by counter vandalism editors who are also admins; blocking vandals requires a post to a user page.

Finally, the **social networking editors** provide a sharp contrast to the other three types—they invest very little in content edits, and invest primarily in their own user page. Their second priority involves user talk, and next are investments in Wikipedia namespace pages, which are often associated with community building and support for social interaction among editors. Partly because their focus is on cultitvating social relationships rather than helping others work, their investment in infrastructure is negligible.

4.2 Network structure

Patterns in relationships can help distinguish between roles in online spaces [2], [12], [29]. Figure 2 illustrates some differences in the structure of relationships in user talk networks, comparing five examples of each role type in our directed sample. These graphs display the interconnections between ego and all of ego's one degree neighbors. The arrow travels from author to person whose user talk page received the edit, red arrows indicate that ego was the author, blue that an alter was the author. Nodes are sized according to total out-degree of each node.

Counter vandalism work also often involves posting warnings on user talk pages, which explains the relative frequency of edits to user talk pages compared to other technical editors. With the exception of these highly specific edits to the user pages of banned vandals, the edit distribution of counter vandalism editors seems very similar to other types of technical editors.

These examples illustrate several patterns that are logically related to role related behavior. Based on those role related insights, researchers could construct wiki related metrics that would help distinguish between different role types. Constructing and testing those metrics is beyond the scope of this paper, however, earlier

research has shown that close attention to local networks and neighbor degree distributions can reveal structural signatures of social roles in online communities [2], [29]. At the most general level, technical editors and vandal fighters have similarly sparse local networks, while the social networkers and substantive experts' networks show larger community structures.

The counter vandalism and technical editors share several attributes with the local network patterns and neighbors degree distributions of "answer people" [29]. Key features include highly skewed neighbors degree distributions (many ties to alters with few other ties), and very few interconnections in their local networks. This makes sense, both the tasks of reverting vandalism and making small technical edits are likely to put one in contact with a diverse range of alters who lack interconnections, and with alters who may be new or otherwise unconnected to many alters in Wikipedia. However, both technical editors and vandal fighters show evidence of small, highly interconnected subgroups with higher degree alters. This makes sense too, given the complex and user generated rule structure of Wikipedia. All of the technical editing tasks in Wikipedia have an implementation side, and a negotiation side. Even simple editing tasks require some negotiation about how, what and when and where to implement particular rules. More complex tasks, require even more negotiation. So, in general we should expect vandal fighters and technical editors to have many one-off interactions on the front stage, with ongoing interactions with homophilous alters in the backstage.

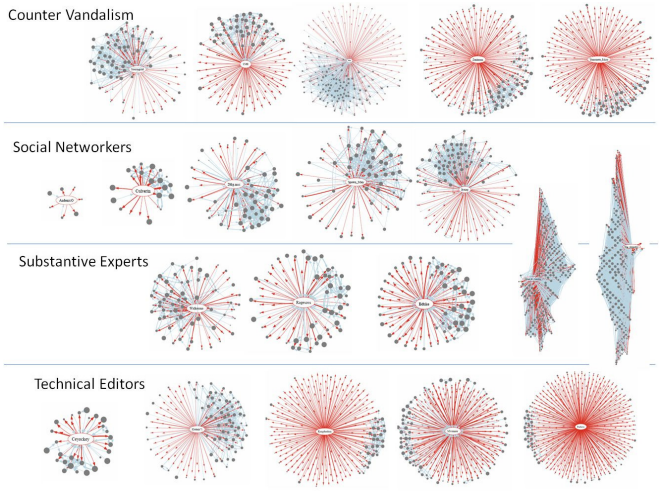

Figure 2. Egocentric user-talk network graphs for role types in directed sample

Social networkers essentially have no front stage. As they focus the vast majority of their energy on sprucing up their user page and interacting with friends, they are likely to develop user talk networks that only include friends who are similar to themselves, or other folks that they run into in the backstage. The social networkers we sampled tended to have densely interconnected communities that occupied a large portion of their local networks. Alters in this local network also had high out-degree, both of these attributes make their network signatures similar to that of the discussion people in Usenet [29].

Substantive experts also show large communities that include high degree alters. However, substantive experts have both front and back stage responsibilities, so they are likely to develop bother relationships within their community of fellow experts, and to outsiders of that interconnected subgroup. In this regard, the substantive expert role is markedly different from that of technical editors and vandal fighters. While small technical changes are likely to put an editor in one-off contact with newbies, substantive experts are much more likely to talk with other experts, as they negotiate the proper content for various pages. Further, changing content is often complex and thus substantive experts should have high rates of mutuality in their contacts to relative isolates in their networks.

4.3 Comparing role prevalence across two samples of Wikipedia editors

How does the role distribution in a "new" cohort of editors compare to the role distribution in a sample of dedicated editors? This section takes the observations about how edit distributions might be associated with particular roles, and constructs a very simplified set of variables that assign Wikipedians in both samples into these roles or not. First, we set an activity threshold of at least 25 total edits, because cases with too few edits spread across many categories of behavior is likely create spurious patterns.

We assigned editors to our four possible roles based on a few distinctive attributes in their edit distribution depicted in the average histograms reported earlier. In those figures and in the following thresholds the proportion content edits are based on total edits, while the remaining proportions employ the sum of all non-content edits as the denominator. Wikipedians were coded as likely role players if they met the 25 edit thresh hold and met all of the binary attributes for the given role type.

Substantive: 30-80% total edits to content pages, <30% to Infrastructure, >45% in content talk and Wiki combined; and >25% content talk.

Technical: >60% of total edits to content pages, >45% in Wiki and Infrastructure combined, and <25% content talk.

Social networkers: less than 45% of total edits to content pages, less than 30% edits to infrastructure, content talk less than 25%, greater than 45% user and user talk combined, greater than 25% wiki pages.

Vandal fighters: >60% content edits, <25% content talk, >30% user and usertalk combined, and >20% Wiki pages.

Coding our samples according to these variables resulted in some suggestive findings about the relative distribution of potential role players in Wikipedia. First, within the nearly two thousand editors in the dedicated sample, almost one third of them show edit patterns consistent with the substantive expert role. About one in ten editors were technical editors, and about six percent were vandal fighters, and only a trivial percentage was coded as potential social networkers. We cannot conclude that the remainder of the dedicated sample does not play any of these roles. First, we know that our indicators are quite primitive and imprecise. But second, even if the indicators were perfectly predictive of role behavior, we would expect many editors to fall outside the predicted role types because many editors play

multiple roles, and thus would exhibit edit distributions that varied outside of the patterns we identified.

The cohort sample included all 5839 editors who created new login identities during the month of January 2006, and tracked their behavior for the subsequent twenty one months. The vast majority of these new editors made fewer than twenty five edits, although a sizable number (1672) did exceed the threshold. Within this set, the relative distribution of predicted role types is very similar to that observed for the dedicated sample. The columns report the absolute number of editors assigned to the potential role category. The percentages reported in Figure 4.4.1 indicate the proportion of active editors in each sample that were assigned to the potential role category. While the proportions are fairly similar in the two samples, the cohort proportions are slightly smaller except for the social networkers.

Using Wikipedia for social networking was actually a relative new development in 2006, and thus very few editors from the dedicated sample matched this edit distribution profile. The cohort sample shows an increase over the dedicated sample, but still reflects a very small absolute number, and very small proportion of the new cohort. With the advent of good social networking systems like Facebook it seems possible that this role type may no longer be especially relevant, but is worth testing on newer data.

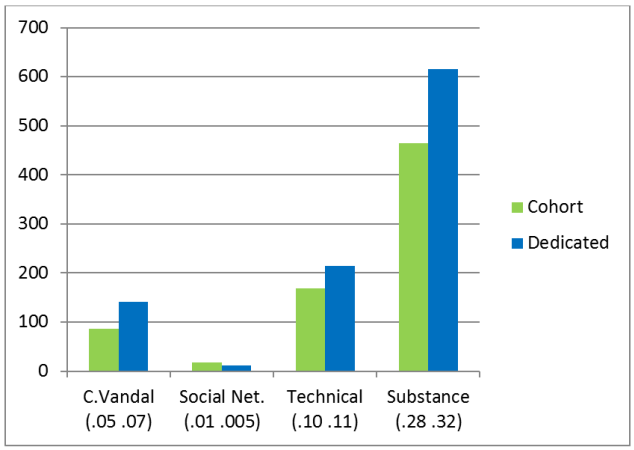

Figure 6. Role distribution in cohort and dedicated samples

Are key role holders being replenished? Quite likely. It seems that potential role players are arriving and developing at a rate that is more than sufficient to supplement and grow the current population. Consider the fact that the cohort sample includes new editors from a single month in 2006, and the absolute number of potential role players nearly equals the number of potential role players from the dedicated sample. Assuming roughly consistent entry rates across months, this implies that in a single year the operation of the Wikipedia social system was cultivating about ten times as many potential role players as had been carried over from all of the previous years. Goldman [15] raises concern over the potential of Wikipedia to replenish its expert role players. This is an important question, because the success of Wikipedia depends on ordinary people playing extraordinary roles in a large and largely uncoordinated system. Our preliminary results suggest that people are able to find their roles in Wikipedia. In fact, because the cohort represents a single month of new role players, the rate of role production seems likely to result in a surplus of potential new role player. Furthermore, we note that administrators, leaders and other organizational players should want to avoid changes that disrupt the flow of new entrants into more complex modes of contribution.

In our discussion of edit distributions related to roles we noted that technical editors were difficult to distinguish from vandal fighters. This proved true also in this analysis. There were 60 cases in the combined data set where the same editor was coded as being both a likely vandal fighter and technical editor. None of the other role potential role combinations were observed. However, a large proportion of editors did not exhibit edit distributions distinctive enough to be coded as playing any of the identified roles. This underscores the need for future research to predict role categories based on degree of performance rather than simple binary assignment.

5. CONCLUSION

Our research began by suggesting that a key to solving the coordination problem inherent in the large, complex task of Wikipedia authorship is provide by the solution to a prior problem, the challenge for each author to find their role(s) in Wikipedia. Through an exploratory analysis we took some steps towards recognizing some important roles in Wikipedia, and we identified potential structural signatures based on name space edit distributions and user talk network features. We tied these potential signatures to role behavior, and provided a preliminary illustration of how some of these features can help assess the role ecology of an online space. In particular we showed evidence that new cohorts of editors were exhibiting edit distributions consistent with role players, and that these new cohorts were generating new potential role players at rates that were probably high enough to meet the replacement demands of the system.

Much room is left for improvement and development in new research. The potential structural signatures identified here need to be refined and tested for predictive accuracy. Wikipedia affords an ideal research site for improving the identification of structural signatures because of the high level of contextual detail that is associated with every edit. We showed how namespace level aggregations can be leveraged, but additional details can aid signature identification. For instance, substantive experts will make multiple edits to related sets of content pages, and will likely return to sets of pages repeatedly. These types of details would help refine role predictions. Similar advances are possible with network ties, where the content of a message could be used to code ties as positive, negative, or neutral, could reveal much more fine grained role signatures. The role status of alters could also greatly aid role prediction. Consider the fact that the edit distributions of vandal fighters and technical editors were very similar. Discerning vandal fighters from other technical editors becomes much easier if we can identify obvious vandals, and thus multiple ties to vandals would be a clear distinguishing factor.

Wikipedia is a complex social system. Although our analysis identified a subset of roles, that short list is neither exhaustive nor are these roles mutually exclusive. Many editors are likely to remain generalists, who dabble in a range of role related tasks. Others, might concentrate on a couple roles, and thus exhibit contradictory patterns. As role signatures become more refined, we should aim for systems that can assess degree of role

performance, and, ideally, to track assessment across time to monitor role change. Finally, we anticipate a moment where standard demographic methods can be applied to test higher level questions about the role ecology of Wikipedias, and ideally, a large scale comparative study across different Wiki systems could be performed.

6. ACKNOWLEDGMENTS

The authors would like to thank the Institute for Social Sciences at Cornell University and the National Science Foundation grants [SES-0537606; and 0835451] for support that made this research possible.

7. REFERENCES

[1] Aberdour, M. 2007. Achieving Quality in Open Source Software. *IEEE Software,* 24(*1*):58-64.

[2] Adamic, L., Zhang J.; Bakshy, E.; Ackerman, M. S. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. *ICWSM2008*. Seattle, WA.

[3] Bird, C., Pattison, D., D'Souza, R., Filkov, V., and Devanbu, P. 2008. Latent social structure in open source projects. In *Proc. of the 16th ACM SIGSOFT Int. Sym. on Foundation of Software Engineering*, 24-35.

[4] Bryant, S. L., Forte, A., and Bruckman, A. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proc. of the 2005 Int. ACM SIGGROUP Con. on Supporting Group Work*, 1-10.

[5] Callero, P. L. 1994. "From Role-Playing to Role-Using - Understanding Role as Resource." *Social Psychology Quarterly,* 57:228-243.

[6] Chesney, T. 2006. An empirical examination of Wikipedia's credibility. *First Monday,* 11-6.

[7] Cortes, C., Pregibon, D. 2001. "Signature-Based Methods for Data Streams." *Data Mining and Knowledge Discovery,* 5:167-182.

[8] Cosley, D., D. Frankoski, L.G. Terveen and J. Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in Wikipedia. *Proc. of the 12th Int. Con. on Intelligent User Interfaces*, 32-41.

[9] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *KDD 2008*.

[10] Friedman, E. and P. Resnick. 2000. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy,* 10(2): 173-199.

[11] Fisher, D., Smith, M., Welser, H. T. 2006. "You Are Who You Talk To: Detecting Roles in Usenet Groups." In *Proc. of the 39th Hawaii Int. Con. on Systems Sciences (HICSS).*

[12] Gleave, E., H. T. Welser, T. M. Lento, and M. A. Smith, 2009. A Conceptual and Operational Definition of 'Social Role' in Online Community, *Proc. of the 42nd Hawaii Int. Con. on Systems Sciences (HICSS).*

[13] Giles, J. 2005. Internet Encyclopedias Go Head to Head. *Nature,* 438: 900-901.

[14] Guseva, A., and Rona-Tas, A. 2001. Uncertainty, Risk, and Trust: Russian and American Credit Card Markets Compared, *American Sociological Review* 66:623-646.

[15] Goldman, E. 2009. Wikipedia's Labor Squeeze and its Consequences. *Journal of Telecommunications and High Technology Law*. Vol 8.

[16] Haythornthwaite, C., Hagar, C. 2005. "The Social World of the Web." *Annual Review of Information Science and Technology 39*: 311-346.

[17] Kittur, A. and Kraut, R. E. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In *CSCW '08. Proc. of the 2008 ACM conference of Computer Supported Cooperative Work.* ACM Press.

[18] Kriplean, T., Beschastnikh, I., & McDonald, D. 2008. Articulations of Wikiwork: Uncovering Valued Work in Wikipedia through Barnstars. *CSCW* San Diego, CA, USA.

[19] Lave, J., Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press.

[20] Merton, R. K. 1968. *Social Theory and Social Structure*. Free Press.

[21] Pinker, S., 1999. *How the Mind Works.* W.W. Norton.

[22] Read, B. 2006. Can Wikipedia Ever Make the Grade? *Chronicle of Higher Education*, 53(1): A31-A35.

[23] Stvilia, B., Twidale, M. B., Smith, L. C., Gasser, L. 2005. Assessing information quality of a community-based encyclopedia. *In Proc. ICIQ* 442-454.

[24] Thom-Santelli, J., Cosley, D., Gay, G. 2009. What's Mine is Mine: Territoriality in Collaborative Authoring. *CHI 2009.*

[25] Turner, T.C., Smith M., Fisher, D., Welser H. T. 2005. Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer Mediated Communication*. 10(4).

[26] Viegas, F., Wattenberg, M., Kriss, J., & van Ham, F. 2007. Talk before you type: Coordination in Wikipedia.. In *Proc. of the 40th Hawaii Int. Con. on Systems Sciences (HICSS).*

[27] Weber, M. 1978. *Economy and Society*. University of California Press.

[28] Weber, S. 2004. *The Success of Open Source*. Harvard University Press.

[29] Welser, Howard T., Eric Gleave, Danyel Fisher, and Marc Smith. 2007. "Visualizing the Signatures of Social Roles in Online Discussion Groups." *The Journal of Social Structure.* 8(2).