Department of Computer Science
George Mason University
Technical Report Series

4400 University Drive MS#4A5
Fairfax, VA 22030-4444 USA
http://cs.gmu.edu/
703-993-1530

Digging Digg : Comment Mining, Popularity Prediction, and Social Network Analysis

Salman Jamali sjamali@gmu.edu
Huzefa Rangwala rangwala@cs.gmu.edu

Technical Report GMU-CS-TR-2009-7

Abstract

Using comment information available from Digg we define a co-participation network between users. We focus on the analysis of this implicit network, and study the behavioral characteristics of users. Using an entropy measure, we infer that users at Digg are not highly focused and participate across a wide range of topics. We also use the comment data and social network derived features to predict the popularity of online content linked at Digg using a classification and regression framework. We show promising results for predicting the popularity scores even after limiting our feature extraction to the first few hours of comment activity that follows a Digg submission.

1 Introduction

The past decade has seen a massive rise in web services and applications that allow users to create, collaborate, and share varied forms of data like articles (web-blogs), pictures (Flickr.com), video (Youtube.com), and status updates (Twitter.com). Social bookmarking websites like *Delicious.com*, *Slashdot.org*, and *Digg.com* allow users to submit links to web content they find interesting along with a short description (referred as stories in this work). Every user in these online communities can provide comments for the posted content (initiating discussions), and also rate the articles that they find interesting. Thus, social bookmarking sites serve as data aggregators, web-based discussion forums, and an online collaborative filtering system that can collectively determine popular online content.

Recently, there have been several studies [9, 1, 12] that have analyzed social networks generated from comment interaction between users. In this work we model a co-participation network similar to the co-authorship and citation networks [10, 11] where users are linked together if they comment on the same discussion thread or submitted story. This implicit relationship between users based on comment information provides an understanding of the complex underlying community structure. We use egonets [17] to capture the local neighborhoods of users within the derived social network, and provide an understanding of the community with multiple interests. We further extract several user-based and comment-based features, and train classification and regression models for predicting popular stories. We evaluate our methods to use features derived from comments that were posted within the first few hours of posting the story. Successful prediction of popular content, allows users to sift through the vast amount of available online data and can also aid in the ranking algorithms pursued by social bookmarking websites.

For our analysis, we use Digg (founded in 2004), a popular social bookmarking website that allows users to share, comment, and rate on diverse online available information. We found that the user community within Digg was highly active in posting comments and found their focus to be spread across a wide range of topics ranging from world business to entertainment. We also showed the ability to predict the popularity index using early available comment and user based features.

The rest of this paper is organized as follows. In Section 1 we provide a brief literature survey. The Digg dataset is described in Section 2, and the definition as well as analysis of the co-participation network is described in Section 3. Section 4 discusses the predic-

tion of story popularity and we provide conclusions in Section 5.

2 Related Work

USENET was one of the first web based message forum developed in 1979 and has seen several works related to development of tools for visualizing the structure of the discussions within these forums [8]. Statistical analysis methods [18] and network analysis [20] methods were developed to understand the characteristics of the different discussion forums.

Recently, researchers have used comment information to define implicit relationships between users, and then used social network analysis methods to understand the characteristics and interaction patterns of several communities and groups [12, 3]. Implicit relationships or links are defined between users who comment or reply on discussion threads to a particular user [12]. Within the context of individual web-blogs, a relationship was defined between the author of the blog and the commenter [3].

Our work is closely related to the analysis of the community participating in the Yahoo Question and Answer forum (Yahoo QA) [1]. In case of the Yahoo QA forum a user posts a question and several users provide an answer which are rated by the community. The work analyzed the interaction patterns between the various users belonging to multiple categories. An interaction or relationship was defined as a directed edge between the user who initiated a question and the users who replied with an answer. Using egonets [17] to characterize the local neighborhood of users within the derived social network, differences in the interaction patterns between users belonging to the technical and advice forums was observed. In our work, we define a weaker undirected interaction between two users who comment on the same story.

Recently, a social network was modeled [9] for the user community in Slashdot (another online bookmarking site). The implicit relationship was defined similar to the reply-answer network above, where an edge was defined between users who would comment directly to a posted comments. Thus, if user A posts a comment, and user B replies to the comment, a relationship exists between users A and B. However, if a user C comments to the story but not to $A's$ comment then there exists no relation between user A and C. Our definition of the implicit relationship between user follows the more traditional definition in co-authorship network [10, 11] and will results in relationships between the three users A, B, and C in the above example.

3 `Digg` Dataset

`Digg` [1] is one of the most active social bookmarking website where registered users submit links, news articles, videos, and images along with an optional short description. Submissions can lead to a discussion amongst the registered users who may post a series of comments regarding the material posted. A registered `Digg` user can rate the submissions (referred to as stories in this work), and support the stories that they find interesting by providing a positive rating referred to as a *digg*. On the other hand users can also provide negative rating known as a *bury*. Using the collaborative effort of millions of registered users, stories get rated to have a `Digg-score` (sum of *diggs* minus sum of *bury*) which serves as a popularity index. The exact algorithm is not revealed, but stories that achieve a high `Digg-score` from a diverse group of users are promoted to the *popular* section of `Digg` [14].

Users also have the option to provide a rating for the individual comments. A positive rating for a comment is a *up* score whereas a negative rating is a *down* score.

We used the `Digg` API to crawl 37185 popular stories from November 16, 2007 to March 10, 2009. The total number of comments in our dataset are 6188266, and the total number of users who posted at least one comment are 253846. The `Digg-score` for the crawled stories ranged from 86 to 37947 with a mean of 1204 and a standard deviation of 1122. The average number of comment made by a user is 24.

Stories at `Digg` are classified hierarchically into two levels, namely eight *categories* and 51 *topics* within the different categories. The eight categories include (i) World Business, (ii) Technology, (iii) Science, (iv) Gaming, (v) Sports, (vi) Entertainment, (vii) Life Style, and (viii) Offbeat. There were a total of 51 topics when we crawled the data. Examples of topics include "Apple", "Microsoft", and "Linux" within "Technology","Football" and "Basketball" within "Sports", and "2008 US Elections" (one of the most popular topic) within "World Business". At the time of this writing however the topic "2008 US Elections" was no longer present. Table 1 provides general statistics about the dataset divided across the eight categories. The table shows the number of stories (S), total number of users who at least commented (U) once, and the average number of comments per story within the eight categories.

We also assign a user membership to one of the eight categories. This is done by assigning the user the category where he/she comments the most. In Table 1 we report the total members per category (M). We similarly assign a user to belong to one of the topics within the

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

Table 1: Digg Dataset Statistics.

Category	S	U	M	C/S
World Business	7341	133468	84220	252
Technology	7536	117441	48567	135
Offbeat	4715	118446	51111	205
Entertainment	3850	90414	19634	150
Science	4924	82575	14765	113
Lifestyle	4221	93161	16465	143
Gaming	2399	69110	13331	177
Sports	2199	51257	5753	90

S denotes the total number of stories within the categories. U indicates the total number of users who commented at least once for the stories within the categories. M indicates the total number of users assigned to the categories (members). C/S denotes the average number of comments per story within the category.

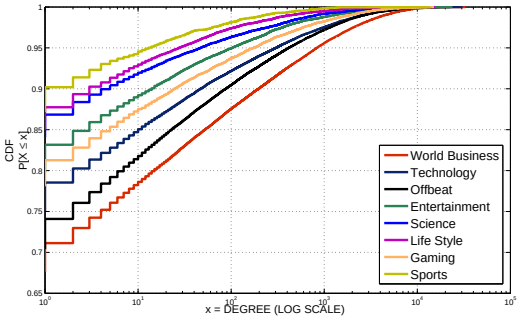

Figure 1: Distribution of degree (Log Scale).

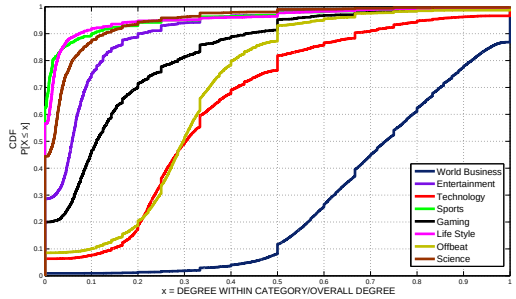

Figure 2: Distribution of the ratio of within-category degree to the overall degree.

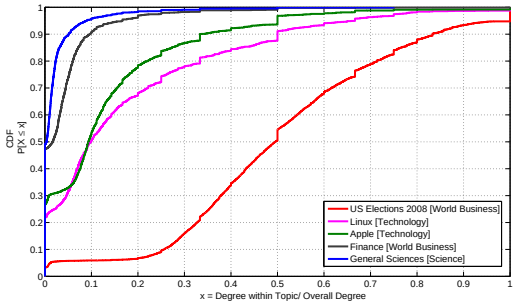

Figure 3: Distribution of the ratio of within-topic to the overall degree.

categories. From columns U and M we notice that there is a large overlap in the categories that users comment

4 User Characterization

Motivated by the work involved with co-authorship and citation networks [10, 11] we define a co-participation network to model the relationships between different users in the Digg community.

4.1 Network Description and Statistics

An undirected graph $\mathcal{G} = (V, E)$ is used to represent the co-participation network. The set of vertices V represent the set of users commenting across the different stories. The sets of edges E represent the interaction between the different users, and an edge $E_{i,j}$ exists between users V_i and V_j if the pair of users co-participate by commenting on n or more stories. We experimented with the threshold parameter n used to define the presence or absence of an edge or relationships between users. The average degree (i.e., number of edges per node) was 2414.5, 114.4, and 26.8 for threshold values of n equal to 1, 4, and 8, respectively. For the results reported here we use a threshold value of $n = 4$ i.e., a pair of users are considered to be connected if they comment on at least four same stories.

A pair of users commenting on the same story may have differing or even opposing views. In the future, we aim to refine our relationship definition between the users based on the polarity of the comments i.e., perform sentiment analysis or opinion mining [13, 7] using text information of the comment.

4.2 Degree Distribution

In Figure 1 we present the cumulative distribution (CDF) of the degrees per user separated based on membership to one of the eight categories. Nodes with more degrees indicate the user participating with several other users. From the Figure 1 we observe the difference between the users in the eight categories. The degree is plotted using a logarithmic scale and indicates a heavy tailed distribution, referring to the high levels of co-participation activity for the various categories. The category "World Business" shows the highest participation amongst the users as seen from the CDF plot. This is primarily because of the high level of activity in terms of posting stories, and discussions due to the 2008 US Presidential Elections (a

topic under World Business), a popular topic when we downloaded the data. We can also see the differences between the other categories.

A user was assigned a category membership based on the category in which he/she would post the maximum comments. A user was free to comment across various categories, and though we compute the degree per user and analyze by category, we do not restrict the neighbors to be in the same topic or category.

4.3 Egonet Analysis

We also use egonet analysis to understand the relationships amongst different users within the different categories. Such an egonet analysis was done previously [17, 1] to differentiate between community of users that were discussion prone or not. A one level egonet for a user is defined as the user, the set of users who interact directly with the user (neighbors), and the relationships between those users. We can extend the definition of egonet to have neighbors who are N hops (links) away from the user in consideration.

In Figure 4 we show the egonets for a set of four active users within each of the eight categories. For each category we identify the most active users, i.e., users who have commented the most on stories posted within a particular category. Figure 4 shows only the 1st, 20th, 80th, and 100th most active users per category. We have upto 200 egonets per category at the project website `http://www.cs.gmu.edu/~mlbio/digg-ego/`. The egonets we present have a two color coding scheme where a co-participation edge $E_{i,j}$ is colored black if both V_i and V_j have the same category membership, whereas edge V_i and V_j is colored green if users V_i and V_j belong to different categories.

We have ordered the categories from top to bottom in decreasing order of the densities of egonets. The egonets of users in categories like "Sports", "Gaming", and "Life Style" (Figure 4 (f)-(h)) have smaller and less denser neighborhood in comparison to categories like "World Business", "Technology", and "Offbeat" (Figure 4 (a)-(c)). The dense nature of egonets for the "World Business" category can be explained by the large number of stories that became popular due to the 2008 US Elections. From this data we can also infer that within the `Digg` community stories within the Sports and Gaming categories do not lead to large user interaction and discussion. The egonets also suggests that users within "Technology" participate by way of commenting in much larger volumes in comparison to "Science".

The egonets for the "World Business" category (Figure 4(a)) show more black edges in comparison to the corresponding egonets for the "Offbeat" category (Figure 4(c)). This suggests that the "World Business" community users are focused and involved with discussing stories that are posted in that category. The "Offbeat" category is a collection of diverse topics (e.g., "comedy", "pets") that do not fit within the other seven categories. As such it is expected that users within the "Offbeat" category are loosely coupled i.e., not focused to comment in the same categories as their category membership indicates. We observe that users in the "Offbeat" category also comment on stories posted in the "Life Style" and "Entertainment" categories.

In Figure 2 we show the cumulative distribution function for that ratio of in-category degree to the overall degree. The in-category degree for a node is the number of one-hop neighbors who have the same membership as the user in consideration. The figure provides complementary results to the ones observed for the hundred most active users seen in the egonet analysis by allowing us to see the percentage of users below a specific value of the in-category ratio.

In Figure 3 we show the cumulative distribution for the ratio of the in-topic degree to the overall degree corresponding to five selected topics within the eight categories. It is interesting to see that the users who comment within "2008 US Elections" topic are highly topic-focused in comparison to the "Finance" topics, both within the "World Business" category. The "Technology" topics "Apple" and "Linux" also have a high degree of in-topic focus. In both the Figures 2 and 3 we neglect users having an overall degree of zero. This does not have an effect on the trends observed and allows us to focus on the users with at least a single neighbor.

4.4 User Membership Analysis

As discussed in Section 2 and observed in Table 1, users within the `Digg` community have overlapping interests and as such participate and comment across multiple areas of interest.

As done previously [1], we computed an entropy measure to capture the focus of the user. Users commenting within a large number of categories in comparison to a user commenting across a fewer number of categories would have a higher entropy and less focus. As such, we can define the entropy for the user with respect to the categories as $H_1 = -\sum_i p_i log(p_i)$ where i iterates over the eight categories and p_i denotes the probability for the user to belong to category i.

Similarly, we can compute an entropy measure for the user with respect to the 51 topics given by $H_2 = -\sum_j p_j log(p_j)$. The sum of the H_1 and H_2 represents the total hierarchical entropy for a user. Using such a two-level hierarchical entropy definition allows us to differentiate between users who would comment on a diverse set of subcategories within a single category (less

entropy because H_1 will be low) and users who would comment on a diverse set of subcategories spread across multiple categories (higher entropy because H_1 will be high).

In Figure 5 we show the hierarchical entropy distribution for the 70753 users who commented at least ten times and 27645 users who commented at least forty times. Computing the entropies for users who comment very few times would bias the analysis (entropy for a user who comments once will be zero). A high percentage of users have a higher entropy i.e., in between 5.0-7.0 and 6.0-7.0 for users that commented at least 10 times and 40 times, respectively. It can be inferred that users have a tendency to participate and comment across multiple discussion topics. This suggests that the user community in Digg is not very focused but this could be due to loosely defined categories and subcategories (called topics). We also use the entropy measures of a comment author as features for predicting the popularity of a story.

5 Predicting Popularity of Stories

Using the comment information associated with posted stories at Digg we predicted the popularity of a particular story, the Digg-score (See Section 2). We wanted to develop a predictive model that would be able to accurately infer the Digg-score using associated comments made by the user community but restricted to the first few hours of the initial posting of the story. As such we trained predictive models using comments only from the first ten hours and the first fifteen hours. We also trained models using all the available comment information for comparison purposes.

5.1 Methods

The prediction was performed by setting up three independent classification problems: (i) a 2-class, (ii) a 6-class, and (iii) a 14-class prediction problem. The bins for the 6-class and 14-class prediction problems were set in Digg-score intervals of 1000 and 500, respectively. For the 2-class prediction problem we split the instances into the first class having all stories with a Digg-score of less than 1000, and the second class with Digg-score greater than 1000. This allowed for a uniform size distribution split. We used the decision tree classifier [16], the nearest neighbor classifier [2], and support vector machines [15] for performing the classification. In this work we present the classification results for the pruned C4.5 decision algorithm denoted by DT, the nearest neighbor classifier using nine neighbors denoted by 9-NN, and the support vector machine classifier using the linear and radial basis kernel function denoted by SVM (L) and SVM (R), respectively. For the K-class classification using SVMs, we trained K binary one-versus-rest classifiers for each of the K classes. We also estimated the Digg-score using ν-SVM regression method [6] (denoted by K=∞).

5.1.1 Feature Description

We used several comment-based and user focused features for predicting the Digg-score . These features capture different aspects of the data and are described in detail below:

Comment Statistics We use the number of comments for a posted story, and the average word length of all the comments as features. A large number of posted comments are directly correlated to high level of user interest and hence, the popularity of stories. Both these features have shown success in predicting the best rated answer for a question within the context of Yahoo's QA Dataset [1].

When users comment they may chose to reply to a specific comment or make a new comment. This posting of comments induces a hierarchical tree structure where we can associate a level with every comment. A comment directly made to the story is considered as the first-level comment and can be thought of as initiating a thread. Analyzing our dataset we observed that a large number of levels are indicative of controversial stories. Controversial stories also have a tendency to be popular as seen in the analysis done Slashdot [9]. Motivated by this we used four features that simply count the number of comments at the first, second, third, and fourth levels for each story.

Digg User Interest Peak We captured the peak in user interest by determining the increase in user level activity within a fixed time span. We denote this feature as burst(X), and is computed as the highest comment activity seen when sliding a window of "X" hours across all the posted comments for a story. Specifically, we can represent burst(X) as

$$burst(X) = \max_{T=0\ldots T'} C(T \ldots T+X)/C(0 \ldots T') \quad (1)$$

where X is the burst window span, T' is the total time since the story was submitted to the community, and C(x) is the number of comments within the x hours. We compute burst(3), burst(4), and burst(5) as three features for predicting the Digg score. The use of different windows captures different types of user interest for a story. A higher burst weight with a shorter time span carries more information towards the popularity prediction.

Digg User Feedback Digg users also have the option of rating the comments. As such, comments that are irrelevant with respect to the posted story or are spam get negative feedback. Comments that are relevant and even cause of controversy are seen to get positive feedback from the user community. For each comment we obtain their *up* score (positive feedback) and the *down* score (negative feedback). We derive two features that sum the "up" scores for all the comments and sum the "down" scores for all the comments associated with the story. Generally, for popular posts it was seen that the sums of up scores were higher than the sum of down scores. A similar comment feedback measure was shown to be positively correlated with the popular stories posted on Slashdot [9].

User Community Structure and Membership We also compute features that are focused on the egonets (described in Section 3.3) of the users with highly rated comments. We define top five comments per story as those comments having the highest comment score given by the difference of ups and downs. For each of these highly rated comment authors we use the degree or number of local neighbors (defined in Section 3.1) as a feature. This results in five features that use local information from the social network.

We also use the two entropies (H_1 and H_2) computed for the categories and topics(Section 3.4) as a measure of knowledge associated with commenter. We use the average entropies for all the comment authors as features, and believe that this captures the knowledge-base and involvement of a user which would be important for predicting the Digg-score .

Overall we use eighteen features to train our prediction models. For training models using the first ten hours, and the first fifteen hours after the story posting we recompute the features. We standardize the feature values by centering around the mean.

5.2 Classification Results

Table 2 shows the classification and estimation results for the different class definitions, and using the comment features extracted for the first ten hours, first fifteen hours, and the complete data. We report a small sampling of the experiments we performed. In particular, we used the default parameters (regularization, width) for the SVM based methods, and report results only for the nine nearest neighbor classifier that showed the best prediction results.

To evaluate the performance of the classification and regression methods we performed 5-fold cross validation. The classification performance was evaluated using the the K-way classification accuracy (Q_K), the area under the receiver operating characteristics curve [4] (ROC), and the F-score (F1). The ROC measures the area under the plot of true positive rate versus the false positive rate, whereas the F1 provides a weighted average between precision and recall. We report the correlation coefficient (CC) between the actual and predicted Digg-score for evaluating the regression results. We used the Weka Toolkit [19] and LibSVM [5] for the popularity prediction.

Firstly, we noticed that the two most discriminative features were the number of comments per story, and the sums of the ups (not shown here). These results are similar to the two similar works related to retrieving the popular posts in Slashdot [9], and predicting the best answer in the Yahoo QA dataset [1].

Analyzing Table 2 we observe that there is a slight improvement in the use of SVM based methods in comparison to the nearest neighbor and decision tree methods as the number of classes are increased. Solving the multi-class classification with higher number of classes is a challenging problem. The prediction performance of the classifiers and estimators when using the ten hours of data as well as the fifteen hours of data are comparable. We observe a 3.5%, 3.7%, and 1.32% decrease in the Q_2, Q_6, Q_14 accuracy when comparing the performance of prediction restricted to ten hours of data in comparison to the complete data, respectively. A similar trend is seen for the fifteen hours of data. The low loss in accuracy suggests a merit in our predictive models for identifying the popularity of posted stories. Our results also show a strong CC for the predicted and original Digg-score using the ν-SVM regression method. The linear kernel is more effective in comparison to the radial basis kernel for the regression problem.

6 Conclusion and Future Directions

In this work, we used comments to define implicit relationships between users of Digg . The users were found to participate in a broad range of topics and exhibit different interaction/relationship patterns based on their interested topics. We also used the available comment as well as information derived from the defined network to predict the popularity of content within the first ten hours of content submission. We reported a 1.0-4.0% loss in multiclass classification accuracy while predicting the popularity score using the first few hours of comment data in comparison to all the available comment data.

We believe that there is lots of opportunity in mining of comment information. We would like to refine our hidden structure by analyzing the polarity or the opinion expressed within the comments [13, 7]. Using the

Table 2: Performance for `Digg-score` Prediction.

	Ten Hours Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.83	0.80	0.80	0.72	0.63	0.62	0.64	0.41	0.41	-
9-NN	0.81	0.75	0.75	0.76	0.59	0.63	0.66	0.37	0.42	-
SVM (L)	**0.88**	**0.81**	**0.80**	0.74	0.63	0.63	0.63	0.44	0.42	**0.73**
SVM (R)	0.84	0.79	0.78	**0.79**	**0.66**	**0.64**	**0.70**	**0.46**	**0.45**	0.60
	Fifteen Hours Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.83	0.80	**0.80**	0.72	0.64	0.63	0.64	0.41	0.41	-
9-NN	0.81	0.75	0.76	0.76	0.59	0.64	0.66	0.37	0.42	-
SVM (L)	**0.89**	**0.82**	**0.80**	0.75	0.63	0.64	0.64	0.44	0.42	**0.75**
SVM (R)	0.84	0.79	0.78	**0.80**	**0.66**	**0.64**	**0.70**	**0.45**	**0.44**	0.61
	All Data									
	K=2			K=6			K=14			K=∞
Method	ROC	F1	Q_2	ROC	F1	Q_6	ROC	F1	Q_14	CC
DT	0.87	0.82	0.82	0.76	0.66	0.67	0.65	0.43	0.44	-
9-NN	0.85	0.79	0.79	0.80	0.63	0.66	0.69	0.38	0.43	-
SVM (L)	**0.91**	**0.84**	**0.83**	0.79	0.65	0.67	0.67	0.46	0.45	**0.80**
SVM (R)	0.86	0.81	0.80	**0.82**	**0.69**	**0.68**	**0.74**	**0.48**	**0.45**	0.64

DT, 9-NN, SVM (L), and SVM (R) denote the decision tree, 9 nearest neighbor classifier, SVM with linear kernel, and SVM with radial basis kernel, respectively. ROC, F1, Q_K denote the average area under the ROC curve, F1 score, and K-way classification accuracy, respectively. CC denotes correlation coefficient. We highlight in bold the methods that perform the best classification or regression. The density estimation was performed using the ν-SVR method.

polarity information we could more correctly model the relationships between commenting users. Further, we are interested in studying the evolution of communities and interests using the implicit definition of relationships and interactions derived from comments.

Acknowledgment

The authors would like to thank the Volgenau School of Information Technology and Engineering at George Mason University for providing a faculty startup grant to Dr. Huzefa Rangwala.

References

[1] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

[2] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, January 1991.

[3] Noor Ali-Hasan and Lada A. Adamic. Expressing social relationships on the blog through links and comments. 2007.

[4] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[5] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] Chih-Chung Chang and Chih-Jen Lin. Training v-support vector regression: theory and algorithms. *Neural Comput.*, 14(8):1959–1977, 2002.

[7] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 231–240, New York, NY, USA, 2008. ACM.

[8] D. Fisher, Marc. Smith, and Howard T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. *Proceedings of the HICSS, Hawaii*, 3:56–59, 2006.

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

[10] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, pages 1019–1031, 2007.

[11] X. Liu, J. Bollen, M. Nelson, and V. Sompel. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41:1462–1480, 2005.

[12] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*, 2006.

[13] Kamal Nigam and Matthew Hurst. Towards a robust metric of polarity. In *Computing Attitude and Affect in Text: Theories and Applications*, 2006.

[14] G. Szabo and B. Huberman. Predicting the popularity of online content. *Technical Report HP Labs*, pages 1–6.

[15] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[16] Geoffrey Webb. Decision tree grafting. In *In IJCAI-97: Fifteenth International Joint Conference on Artificial Intelligence*, pages 846–851. Morgan Kaufmann, 1997.

[17] Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure*, 8(2), 2007.

[18] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 257–264, New York, NY, USA, 1998. ACM.

[19] I. Witten and E. Frank. Data mining: Practical machine learning tools and techniques. 2005.

[20] Kou Zhongbao and Zhang Changshui. Reply networks on a bulletin board system. *Phys. Rev. E*, 67(3):036117, Mar 2003.

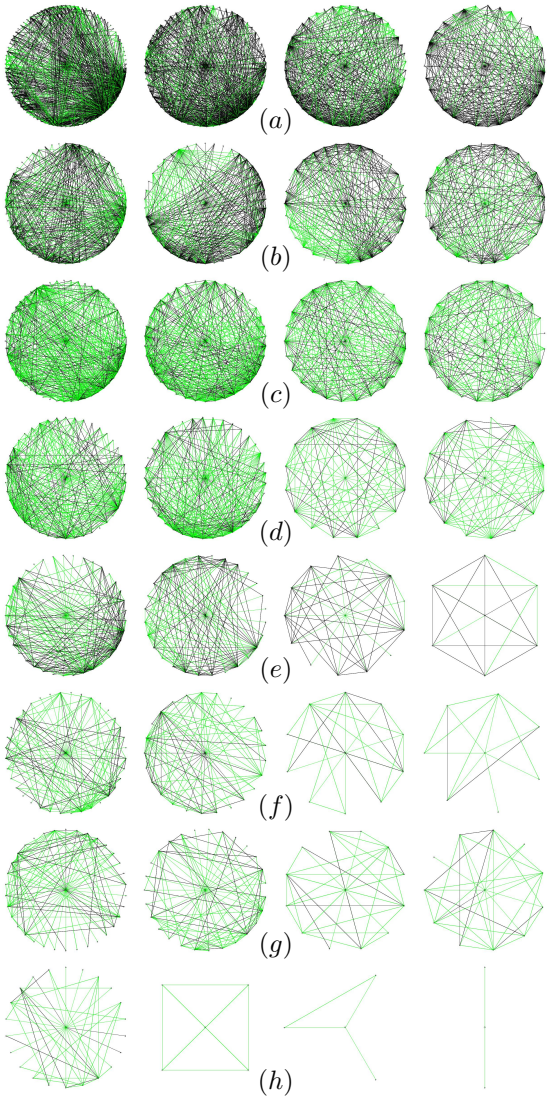

Figure 4: Egonets. Each row L to R shows egonets for 1st, 20th, 80th, and 100th most active users in Categories: (a) World Business, (b) Technology, (c) Offbeat, (d) Entertainment, (e) Science, (f) Life Style, (g) Gaming, and (h) Sports.

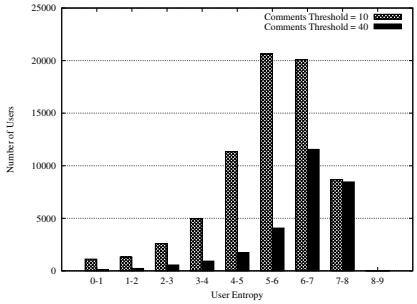

Figure 5: Hierarchical Entropy Distribution for users who commented at least 10 and 40 times.