# Co-Participation Networks Using Comment Information

**Huzefa Rangwala** and **Salman Jamali**
Department of Computer Science, George Mason University, USA
rangwala@cs.gmu.edu, sjamali@gmu.edu

## 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. We 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 also compare network properties of our co-participation network to a previously defined reply-answer network on news forums.

## 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 (Gómez, Kaltenbrunner, and López 2008; Adamic et al. 2008; Mishne and Glance 2006) 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 (Liben-Nowell and Kleinberg 2007; Liu et al. 2005) 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 (Welser et al. 2007) 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.



## 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 (Fisher, Smith, and Welser 2006). Statistical analysis methods (Whittaker et al. 1998) and network analysis (Zhongbao and Changshui 2003) 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 (Mishne and Glance 2006; Ali-Hasan and Adamic 2007). Implicit relationships or links are defined between users who comment or reply on discussion threads to a particular user (Mishne and Glance 2006). Within the context of individual web-blogs, a relationship was defined between the author of the blog and the commenter (Ali-Hasan and Adamic 2007).

Our work is closely related to the analysis of the community participating in the Yahoo Question and Answer forum (Yahoo QA) (Adamic et al. 2008). 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 (Welser et al. 2007) 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 (Gómez, Kaltenbrunner, and López 2008) 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 (Liben-Nowell and Kleinberg 2007; Liu et al. 2005) and will results in relationships between the three users $A$, $B$, and $C$ in the above example.

## 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 (Szabo and Huberman 2008). 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 categories. From columns U and M we notice that there is a large overlap in the categories that users comment

[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.

Table 2: Social network statistics for co-participation and reply-answer network.

| Network | $\|V\|$ | $\|E\|$ | $MC$ | $\langle k \rangle$ | $L$ | $D$ | $r$ |
|---|---|---|---|---|---|---|---|
| Digg Networks | | | | | | | |
| $CN_4$ | 253846 | 14519792 | 99.9% | 114.4 | 2.4 | 10 | -0.45 |
| $CN_8$ | 253846 | 3397267 | 99.8% | 26.8 | 2.3 | 6 | -0.39 |
| RN | 188494 | 3084333 | 76.0% | 16.4 | 3.8 | 10 | -0.001 |
| Slashdot Networks* | | | | | | | |
| RN | 80962 | 1052395 | 73.0% | 13.0 | 3.6 | 10 | -0.016 |

$CN_4$ and $CN_8$ denote the co-participation network with thresholds for edges set to 4 and 8, respectively. RN denotes the directed reply-answer network (Gómez, Kaltenbrunner, and López 2008). $|V|$ and $|E|$ denote the total number of nodes and edges, respectively. The other network statistics include $MC$ (the maximum cluster size), $r$ (mixing coefficient or degree correlation), $\langle k \rangle$ (the average degree), $L$ (average path length in the most giant component), and $D$ (the maximal distance between two nodes). *The results of the Slashdot network analysis were taken from the published study by Gomez et. al(Gómez, Kaltenbrunner, and López 2008).

## User Characterization

Motivated by the work involved with co-authorship and citation networks (Liben-Nowell and Kleinberg 2007; Liu et al. 2005) we define a co-participation network to model the relationships between different users in the Digg community.

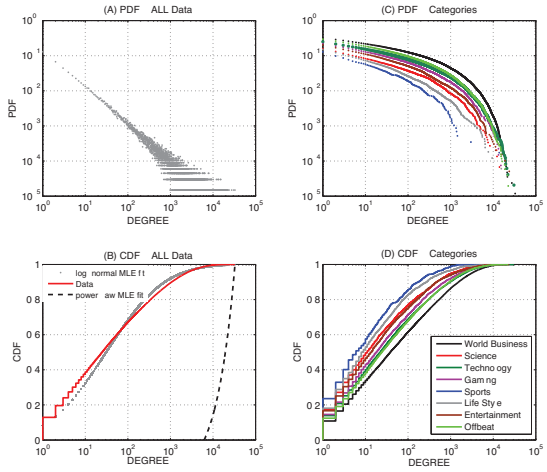


Figure 1: (A) Probability distribution function (PDF) plot of the degree for all the users. (B) Cumulative distribution function (CDF) plot of the degree for all the users along with fitted power-law and log-normal distributions. (C) Probability distribution function (PDF) of the degree for users separated based on categories. (D) Cumulative distribution function (CDF) of the degree for users separated based on categories.

## 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.

In Table 2 we report several network statistics to characterize the derived graphs (Newman 2003) for the co-participation networks defined across the `Digg` dataset for edge threshold values of 4 and 8. We also report these statistics for the reply-answer network (RN) defined in the Slashdot data study (Gómez, Kaltenbrunner, and López 2008). The reply-answer network represents a directed edge going from user $V_j$ to user $V_i$ if the user $V_j$ directly comments on a thread/comment contributed by user $V_i$. We compute the directed reply-author network for our `Digg` dataset for comparative purposes.

Table 2 shows the total number of nodes $(|V|)$ and edges $(|E|)$ for the different network representations. As expected the co-participation networks has more edges or interactions between the different users in comparison to the directed reply-answer networks. We also report the maximum cluster size $(MC)$ or the giant component size (Newman 2003). In case of the co-participation networks defined for the `Digg` data we see that 99% of the users are within a single giant cluster and are connected to each other. In comparison, 73% and 76% of the users form the giant component for the reply-answer networks defined for the Slashdot and `Digg` dataset, respectively.

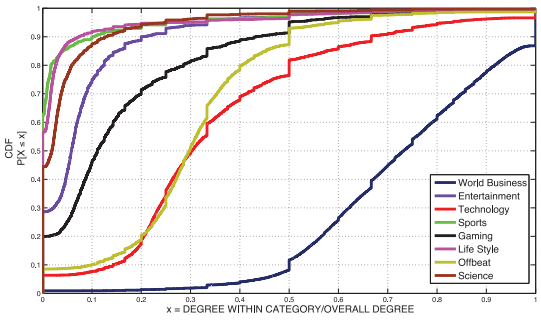


Figure 2: Distribution of the ratio of within-category degree to the overall degree.

The average degree $(\langle k \rangle)$ (i.e., number of edges per node) for the co-participation network was 2414.5, 114.4, and 26.8 for edge threshold values of $n$ equal to 1, 4, and 8, respectively. We also report the average path length $(L)$ and the maximal distance $D$ in the largest cluster across the `Digg` and Slashdot networks. The co-participation network with edge threshold weight set to 8 has a smaller maximal distance.

We also compute the degree correlation or mixing coefficient (Gómez, Kaltenbrunner, and López 2008; Newman 2003) $(r)$ that determines if highly connected users are preferentially linked to other highly connected users. Comparing the different derived networks we can see that the `Digg` co-participation networks are characterized by dissortative mixing which is not a general trend for social networks (Newman 2003). In contrast the reply-answer networks are neither assortative or dissortative.

Henceforth, for the results reported we use a co-participation network introduced here with 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.

## Degree Distribution

In Figure 1 we show the probability distribution (PDF) and cumulative distribution (PDF) of the degrees. Figures 1(A) and (B) show the PDF and CDF plots for the entire community of users, whereas Figures 1 (C) and (D) show the PDF and CDF plots for the users separated based on membership to one of the eight categories. We observe that the obtained distributions are heavy-tailed indicating a high level of heterogeneity between the users. Using the approach described in the analysis of reply-author network (Gómez, Kaltenbrunner, and López 2008) we fitted the observed data to the power law and log-normal distribution. The optimal parameters were selected using the maximum likelihood estimation (MLE) and a truncated power-law approximation. Figure 1 (B) shows the fit of the two distributions with respect to the data, and as observed in the Slashdot analysis (Gómez, Kaltenbrunner, and López 2008) the data fits the log-normal

distribution as determined by the Kolmogorov-Smirnov significance test.

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. 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.

We also analyze the egonets for the sixteen most active users in two month time windows starting from November 16, 2007 to January 9, 2009 (duration of our crawl for the dataset). In Figure 3 we present here the average degree for the users. In essence, we build the implicit social network between the users for the posts and comments posted within the two month period windows. We notice that there is an apparent increase in the average degrees for periods of four months from March 15, 2008 to July 14, 2008. Within the particular period there might be topics that have indulged users to increase in their usual pattern of activity. We analyzed the titles of the stories posted within that timeframe. Firstly, we removed the stop words, and then compiled a list of unique words sorted by frequency. The most active five words in the title included "Obama", "Clinton", "Mccain", "iPhone", and "Bush". The period was known for the current president, Barack Obama winning the Democratic candidate nomination and transitioning to campaign for the President's office.

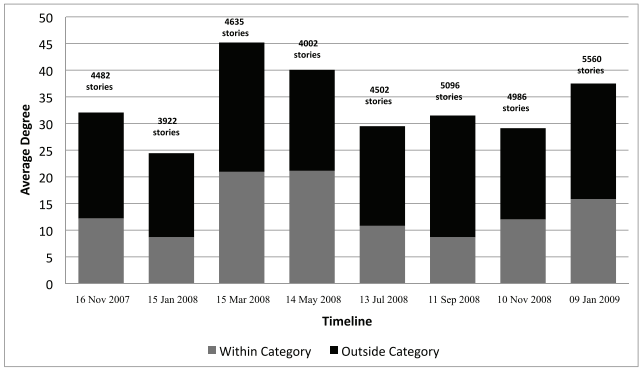


Figure 3: Average Degree for the 16 most active users in two month periods.

## 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 observed 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. Using the 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 interactions.

## References


Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 665–674. New York, NY, USA: ACM.

Ali-Hasan, N., and Adamic, L. A. 2007. Expressing social relationships on the blog through links and comments.

Fisher, D.; Smith, M.; and Welser, H. T. 2006. You are who you talk to: Detecting roles in usenet newsgroups. *Proceedings of the HICSS, Hawaii* 3:56–59.

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

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

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

Mishne, G., and Glance, N. 2006. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45(2):167–256.

Szabo, G., and Huberman, B. 2008. Predicting the popularity of online content. *Technical Report HP Labs* (0):1–6.

Welser, H. T.; Gleave, E.; Fisher, D.; and Smith, M. 2007. Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure* 8(2).

Whittaker, S.; Terveen, L.; Hill, W.; and Cherny, L. 1998. The dynamics of mass interaction. In *CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work*, 257–264. New York, NY, USA: ACM.

Zhongbao, K., and Changshui, Z. 2003. Reply networks on a bulletin board system. *Phys. Rev. E* 67(3):036117.