Identifying Social Roles in reddit Using Network Structure

Cody Buntain
Department of Computer Science
University of Maryland
College Park, Maryland 20742
cbuntain@cs.umd.edu

Jennifer Golbeck
College of Information Studies
University of Maryland
College Park, Maryland 20742
golbeck@cs.umd.edu

ABSTRACT

As social networks and the user-generated content that populates them continue to grow in prevalence, size, and influence, understanding how users interact and produce this content becomes increasingly important. Insight into these community dynamics could prove valuable for measuring content trust, providing role-based group recommendations, or evaluating group stability and growth. To this end, we explore user posting behavior on reddit, a large social networking site comprised of many sub-communities in which a user may participate simultaneously. We demonstrate that the well-known "answer-person" role is present in the reddit community, provide an exposition on an automated method for identifying this role based solely on user interactions (foregoing expensive content analysis), and show that users rarely exhibit significant participation in more than one communities.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; G.2.2 [**Graph Theory**]: Network Problems

General Terms

Social Networks, Social Roles, Online Communities

1. INTRODUCTION

Increasingly prevalent and influential social networks facilitate community development and communication in unprecedented ways. This resulting connectedness is beneficial through the communities fostered across physical, national, and ideological boundaries, but dangers exist given our insufficient understanding of the space. Answers to how users generate the content that fills these networks, their behavioral patterns, and what this data says about us are all still nebulous and unclear. For instance, the anonymity the Web provides serves as a boon for political activists in fear of reprisal but also serves as a mask of unaccountability for delinquents, troublemakers, and "trolls." By enhancing our understanding of these behavioral patterns, or roles [1], in communities, we can assist users in navigating online interactions and promote higher-quality content.

This idea of identifying user roles in online communities is not new; researchers have been exploring roles in email and newsgroups since the 1990s. Much of this existing work, however, focuses on identifying a user's role in the entire network rather than identifying roles in the individual communities in which that user participates. The implicit assumption in existing literature is that user behavior is consistent across all of these communities (that is, a user who assumes role X in one community would also assume role X in all the communities in which he participates). While such an assumption is intuitively invalid given how people interact differently in various social groups (a workplace community versus a community of friends), before we can establish or refute this assumption, we must first answer whether we can identify social roles in a social network with multiple communities and whether users participate in multiple communities at all (the anecdotal "1% Rule" might suggest users do not). To explore these social dynamics, we must first identify a target social network for analysis that is composed of many smaller communities. This dynamic is readily apparent in reddit [2], a large web-based community where many varied communities (or "subreddits") exist underneath the umbrella of the larger reddit community. reddit's inter-communal behavior offers an interesting microcosm of social interaction by allowing us to explore a user's interactions in a single community and across many sub-communities.

Our investigation draws on existing literature and focuses on a specific social role, the "answer-person" role, in which a user's dominant behavior is to respond to questions posed by other users but engage in only limited discussion [16]. By focusing on this role, we can leverage established research that shows this can be classified without the need for computationally expensive content-based classification. Furthermore, this role plays a key role in web-based resources and is likely to be present in at least one of reddit's many subreddits. From this foundation, this paper seeks to answer three questions: 1) whether the answer-person social role exists in the reddit community, 2) if it is possible to identify this role in an automated fashion, and 3) whether users participate significantly in multiple communities. We begin this analysis with a discussion on related work to establish a basis

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579231.

[1] A "role" here refers to a pattern of behavior and not to a position or job title as it does in some other literature.

[2] http://www.reddit.com

for role identification in electronic communities (Section 2) and then lay out the structure through which we conducted our experiments (Section 3). We follow up with a discussion on data gathering, results, and analytics (sections 4 and 5) before closing (Section 6).

2. RELATED WORK

Since Milgram's work on the value of weak ties in social networks, social science scholars have displayed a keen interest in analyzing social networks, their evolution, and the parts contributors play in them [8]. As electronic communities began to emerge with Usenet and bulletin board systems (BBSs) in the late 1970s and early 1980s, these scholars soon had access to valuable stores of information that facilitated direct analysis of such social networks without the need for lengthy, costly surveys on face-to-face interaction. In the 1990s and early 2000s, much of the analysis surrounding these electronic communities focused on content analysis to understand contributor types, community topics, and user interactions in these large collaborative environments. Donath's work on visualizing these interactions laid the foundation for the structural signatures we leverage in this work to identify specific roles [2].

Building off the works of Donath and others, Agrawal et al. explored the social network characteristics present in electronic communities [1]. In his work, Agrawal modeled the Usenet community as a graph in which nodes represented contributors, and links between contributors represented responses. Agrawal then leveraged graph theoretic algorithms to partition this Usenet social network model into binary sets representing whether a contributor was for or against some topic. By relying exclusively on network analysis, Agrawal was able to delineate between proponents and opponents of his test topics with higher accuracy than methods that relied on content analysis. Agrawal ended the paper with a brief supposition that these network links are perhaps more informative than the actual content for such classification tasks. The success in modeling these electronic communities as networks furthered the way for several new research techniques in areas such as topic modeling, sentiment analysis, and role detection, which in turn provided our work with models for representing the reddit sub-communities as social networks [13, 11].

Though research into the dynamics and evolution of electronic communities is broad, of particular relevance to our work is social role detection in these communities. Though role detection in the physical realm has been an area of research for many years, Golder and Donath's 2004 work applied much of the existing sociological and psychological theory to electronic communities, Usenet in particular [4]. Their analysis included user participation, types of interactions, speech patterns, and several other behavioral dimensions to deconstruct the components of one's role in the digital realm. From that investigation, they constructed a new taxonomy of roles specific to electronic communities, which included roles for the celebrity, newbie, lurker, and troll. Though this taxonomy lacks the answer-person role on which we concentrate, the celebrity role's influence in the community is similar in some aspects.

Fisher, Smith, and Welser of Microsoft Research then built on this work by applying Agrawal's structural analysis techniques and 2-degree egocentric network visualizations to identify roles in a set of Usenet communities without the intensive content analysis techniques used in the aforementioned works [3]. Fisher targeted a range of communities, each with a distinct purpose from question-and-answer to discussion to support to malicious disruption. By visualizing these ego networks, Fisher identified significant differences between contributors that suggested specific communal roles as well as unique cross-community differences. Fisher, Wesler, and Gleave then expanded on their results on community- and behavior-specific network structures in the following year [16]. In this new paper, Wesler focused exclusively on classifying roles via users' structural signatures, forgoing content analysis completely, which lead to substantial performance enhancements. Owing to their importance in different aspects of community building, content generation, and information quality, this new work concentrated only on two specific social roles: the answer-person role and the discussion-person role. Because of their distinct structural characteristics and prevalence in electronic communities, we adopt this restricted view and concentrate only on these roles as well.

Over the past five years, researchers have applied these social role analysis techniques to many other communities besides Usenet. In 2008, Gómez et al. explored user respondent activity in the Slashdot community as a mechanism for identifying controversy [5]. Welser et al. published additional work in 2011 in which they explored the Wikipedia network to identify more community-specific roles with a good deal of success [15]. Similarly, 2012 saw a pair of papers concerning role identification in the Twitter social network and in Governor Sarah Palin's email network [12, 7]. Interestingly, many of these works identify sets of behaviors that are often unique to the target communities, which suggests our question of multi-community interaction and others like whether a user's roles are conserved across communities are valid. For instance, it may be reasonable for a user to participate in both the Slashdot and Wikipedia communities, but will that user's roles be identical in those communities given the unique characteristics present in each community?

Rossi and Gallagher partially addressed role dynamism in their paper on learning the structural dynamics of roles [9]. Though Rossi's work is very general and does not restrict itself specifically to users and social networks, Rossi's use of the term "role" to refer to a pattern of behavior is consistent with our own. Rossi focuses on the temporal and evolutional aspects of roles, which moves the state of the art away from the idea that roles are static over time. While our work is differentiated by our interest in the dynamics of roles across communities, it is similar in spirit in that we also reject the notion that roles are static, either temporally or across communities. Rossi's research into temporal dynamics combined with the cross-community research described herein could represent a more refined and accurate model of the ways in which users interact for the future.

Despite this extensive body of work, it is worth noting that such efforts are often hindered by the small portion of online populations that actually interact often within online communities. This phenomenon is colloquially known as the "1% Rule," or the 90-9-1 principle, and states that only 1% of users actually contribute new content to these communities (with 90% lurking, and 9% providing edits, up votes, likes, or similar behavior). While a relatively informal rule, recent research has shown behavior consistent with this tenet across four separate social networks [14]. One should then expect to find a relatively limited number of users who participate in multiple communities.

3. METHODS

To reiterate, our objectives are to answer the following three questions: 1) Is the answer-person role present in reddit? 2) Can we automatically identify this role? 3) Is there significant multi-community participation?

To answer these questions, we first needed to gain access to data describing the social interactions between reddit users in a way that supported reconstructing the underlying social networks. Once we could build these networks, we then performed manual analysis on a subset of reddit users to determine whether the target role existed and developed a feature set that would allow us to learn and identify the structure. From there, we could then build a machine learning algorithm from these hand-labeled feature sets to identify the answer-person role across our entire data set.

3.1 Gathering the Data

Since no such repository of reddit user interaction data currently exists, nor is there an easy means of retrieving this data, we had to develop our own data collection methods. For these reasons, we built a toolkit to crawl specific reddit communities, gather posting statistics on specific users, and extract this information [3]. reddit also provides support for third-party browsers and tools through an API, which several entrepreneurial developers conveniently wrapped in a Python package called PRAW [4]. Through this wrapper, we were able to access submissions from any given subreddit and extract the associated reply threads.

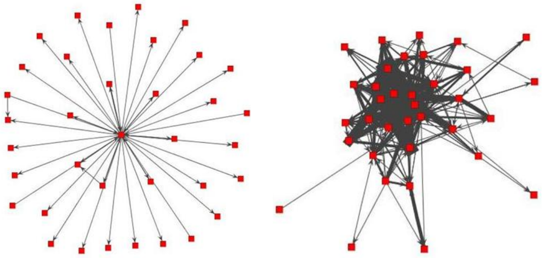
reddit is a sizable and active community, rendering it too large to capture fully given the many contingencies in place to support the site's load balancing and throttling needs. To support these availability requirements, reddit imposes strict limitations on request frequency in its API (no more than two requests per second). Furthermore, reddit limits the number of objects the API will return for a single request (between 100-1500 depending on the type of object and user level). These constraints obstructed high-speed data gathering, so to gather our data, we sampled from reddit's listing of the top replies to the top submissions for the month.

Since our research questions presuppose the answer-person role's existence in reddit, we first targeted several question-and-answer-based communities where such people might be most prevalent. This line of research identified a few users who had taken part in "IAmA" events, or events in which someone identifies some particular characteristic, trait, or occupation about himself (e.g., an actor, scientist, director, or a specific person) and invites reddit users to ask questions. These events are hosted by people from many walks of life and varying degrees of fame, from President Barack Obama to the team controlling the Mars rover Curiosity to a cook in a seafood restaurant. Users who conduct these question-and-answer sessions tend to fit the answer-person role well, which provides a good baseline for further analysis. To vary our data and explore communities whose formats might differ from the question-answer behavior of the "Ask*" communities and to ensure negative samples (or non-answer people), we also identified a collection of discussion-oriented communities for inclusion. These subreddit communities included: AskScienceDiscussion, AskMen, AskScience, AskWomen, CompSci, DesMoines, IAmA, MachineLearning, Movies, MyLittlePony, PersonalFinance, TalesFromTechSupport, and WashingtonDC.

To obtain a sufficiently large data set, we looked at the top 100 submissions from the month of July, 2013 and commenters' behaviors within those submissions. For each submission, we extracted on average the top 200 comments and put their authors into a graph structure supported by the NetworkX [5] framework [6]. From these submissions and their associated comments, we created a separate directed graph G for each subreddit. For each graph $G = \{V, E\}$, nodes in set V corresponded to users who either posted a submission or a comment, and the directed edges between these nodes in E corresponded to comment replies directed from author to recipient. Edges in G were weighted by the number of responses between the two users. We then exported these graphs to the Graph Exchange XML Format (or GEXF) for analysis and visualization in the Gephi[6] toolkit.

3.2 Identifying Role-Consistent Users

After constructing these networks, we then explored each graph to determine whether it contained nodes conforming to the answer-person role. To identify nodes with this role, we inspected the shape of the 1.5-degree egocentric networks of the higher-degree nodes to determine the similarity to the visual structural signatures identified by Wesler et al. in 2007 [16]. Such nodes were characterized by a hub-and-spoke pattern with the target node in the center with many edges radiating out toward nodes that otherwise have few connections, like that shown in Figure 1a. This structure is in contrast to a non-answer-role, such as a discussion person shown in Figure 1b.



(a) Answer Person (b) Discussion Person

Figure 1: Exemplary Role Signatures

As we searched through these high-degree nodes in these graphs, we hand-labeled each vertex indicating whether the corresponding user conformed to the answer-person role. reddit is not without other social roles, however, as several of the discussion-oriented subreddit communities contain non-answer-person roles. There also exists ambiguous structural signatures that seem to be hybrids of the answer-person and non-answer-person roles (which is perhaps why Wesler et al. chose a continuous scale for answer-role conformance rather

[3] This code is available on GitHub at https://github.com/cbuntain/redditResponseExtractor

[4] https://github.com/praw-dev/praw

[5] http://networkx.github.io/

[6] http://gephi.org/

than the binary label used here). In these instances, we count these vertices as non-answer-person roles.

After collecting a sizable node set containing both answer-person and non-answer-person users, we culled the remaining nodes by removing those with less than 20 outgoing edges (note that we chose 20 as our threshold value to ensure statistical significance). For the remaining nodes, we calculated a collection of metrics on each node's ego network similar to those used by Wesler for use as feature vectors in the learning process. First, we calculated undirected network density (Eq. 1) since ego networks for answer-person users should have lower density than non-answer/discussion person users given the sparse interactions between nodes. Low degree distribution (Eq. 2) was another metric we used to determine a smoothed ratio on the number of rarely interacting neighbors where answer-person users should have more such neighbors than other roles. Similarly, the proportion of neighbors with low degree (Eq. 3 where v_e is the center node of the ego network) also provided some insight into neighbor interaction sparsity. Proportion of intense ties (where "intense ties" are defined as ties with weights greater than 1) (Eq. 4), clustering coefficients [10], and triangle density were several other metrics we used to gauge how much repeated interaction occurred within the target user's ego network. We define triangle density as the number of triangles present over the maximum number of triangles, shown in Eq. 5 where G_e is the ego network of vertex v_e, and triangles(G_e, v_e) is the number of triangles in that graph. For completeness, we also included features for the source community, so we could determine whether role behavior is dependent more on the community than the network's structure.

$$\frac{2|E|}{|V| \cdot (|V| - 1)} \tag{1}$$

$$s = |v \in V \text{ s.t. degree}(v) < 4|, \frac{s+1}{|V| - s} \tag{2}$$

$$succ(v_e) = \{v' \in V \text{ s.t. } (v_e, v') \in E\},$$

$$\frac{|v \in succ(v_e) \text{ s.t. degree}(v) < 2|}{|succ(v_e)|} \tag{3}$$

$$\frac{|v \in succ(v_e) \text{ s.t. weight}(v) > 1|}{|succ(v_e)|} \tag{4}$$

$$\frac{2 \cdot \text{triangles}(G_e, v_e)}{\text{degree}(v_e)(\text{degree}(v_e) - 1)} \tag{5}$$

3.3 Automatically Classifying Roles

Moving forward, we then leveraged the Scikit-Learn Python library [7] to construct a supervised classifier capable of predicting these answer-role labels with acceptable accuracy. While we tested a few different learning algorithms, we settled on a standard decision tree (d-tree) to facilitate inspection of the resulting decision rules. We trained a set of 100 d-trees using 85% of our labeled data set (which was drawn randomly from the full set) and evaluated accuracy on the remaining held-out data.

For a subset of those users who we classified as answer people, we attempted to capture their posting behaviors across other communities to determine whether they assumed the same role in these other communities. Though reddit's API made this task difficult in that it is difficult to extract a user's posting behavior without making many requests to the reddit servers, we again leveraged PRAW to capture the subreddit communities in which these users most often participated. From there, we could at least determine whether the target user was a prevalent participant in the new community, and if so, we could use our existing infrastructure to capture the top submissions and posts from that new community. We then built feature vectors using the metrics described above for these additional communities and ran them through our classifier to determine whether the original user's role was conserved across communities.

4. RESULTS

4.1 Data Sets and Role Distribution

We captured 279 unique users across 10 subreddits who had more than 20 outgoing edges. The remaining three subreddits (AskScienceDiscussion, CompSci, and TalesFromTechSupport) lacked users who met our thresholds, so those communities were discarded. Of these 279 users, seven appeared in more than one subreddit community, yielding a total of 286 vertices across 10 graphs. As mentioned previously, four of these dual-subreddit users were held back from the training set, so our training set contained 275 labeled samples. This small user set is unlikely to produce statistically significant results as to whether roles are conserved, but it does provide some insight into the fraction of users who participate in multiple communities. Figure 2 depicts the number of users we analyzed in each subreddit (blue bars) as well as the number of subscribers in each subreddit in a log-10 scale (red line).

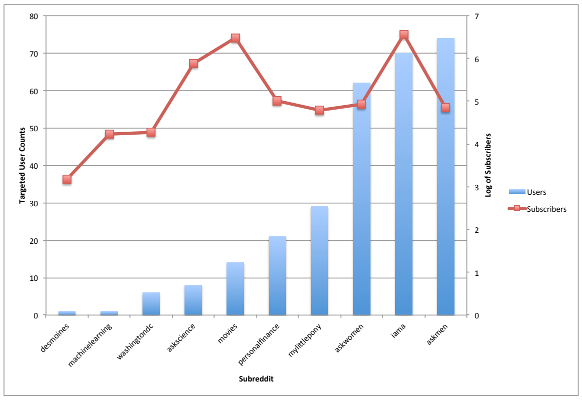


Figure 2: Users and Subscribers per Subreddit

For each vertex in each subreddit graph, we visualized the corresponding 1.5-degree ego network per the methodology we described. This inspection answered the first research question of whether the answer-person role exists in the reddit community with a definitive "yes." In fact, Mark Shuttleworth, the founder of Ubuntu's Canonical, hosted a popular IAmA submission, and the resulting visualization of his ego network (Figure 3a) presents a nearly identical structural signature to that proposed by Wesler et al. Our supposition that subreddits like IAmA would contain many

[7] http://scikit-learn.org/

such answer-person examples turned out to be quite correct. In contrast, Figure 3b illustrates a classically non-answer-person role taken from a prevalent user in the MyLittlePony subreddit.

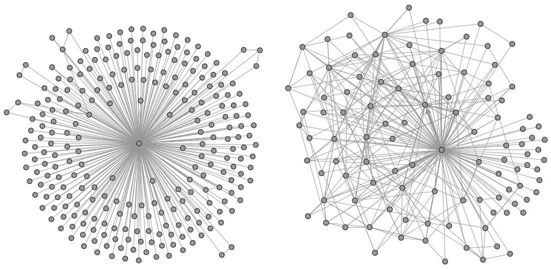

(a) Answer-Person (b) Non-Answer-Person

Figure 3: Role Structures

After manually inspecting and labeling each vertex, we found that the distribution of answer-person and non-answer-person roles were relatively evenly distributed across our data sets. Approximately 150 users were labeled with the answer-person role, and around 130 were labeled with the non-answer-person role. Figure 4 shows the proportions of these roles distributed across the subreddits (red and blue bars) as well as the absolute total of answer people in each subreddit (green line).

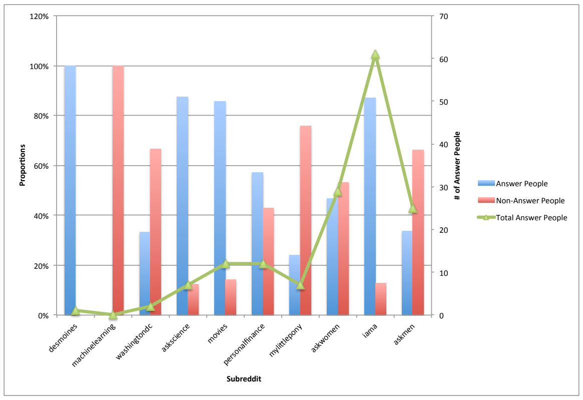


Figure 4: Roles per Subreddit

4.2 Social Role Decision Trees

After labeling each vertex, we then trained 100 instances of Scikit's decision tree algorithm using the features described in Section 3.2 and 85% of the entire data set across all subreddits and evaluated accuracy using the remaining 15%. This set of trees achieved accuracies with a minimum of 0.66, maximum of 0.92, and mean of 0.7966.

An interesting question, however, is the predictive quality of the subreddit in which a user posts; that is, can we determine a user's most likely role based only the community in which she interacts? Intuition would suggest that active users in the Ask* subreddits are likely to adhere to the answer-person role. To explore this possibility, we trained a similar set of decision trees based only on the user's subreddit to determine whether that contributed more than the network structure properties we explored above. This set of trees achieved accuracies with a minimum of 0.48, maximum of 0.84, and mean of 0.6854, which indicates the parent subreddit is not as predictive as social structure.

5. DISCUSSION

To review, we first sought to determine whether the answer-person role exists in the reddit web-based community. If so, can users who assume this role be identified by characteristics of the social network structure rather than with content analysis? Lastly, do users of these multi-community sites participate significantly in across community boundaries?

5.1 Existence of the Answer-Person Role

Concerning whether the answer-person role exists in reddit, our results clearly indicate the affirmative. Communities such as the IAmA subreddit include scientists from the Mars Curiosity Rover Mission, the actor Sharlto Copley (from movies like "District 9," "The A-Team," and the new "Elysium"), and Colin Mochrie from "Whose Line Is It Anyway?" who post specifically to answer questions and therefore almost intrinsically satisfy the conditions of the answer-person role. Beyond these almost trivially satisfying subreddits, we do see the answer-person role present in other sub-communities as well; many users in the Movies and WashingtonDC subreddits also assume this role for example.

5.2 Classifying the Answer-Person Role

Through Welser's metrics in conjunction with other network structures provided by the NetworkX package, we were able to positively identify on average 80% of the answer-person users. This result is consistent with the existing literature in that expensive content-based analysis is not necessary to identify this specific role. Quickly and automatically identifying answer-person users in this way could facilitate content promotion in subreddits or communities where answers are prioritized over discussion (perhaps in technical support communities). Furthermore, users adhering to this role could be automatically alerted when questions specific to their areas or communities are posted to decrease the time between question and answer.

One potential issue in this investigation was whether the subreddit communities from which we drew our data would unduly influence these roles based on the purpose of the source community. For instance, as we have discussed, the IAmA subreddit has a particular question-and-answer format that basically forces user behavior into the answer-person role. Our results from trees trained on structural features versus those trees trained on affiliation (around 80% versus 69% respectively) suggest that users' structural features are more predictive than the source community. This result lends some weight to the notion that user behavior might vary between communities even if the communities are similar. While an answer person may be an answer person in multiple communities, even if she participates exclusively in one type of community, there may be no guarantee she will exhibit the same behavior across *all* communities.

5.3 Multi-Community Interactions

It seems users tend to avoid prolific interactions across distinct communities. It is exceedingly rare to find a user who posts often in something like the AskScience subreddit *and* the Movies subreddit (though someone posting in AskScience may be slightly more likely to be a major contributor in AskPhysics or ExplainLikeImFive). Our data supports this result in that only seven of our 279 unique users, or around 3%, were identified in multiple communities. reddit also has an interesting dynamic in that many users who post submissions in larger attention-gathering subreddits such as IAmA create "throwaway" accounts for the sole purpose of answering questions without drawing additional attention to their regular activities, which may hinder tracking behavior across multiple communities. Fortunately, the answer role plainly exists in communities where throwaway accounts are less prevalent, providing some measure of insight into cross-community interactions and roles.

These results are seemingly consistent with the intuition provided us by the 1% rule in that only a small fraction of the population that exhibits significant participation will also participate across multiple communities. We should then expect approximately 1% of the 1% to exhibit this behavior, which suggests that user interactions in voluntary social networks (e.g., Facebook, reddit, Twitter, etc.) will follow a similar pattern of limited multi-community interaction. That is, without incentives for users to participate in multiple communities, only a small fraction of the already small fraction of significant users are likely to exhibit this behavior. In non-voluntary networks that develop in organizations with boundaries between units or departments, however, this behavior may differ significantly. For instance, networks in which a participant has responsibility in multiple segments of the community (as with a manager and her interactions with subordinates and upper management) might require multi-community interaction.

6. CONCLUSIONS

Multipurpose, digital communal spaces similar to reddit continue to experience rapid growth and popularity. While a significant portion of existing social networking literature is applicable to these new spaces, much of it assumes consistent user behavior throughout the network. This research builds the foundation for further analysis of this assumption by identifying well-known behavioral patterns and social roles in these multi-community environments. We clearly demonstrated the presence and identifiability of the answer-person role in reddit and showed that only a very small number of users participate across community boundaries. We have yet to show, however, whether users that *do* cross boundaries behave consistently across all the communities in which they interact, leaving the door open for future work that could enhance community detection and user classification.

7. REFERENCES

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 529–535, New York, NY, USA, 2003. ACM.

[2] J. Donath, K. Karahalios, F. Viegas, and A. Street. Visualizing conversation. *Journal of Computer-Mediated Communication*, 4(4):0, 1999.

[3] D. Fisher, M. Smith, and H. T. Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03*, HICSS '06, pages 59.2—-, Washington, DC, USA, 2006. IEEE Computer Society.

[4] S. A. Golder and J. Donath. Social roles in electronic communities. *Internet Research*, 5:19–22, 2004.

[5] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 645–654, New York, NY, USA, 2008. ACM.

[6] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using {NetworkX}. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008.

[7] X. Hu and H. Liu. Social status and role analysis of palin's email network. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 531–532, New York, NY, USA, 2012. ACM.

[8] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[9] R. Rossi, B. Gallagher, J. Neville, and K. Henderson. Role-dynamics: fast mining of large dynamic networks. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 997–1006, New York, NY, USA, 2012. ACM.

[10] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, J. Kertesz, and J. Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):27105, Feb. 2007.

[11] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405, New York, NY, USA, 2011. ACM.

[12] R. Tinati, L. Carr, W. Hall, J. Bentwood, and V. Street. Identifying communicator roles in twitter. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1161–1168, New York, NY, USA, 2012. ACM.

[13] V. H. Tuulos and H. Tirri. Combining Topic Models and Social Networks for Chat Data Mining. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 206–213, Washington, DC, USA, 2004. IEEE Computer Society.

[14] T. van Mierlo. The 1% rule in four digital health social networks: An observational study. *J Med Internet Res*, 16(2):e33, Feb 2014.

[15] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011. ACM.

[16] H. T. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32, 2007.