ISSN 2007-9737

Co-Comment Network: A Novel Approach for Construction of Social Networks within Reddit

Mrinal Kanti Baowaly[1], George William Kibirige[2], Bikash Chandra Singh[3]

[1] Bangabandhu Sheikh Mujibur Rahman Science and Technology University,
Department of Computer Science and Engineering,
Bangladesh

[2] Sokoine University of Agriculture,
Department of Informatics and Information Technology,
Tanzania

[3] Islamic University,
Department of Information and Communication Technology, Kushtia,
Bangladesh

baowaly@gmail.com, baowaly@bsmrstu.edu.bd,
georgek@sua.ac.tz, bikash.singh@ice.iu.ac.bd

Abstract. This work analyzes the online Reddit comments and proposes a novel approach to construct a social network automatically within Reddit that connects users to it. The investigation reveals that users tend to form links to others who have a mutual interest in interacting with some topics or posts. This mutual interest among users can make connections among them in the network. This study shows that Reddit contains a large, strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes which the authors define as the Co-Comment network. One of the important research focuses is to examine whether the networks built from the Reddit comments show the desirable social network properties, e.g., power-law distribution, small-world effect, and clustering coefficient. Results obtained from the experiment confirm that the Co-Comment network conforms to social network properties. The authors also discover communities within the Co-Comment network and verify them to validate the constructed network.

Keywords. Social network, communities in social network, network properties, power-law, small-world effect, clustering coefficient.

1 Introduction

A social network concerning human beings is interconnection of communities of people who share common interests, in the same social group either in work-related area, academic, religion, etc. When such social interaction is done on the internet it is called an online social network. Online social networks serve many purposes, but three primary roles stand out as common across most social groups [5]. First, online social networks are used to maintain and strengthen existing social ties or make new social connections. Second, online social networks are used by members to share information.

The content shared may vary from network to network. Third, online social networks are used to find information of interest in filtering, recommending, and organizing the content uploaded by users. Most social networks such as the one shown in Figure 1 are represented in the form of nodes that indicate users, and edges that indicate relationships between users [2]. Depending on the social media platform, one user may be able to contact any other user(s).

doi: 10.13053/CyS-26-1-4175

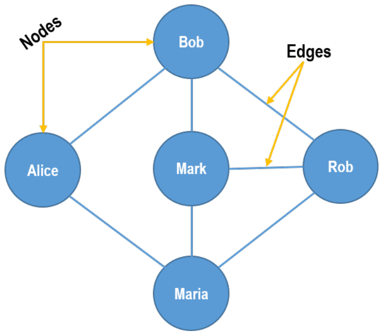


Fig. 1. Example of a social network

In other cases, users can contact anyone they have a connection to, and subsequently anyone the contacted user has a connection to, and so on. Some services require members to have a pre-existing connection to contact other members.

Some basic network properties for such real-world social networks include average shortest path (small-world effect), clustering coefficient, power-law distribution, etc. [14], which are used to analyze and evaluate social networks.

Within the social networks, diverse groupings may evolve based on user orientation or interests. Users with common interests may regroup to form smaller groups within the network called communities. The concept of a community is central to online as well as offline social networks [15]. A community is a subset of the users in a social network that is more tightly interconnected than the overall network. Communities are interesting for a variety of reasons. For example, users in a community tend to interact frequently, often share interests, and trust each other to some extent. In this research, our objective is to construct a social network and discover some potential communities within the network.

Reddit [23] is an entertainment, social networking, and news website where registered community members can submit content, such as text posts or direct links, making it essentially an online bulletin board system. Registered users can then vote submissions up or down to organize the posts and determine their position on the site's pages [30].

Unlike some most popular social media like Facebook and Twitter, Reddit does not have active social networking features, e.g. friendship, following, messaging between users. People that you mark as friends on Reddit are not notified that you have marked them as a friend [24]. You can only see your friends' posts in the r/friends subreddit (a community, and the posts associated with it, on the Reddit) although they can not see your post there. Moreover, you can not communicate with them as a friend. Hence, Reddit is used less for social networks or communities and more for talking with the world.

In this research, we aim to discover social networks among users within Reddit based on their common interest in commenting on other users' posts. The idea of making connections between two users is when some users discuss a topic or post with their comments and replies, it means they have a mutual interest in that topic or post. We hypothesize that this mutual interest in the discussion can lead to the formation of social networks, which we define as the Co-Comment network. Once we can form this kind of network, we can also find some communities in the network. The overall contributions of this research can be summarized in the following ways:

— Proposing a novel approach to construct social networks automatically within Reddit based on users' mutual interest in interacting with different posts.

— Analyzing the network properties such as power-law distribution, average shortest path (small-world effect), and clustering coefficient.

— Discovering communities in the constructed Co-Comment network.

The remaining sections of the paper are organized here. In section 2, related works are described. Section 3 discusses the proposed model to construct social networks and the methods to analyze these networks. Section 4 describes the dataset and does some pre-processing on the dataset. The experiment and results of the proposed method have been described and presented in section 5. Analysis and evaluation of the constructed social networks are also given in section 5. This paper includes the conclusion and some future work directions in section 6.

2 Related Works

As a result of its vast popularity, online social networks are the subject of an investigation by researchers all around the world today. In this section, we have described some past works relating to this paper of constructing social networks as follows.

Adamic, Buyukkokten, and Adar studied an early online social network at Stanford University [1] and found that the network has a small-world effect as well as a significant clustering. Using the rich profile data provided by the users they were able to infer the attributes contributing to the formation of friendships and to define how the similarity of users decays as the distance between them in the network increases.

Mining social network graphs appear as dominant structures in many fields, including sociology, biology, neuroscience, and computer science. In most of the mentioned cases, graphs are undirected – in the sense that there is no directionality on the edges, making the semantics of the edges symmetric as the source node transmits some property to the target and vice versa [6]. Liben-Nowell, Novak, et al. in their study [10] found a strong correlation between friendship and the geographic location of users by using data from Live-Journals.

Marc Barthélemy investigated the spatial characteristics of social networks in [3]. He analyzed that characterizing and understanding the structure and the evolution of spatial networks is crucial for many different fields, e.g, transportation and mobility networks, Internet, mobile phone networks, power grids, social and contact networks, and neural networks. Won-Yong Shin and et al. developed a novel framework [25] for analysing the degree of bidirectional online friendship via Twitter, while not only utilizing geo-tagged mentions but also introducing a definition of bidirectional friendship.

Martinez-Callaghan and Gil-Lacruz in [13] focused on how Japanese immigrants who settled in Spain and Scotland are building new social networks and explored how this process changes the way they relate to Japan and their Japanese identities. They found that both groups of immigrants emphasized five key elements in developing their new sense of belonging. These were: having a local partner; interest in the host country; workplace experiences; number of local and/or Japanese friends in their social network; and foreign language proficiency. An interesting feature that presents in real social networks is the clustering or community-structure property, under which the graph topology is organized into components commonly called communities or clusters. Numerous algorithms for automatically detecting communities in social networks have been proposed in [7, 18, 27, 12, 4, 8, 26].

As Lu, Wahlström, and Nehorai discussed in [11], community detection is important for many reasons. It allows the classification of the functions of nodes as per their structural positions in the communities. It reveals the hierarchical organization that exists in many real-world networks. It improves the performance and efficiency of processing, analyzing, and storing networked data. Some applications of community detection are topic/domain detection, friend suggestion, viral marketing, graph compression, parallel processing on graphs, etc. However, this study is different from all of the aforementioned past studies for constructing social networks. According to our knowledge, there has not been any study that proposed to build social networks based on users' mutual interest in interacting with different posts on Reddit.

3 Material and Methods

In this section, the authors present the project idea and discuss the design process. First, they give an overview of the proposed model and then the idea behind constructing the Co-Comment network.

3.1 The Proposed Model of Constructing Co-Comment Network

3.1.1 The Project Idea

This study used a dataset from Reddit.com.

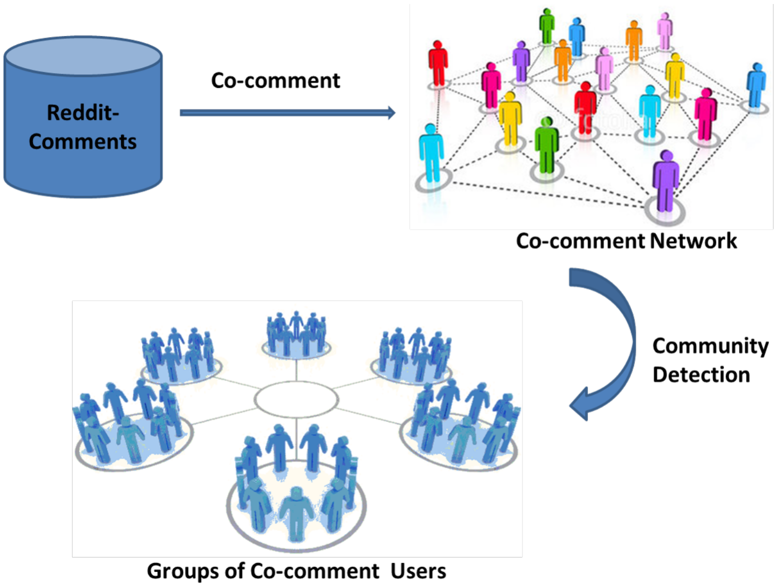


Fig. 2. Overview of the project idea

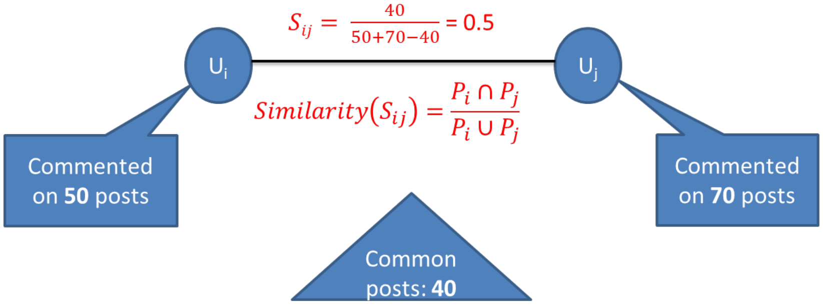

$S_{ij} = \frac{40}{50+70-40} = 0.5$
U_i
U_j
$Similarity(S_{ij}) = \frac{P_i \cap P_j}{P_i \cup P_j}$
Commented
on **50** posts
Commented
on **70** posts
Common
posts: **40**

Fig. 3. Establishing the connection between two users

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1j} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2j} & \cdots & W_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ W_{i1} & W_{i2} & \cdots & W_{ij} & \cdots & W_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nj} & \cdots & W_{nn} \end{bmatrix}$$

Fig. 4. Weight matrix

The content of this website is a collection of links e.g. news, text posts, articles, pictures, videos, etc. Users post links here and they also comment on other links. The authors first aim to build a social network (Co-Comment network) from the reddit-comments dataset [9] and then find communities in the Co-Comment network. The overview of the project idea is as shown in Figure 2.

3.1.2 Establish Connection between Users

It is already mentioned that the users can freely comment everywhere in the posts on Reddit. But if you observe the commenting behavior of the users, you will notice that they usually like to comment on some specific type of posts not everywhere. On the other hand, if two users both comment on some posts, it can say that they have a mutual interest and hence it can build a connection between them. Let us consider an example presented in Figure 3. Here, two users U_i and U_j commented on some posts on Reddit. If they commented on some common posts, it can make a connection between them based on their similarity (mutual interest). P_i denotes the number of posts the user U_i commented on and P_j denotes the number of posts the user U_j commented on. The similarity is computed according to Jaccard's coefficient [29].

In this example, the computed similarity between users is 0.5. That means they have a 50% common interest.

3.1.3 Building a Co-Comment Network

Calculating weight: Once we have built a connection or link between users based on their similarity, we then compute the weight of the link by just multiplying the similarity with their total posts. The equation is as follows:

$$\begin{aligned} Weight(W_{ij}) = Similarity(S_{ij}) \times \\ Total\ comments(P_i + P_j). \end{aligned} \tag{1}$$

For the previous example shown in Figure 3, we got weight:

$$W_{ij} = 0.5 \times (50 + 70) = 60.$$

Let us consider another example:

$$\begin{gathered} P_i = 12, P_j = 18, P_i \cap P_j = 10, S_{ij} = 0.5, \\ W_{ij} = 0.5 \times (12 + 18) = 15. \end{gathered}$$

In this case, you can observe that although in both cases the value of similarity is the same (0.5), their weights are different (60 and 15 respectively) because of their contribution (total number of comments) to the network.

Computing weight matrix: Following this method of computing weights for every user pair in the Reddit dataset we compute the weight matrix as shown in Figure 4. This matrix represents our Co-Comment network (graph).

Note that our network is undirected because we are making a connection between users is based on their common posts.

3.1.4 Steps of Constructing Co-Comment Networks

Overall, constructing the Co-Comment networks take the following steps in our study:

- First, compute the number of common posts for every user pair where they both commented on.
- Second, calculate similarity based on Jaccard's coefficient between users.
- Third, calculate link weight multiplying link similarity and total posts.
- Finally, compute the desirable weight matrix (i.e. Co-Comment network)

3.2 Analyzing Co-Comment Networks

After constructing the Co-Comment networks, it is needed to evaluate the networks for verifying whether the networks maintain the social network properties. In this investigation, we evaluate the Co-Comment networks using two different methods:

3.2.1 Network Properties

Commonly used measures such as average shortest path (small-world effect), clustering coefficient, power-law distribution are the attributes of networks that can be calculated to analyze network properties. The Co-Comment graph consists of undirected edges and nodes. The edges are the measured weights that connect nodes. The graph may consist of connected parts and disconnected or disjoints set of nodes. A social network focuses on the relationship among nodes and their attributes and not the individual node [14].

3.2.2 Community Detection

Community is also one of the fundamental aspects of social network analysis. The number of communities, if any, within the network is typically unknown and the communities are often of unequal size and/or density.

Despite these difficulties, however, several methods for community finding have been developed and employed with varying levels of success. In this study, the researchers applied the Louvain method [19] for detecting communities in the Co-Comment network. This method is a heuristic method that is based on modularity optimization. It outperforms all other known community detection methods in terms of computation time.

The quality of the communities detected is measured by the modularity, which is a scale value between -1 and 1 that measures the density of edges inside communities to edges outside communities [17].

Optimizing this value theoretically results in the best possible grouping of the nodes of a given network, however going through all possible iterations of the nodes into groups is impractical so heuristic algorithms are used.

In the Louvain Method of community detection, first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated.

4 Data Description and Pre-processing

We collected the Reddit dataset of their publicly available comments from the Kaggle website [9]. The comments were of May 2015 on Reddit. The basic statistics of the dataset are shown in Table 1.

Dataset such as Reddit comments obtained from Kaggle requires pre-processing such as cleaning before use. Some pre-processing techniques applied to our dataset are discussed as follows.

Table 1. Basic statistics of the dataset

Description	Count
No. of comments	54,504,410
No. of attributes/columns	22
No. of links/posts	3,839,004
No. of commenting authors/users	2,611,449

Table 2. Basic statistics of the co-comment networks

Data Sample	Constructed Network Type	# Nodes	# Edges
1	Undirected	9,710	677,694
2	Undirected	9,695	670,171
3	Undirected	9,712	677,105

4.1 Selecting Frequent Users

Since the total number of users in the dataset was 2,611,449 (over 2.6 million), it is very tough and time consuming to work on such a large number of users. Because of the time limit and resource constraint we decided to find only the frequent users who comment frequently. Our constraint was selecting users who commented on at least 10 posts. After completing this process, the users who have fewer commented posts were discarded. Therefore, the total number of frequent users $(Posts \geq 10)$ was: 650,169.

4.2 Data Cleaning

The social network is a crowd of users. As it can be seen in market places where we see so many people but with different motives. Some of them may be campaigners, promoters. Some evil persons, frauds, thieves may also be there.

This type of behavior can be seen in the online social network as well. Similarly, all the users are not real on Reddit. Some of them are robots, campaigners, and promoters. They usually comment on nearly every topic and a countless number of times. For this reason, we filtered out those frequent users who commented on more

Table 3. Average shortest path

Co-Comment Network	# Nodes	# Edges	l
1	9,710	677,694	2.6567
2	9,695	670,171	2.6329
3	9,712	677,105	2.6577

than 200 posts. Hence, the total number of cleaned frequent users (Posts: 10 - 200) is 636,537.

4.3 Feature Selection

There were 22 fields or features in the Reddit comment dataset. We checked each of them closely and found most of them were not required for our study. We selected only 7 fields that are relevant for constructing the Co-Comment network. These are Id, Name, link_id, author, subreddit_id, subreddit, and parent_id.

4.4 Data Sampling

Since our major task is to find out the common posts of every user pair, the current number of frequent users (i.e. 636,537) is still big. Hence, we picked 10,000 users randomly from the frequent set of users. We took three such random samples from the dataset.

5 Experiment, Result Analysis and Evaluation

The network obtained from the social domain exhibits great connectivity patterns that are often labeled with attributes and features to extract knowledge from such a network. This section addresses the construction of Co-comment networks, analysis and evaluation results of the networks after applying the proposed method and network analyzing techniques.

5.1 Co-Comment Networks

As we mentioned we took three random samples from the frequent set of users. Each sample consists of 10,000 users. We then applied our proposed method to each of these data samples and constructed three different Co-Comment networks. Note that each user in the data sample refers to a separate node in the network. The simulation results for all three networks obtained from the data samples are summarized in Table 2.

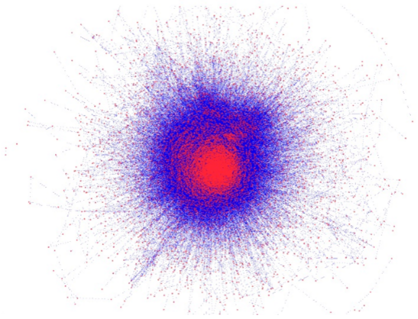


Fig. 5. Co-Comment network 1

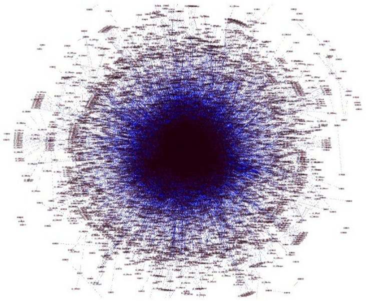


Fig. 6. Co-Comment network 2

The first data sample produces an undirected network of 9,710 nodes and 677,694 edges. The number of nodes in the network (i.e. 9,710) being less than 10,000 is because some isolated nodes have no connections in the constructed network. Figure 5 shows a graphical representation of the Co-Comment network obtained from data sample 1. The second data sample makes an undirected network of 9,695 nodes and 670,171 edges. Figure 6 shows the Co-Comment network obtained from data sample 2. The third data sample creates

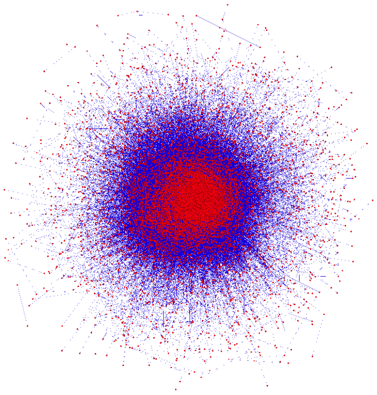


Fig. 7. Co-Comment network 3

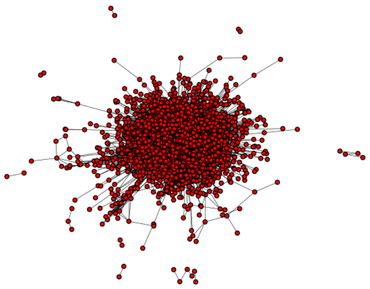


Fig. 8. Small-world effect in co-comment network

an undirected network of 9,712 nodes and 677,105 edges. Figure 7 shows the Co-Comment network built from data sample 3.

5.2 Analyzing the Networks

As we discussed the methods in section 3, we were required to analyze the network properties and detect communities in the Co-Comment networks to validate them as social networks. The processes are discussed as follows.

5.2.1 Verifying Network Properties

We analyzed three different network properties for the constructed Co-Comment networks.

Average shortest path: In the small-world effect, the shortest distance between one node to another is calculated and the average shortest distance of the whole graph can be found. The Co-Comment graph consists of a disjointed set. In this calculation, we consider the connected components only. Figure 8 shows a picture of the small-world effect on the first Co-Comment network. It shows that most nodes can be reached from every other node by a small number of hops or steps. The other two Co-Comment networks also show a similar result for the small-world effect.

Table 4. Average clustering coefficient

Co-Comment Network	# Nodes	# Edges	C
1	9,710	677,694	0.3667
2	9,695	670,171	0.3635
3	9,712	677,105	0.3650

However, the formula used to calculate the average shortest path is:

$$l = \frac{1}{0.5n(n+1))} \sum_{i \geqslant j} d_{ij}, \tag{2}$$

where l is the average shortest distance between all node pairs that have connected paths and d_{ij} is the shortest distance from node i to node j. The average shortest path (l) for the three constructed Co-Comment networks obtained from three data samples are presented in Table 3.

Clustering coefficient: The clustering coefficient is a parameter used to measure the network density of a network. The value of the clustering coefficient shows the probability that any two nodes e.g. i and j which are connected to another node k are also connected to each other making a triangle. The clustering coefficient C_i of a single node i is calculated using the Equation 3 and the average clustering coefficient C of the whole network is computed using the Equation 4 as follows:

$$C_i = \frac{Number\ of\ triangles\ connected\ to\ node\ i}{Number\ of\ triples\ centered\ on\ node\ i}, \tag{3}$$

$$C = \frac{1}{n} \sum_i C_i. \tag{4}$$

The average clustering coefficient C for the three Co-Comment networks are shown in Table 4.

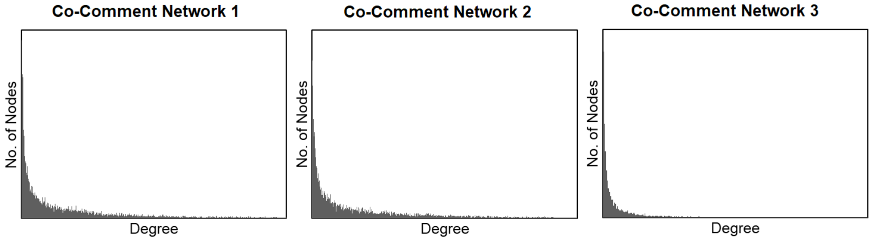
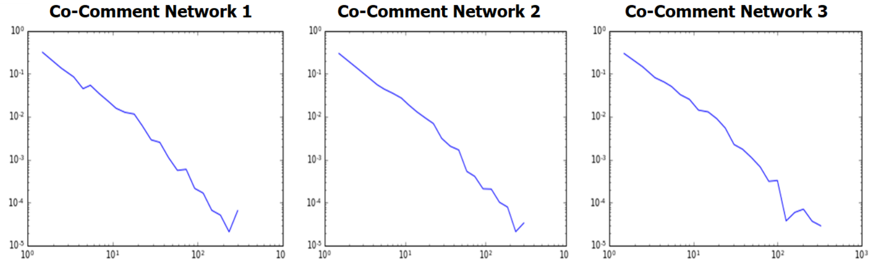


Fig. 9. Node degree distribution



α = 2.2593 α = 2.4026 α = 2.2440

α : power-law exponent (typically 2 ≤ α ≤ 3)

Fig. 10. Log-to-Log graph for power law distribution

Power-law distribution: The power-law distribution checks if a network graph conforms to the power-law distribution principle. We need to verify that Reddit Co-Comment networks satisfy the power-law distribution.

Figure 9 shows node degree distribution for verifying power-law distribution. We observe that the node degree distribution of each network is a long-tailed distribution that conforms to the power-law distribution.

We have also another way of verifying the power-law distribution using the value of the power-law exponent (α) in the log-to-log graph of each network as shown in Figure 10. This value is typically in the range $2 \leqslant \alpha \leqslant 3$ for the social networks.

With the growing popularity of online social networks, some studies have examined the properties of the networks over time. M. E. J. Newman in his study [16] showed different measurements of the network properties for a number of published social networks.

In this study, we display basic network properties for those published social networks vs our Co-Comment networks' properties in Table 5.

We notice from this table that the values of Co-Comment networks' basic properties, the average shortest path (l), average clustering coefficient (C), power-law exponent (α) are within the range of the values observed by the published social networks.

From this observation, we can assert that our Co-Comment networks constructed using the proposed model are capable of maintaining social networks' properties.

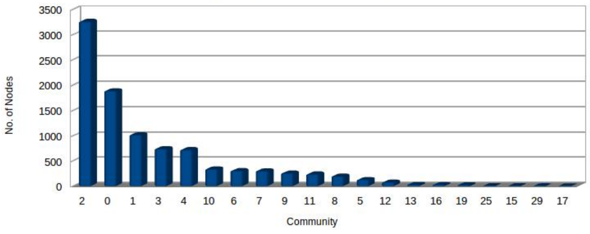


Fig. 11. Distribution of communities in Co-Comment network 1

Table 5. Basic network properties for some published social networks vs Co-Comment networks

	Network	Type	n	m	l	α	C
Some published social networks	film actors	undirected	449913	25516482	3.48	2.3	0.78
	company directors	undirected	7673	55392	4.60	-	0.88
	math coauthorship	undirected	253339	496489	7.57	-	0.34
	physics coauthorship	undirected	52909	245300	6.19	-	0.56
	biology coauthorship	undirected	152251	11803064	4.92	-	0.60
	telephone call graph	undirected	47000000	80000000	-	2.1	-
	email messages	directed	59912	86300	4.95	1.5/2.0	0.16
	email address books	directed	16881	57029	5.22	-	0.13
	student relationships	undirected	573	477	16.01	-	0.001
	sexual contacts	undirected	2810	-	-	3.2	-
Co-Comment networks	Co-Comment network 1	undirected	9710	677694	2.6567	2.2593	0.3667
	Co-Comment network 2	undirected	9695	670171	2.6329	2.4026	0.3635
	Co-Comment Network 3	undirected	9712	677105	2.6577	2.2440	0.3650

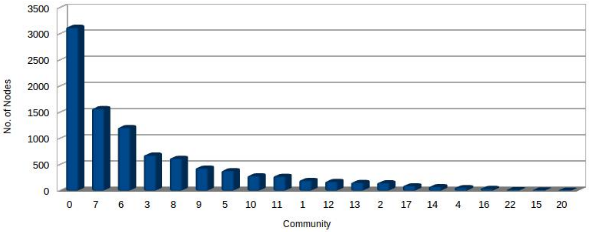


Fig. 12. Distribution of communities in Co-Comment network 2

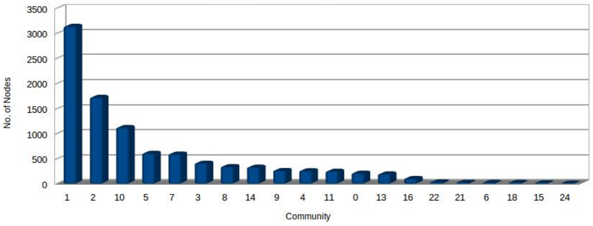


Fig. 13. Distribution of communities in Co-Comment network 3

5.2.2 Detecting Communities

As we discussed earlier, we used the Louvain method to discover communities within the Co-Comment networks. We listed the number of communities found within each network and the modularity in Table 6. Each modularity score for the discovered communities agrees with the expected value in the range between -1 and 1 that confirms the quality of the communities within the Co-Comment networks. The distribution of communities within each Co-Comment network is also presented here. Figure 11, Figure 12, and Figure 13 show the community distribution for the Co-Comment network 1, Co-Comment network 2, and Co-Comment network 3 respectively. The X-axis shows different communities, labeled by 0, 1, 2... and the Y-axis represents numbers of nodes/users.

5.2.3 Evaluating Communities

There are more than forty communities within each Co-Comment network as shown in Table 6. Due to the space constraint in this paper, a few communities from Co-Comment network 2 and Co-Comment network 3 were selected for evaluation. The community detection technique does not give actual communities always. Hence, we need to verify whether extracted communities satisfy their definition. There are different ways to do it. We followed the semantic evaluation technique. We will manually verify whether the extracted communities are coherent. For instance,

Table 6. Communities within Co-Comment networks

Co-Comment Network	# Nodes	# Edges	# Communities	Modularity
1	9,710	677,694	41	0.2546
2	9,695	670,171	46	0.2566
3	9,712	677,105	42	0.2568

we will first observe the distributions of subreddit categories in the community. Next, from this subreddit distribution, we may find out the nature of the community and then validate it.

Community 4 in Co-Comment network 2: This community belongs to Co-Comment network 2 and has 61 users.

Table 7 shows this community structure: subreddit categories, their ranking within the community 4, the number of comments posts by the users under each subreddit.

Table 7 shows that the electronic_cigarette subreddit has the highest frequency of comment posts. Therefore this community is considered as an electronic cigarette community. The users of this community commented on most of the posts relating to the electronic cigarette. The following subreddits: ecigclassifieds, EJuicePorn, DIY_eJuice, ecr_eu, Vaping are also related to e-cigarette [28]. Therefore, users of this community are most probably consumers of electronic cigarettes.

Community 10 in Co-Comment network 3: This community has 1,119 people and it belongs to Co-Comment Network 3.

Table 8 shows this community structure: subreddit categories, their ranking within the community 10, the number of comments posts by the users under each subreddit.

Table 8 shows that nba is the subreddit with the highest number of comment posts, 7,186. This is followed by nfl with 6,430 posts.

Table 8 shows only the top 12 ranking subreddits in this community.

The nba subreddit deals with basketball posts from nba league in the USA [19]. The second subreddit deals with nfl posts which is the national football league of the USA [20]. The third subreddit Soccer is involving football posts [21]. The SquaredCircle subreddit deals with another sporting event called wrestling [22]. The majority of the subreddits in this community are related to sports in general. Therefore this community can be considered as Sports community.

Table 7. Subreddit categories in community 4 of Co-Comment network 2

Ranking	Subreddit	# Comments
1	electronic_cigarette	481
2	supremeclothing	30
3	ecr_eu	10
4	DIY_eJuice	7
5	Vaping	4
6	Frat	3
7	JusticePorn	1
7	malefashionadvice	1
7	BTFC	1
7	ADHD	1
7	ecigclassifieds	1
7	vapeitforward	1
7	cars	1

6 Conclusion and Implications

In this research, the authors proposed a novel approach to construct social networks — designated as Co-Comment networks within Reddit based on the users' mutual interest in interacting with different topics or posts. This study first builds the Co-Comment networks from the Reddit comments

Table 8. Subreddit categories in community 10 of Co-Comment network 3

Ranking	Subreddit	# Comments
1	nba	7,186
2	nfl	6,430
3	Soccer	1,901
4	SquaredCircle	1,394
5	Boxing	1,190
6	Hockey	1,059
7	Asoiaf	493
8	Hiphopheads	457
9	gameofthrones	451
10	AskReddit	418
11	FIFA	328
12	MMA	235

dataset. It then analyzes and verifies these networks with the basic social network properties and community detection. The constructed social networks are proved realistic. Based on this extensive investigation it can be concluded that the authors can successfully create a social network within Reddit using the model proposed in this study. Furthermore, this research focused on Reddit as the case study to build social networks, the proposed approach is not restricted to Reddit only. The authors believe that the proposed model can be effectively used for other applications that pose similar mutual interests of the users to those explored in this study. Such kind of work can be done as future work.

References

1. **Adamic, L., Buyukkokten, O., Adar, E. (2003).** A social network caught in the web. First Monday, Vol. 8, No. 6. DOI: 10.5210/fm.v8i6.1057.

2. **Aggarwal, C. C. (2011).** An Introduction to Social Network Data Analytics, chapter 1. Springer US, Boston, MA, pp. 1–15. DOI: 10.1007/978-1-4419-8462-3_1.

3. **Barthelemy, M. (2011).** Spatial networks. Physics Reports, Vol. 499, No. 1, pp. 1–101. DOI: 10.1016/j.physrep.2010.11.002.

4. **Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008).** Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008, No. 10, pp. P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

5. **Boyd, D. M., Ellison, N. B. (2007).** Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, Vol. 13, No. 1, pp. 210–230. DOI: 10.1111/j.1083-6101.2007.00393.x.

6. **Broekstra, J., Kampman, A., van Harmelen, F. (2002).** Sesame: A generic architecture for storing and querying rdf and rdf schema. The Semantic Web – ISWC 2002, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 54–68. DOI: 10.1007/3-540-48005-6_7.

7. **Girvan, M., Newman, M. E. J. (2002).** Community structure in social and biological networks. Proceedings of the National Academy of Sciences, Vol. 99, No. 12, pp. 7821–7826. DOI: 10.1073/pnas.122653799.

8. **Gong, M., Ma, L., Zhang, Q., Jiao, L. (2012).** Community detection in networks by using multiobjective evolutionary algorithm with decomposition. Physica A: Statistical Mechanics and its Applications, Vol. 391, No. 15, pp. 4050–4060. DOI: https://doi.org/10.1016/j.physa.2012.03.021.

9. **Kaggle (2020).** May 2015 reddit comments. `https://www.kaggle.com/reddit/reddit-comments-may-2015`.

10. **Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A. (2005).** Geographic routing in social networks. Proceedings of the National Academy of Sciences, Vol. 102, No. 33, pp. 11623–11628. DOI: 10.1073/pnas.0503018102.

11. **Lu, Z., Wahlström, J., Nehorai, A. (2018).** Community detection in complex networks via clique conductance. Scientific Reports, Vol. 8, No. 1, pp. 5982. DOI: 10.1038/s41598-018-23932-z.

12. **Luo, F., Wang, J. Z., Promislow, E. (2006).** Exploring local community structures in large networks. 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), pp. 233–239. DOI: 10.1109/WI.2006.72.

13. **Martinez-Callaghan, J., Gil-Lacruz, M. (2017).** Developing identity, sense of belonging and social networks among japanese immigrants in scotland and spain. Asian and Pacific Migration Journal, Vol. 26, No. 2, pp. 241–261. DOI: 10.1177/0117196817706034.

14. **McGlohon, M., Akoglu, L., Faloutsos, C. (2011).** Statistical Properties of Social Networks, chapter 2. Springer US, pp. 17–42. DOI: 10.1007/978-1-4419-8462-3_2.

15. **Mislove, A. E. (2009).** Online social networks: Measurement, analysis, and applications to distributed information systems – PhD Thesis, Rice University. https://hdl.handle.net/1911/61861.

16. **Newman, M. E. J. (2003).** The structure and function of complex networks. SIAM Review, Vol. 45, No. 2, pp. 167–256. DOI: 10.1137/S003614450342480.

17. **Newman, M. E. J. (2006).** Modularity and community structure in networks. Proceedings of the National Academy of Sciences, Vol. 103, No. 23, pp. 8577–8582. DOI: 10.1073/pnas.0601602103.

18. **Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004).** Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, Vol. 101, No. 9, pp. 2658–2663. DOI: 10.1073/pnas.0400054101.

19. **Reddit (2020).** /r/nba - Subredit. https://www.reddit.com/r/nba/.

20. **Reddit (2020).** /r/nfl - Subredit. https://www.reddit.com//r/nfl.

21. **Reddit (2020).** /r/soccer - Subredit. https://www.reddit.com/r/soccer/.

22. **Reddit (2020).** /r/squaredcircle - Subredit. https://www.reddit.com/r/SquaredCircle/.

23. **Reddit Inc (2005).** Reddit. https://www.reddit.com/.

24. **Reddit wiki (2020).** Making friends. https://www.reddit.com/wiki/friends.

25. **Shin, W.-Y., Singh, B. C., Cho, J., Everett, A. M. (2015).** A new understanding of friendships in space: Complex networks meet twitter. Journal of Information Science, Vol. 41, No. 6, pp. 751–764. DOI: 10.1177/0165551515600136.

26. **Singh, B. C., Rahman, M. M., Miah, M. S., Baowaly, M. K. (2017).** Community detection using node attributes and structural patterns in online social networks. Computer and Information Science, Vol. 10, No. 4. DOI: 10.5539/cis.v10n4p50.

27. **Tyler, J. R., Wilkinson, D. M., Huberman, B. A. (2005).** E-mail as spectroscopy: Automated discovery of community structure within organizations. The Information Society, Vol. 21, No. 2, pp. 143–153. DOI: 10.1080/01972240590925348.

28. **Wang, L., Zhan, Y., Li, Q., Zeng, D., Leischow, S., Okamoto, J. (2015).** An examination of electronic cigarette content on social media: Analysis of e-cigarette flavor content on reddit. International Journal of Environmental Research and Public Health, Vol. 12, No. 11, pp. 14916—-14935. DOI: 10.3390/ijerph121114916.

29. **Wikipedia contributors (2020).** Jaccard index – Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=999247758.

30. **Wikipedia contributors (2020).** Reddit – Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Reddit&oldid=997746955.

Article received on 16/08/2021; accepted on 20/10/2021.
Corresponding author is Mrinal Kanti Baowaly.