Contextualized versus Structural Overlapping Communities in Social Media

Mohsen Shahriari, Sabrina Haefele, Ralf Klamma
Advanced Community Information Systems (ACIS)
RWTH Aachen University
Ahornstr. 55, 52056 Aachen, Germany
{shahriari, haefele, klamma}@dbis.rwth-aachen.de

ABSTRACT

Overlapping community structures are fundamental building blocks of Question Answer Forums (QAF) which are extremely linked with every social platform. Keeping in mind the value of structural-based Overlapping Community Detection (OCD) techniques, our imperceptible knowledge of content and contextualized information persuades us to do further research in this direction. To investigate community content, we devise two simple OCD algorithms taking use of posts in forums. In addition, we crawl data of an open source software project on modelling of chemical structures and make it available online for additional analysis. Furthermore, we analyse the devised methods and baseline algorithms on this dataset over the periods of its releases. Not only content could improve the performance of structural techniques but also two innovative algorithms performed competitive in comparison to other methods. Results also indicate reverse correlation between modularity and similarity of content in this forum, which show modularity and content as significant analytic indicators. Our content-based techniques and investigations can be applied by researchers in analytic and recommender systems.

Keywords

Overlapping community detection; Question answer forums; Structure and content-based community detection

1. INTRODUCTION

Nowadays, social networks are conflated with our daily life. Students check their Facebook profiles during the recess in classrooms or young graduates are looking for jobs searching in LinkedIn. In this regard one may notice, people liking the same page or being member of the same social group are perhaps more similar because of sharing same interests, innovations and content. To be more precise, they belong to the same community [14] [20]. People in a social network encounter some temporal challenges; they may leave the community or, they may stay and activate more in the context of another community. In other words, they leave the community and look for something new and suitable for their interests [16] [9]. Due to inherent contradiction of community detection criteria, research may fail to render a unique definition for **communities**; they appear in different resolutions from small to big ones in forums and peer production systems [2].

Community Detection (CD) algorithms detect clusters based on some predefined criteria and assign nodes to communities. Community structures are inherently overlapping; nodes participate actively in various domains and benefit from each community context. In addition to overlapping community structures, networks experience other properties such as shrinking diameter, small world-ness, motifs and dynamism [13] [15] [4]. Researchers in this domain have made a lot endeavour to devise algorithms and detect communities, however these algorithms may not be suitable for all existing contexts and they may be content-blind. So far, research has mainly studied structural methods that only take into consideration how nodes are interconnected among themselves. For instance, the map equation method may be suitable to system behaviour, network structures and local interactions. In contrast, stochastic and modularity techniques investigate network processes and their formations [17]. Furthermore, some of the algorithms work based on identification of influential members suitable for networks with hierarchical structures [18]. However, the structural-based algorithms do not consider whether the communities are meaningful or they may not reflect actual changes. In real life, people may initiate new connections or message each other while they have some thoughts, innovations and talks to share and communicate. Hence, conversations (here posts) and how much people tend to discuss together, may cause communities to form and hold people tightly connected [1].

To initiate the research, we assume that content, interesting debates and contextual similarities among people, contribute highly to bring people together and give rise to communities/clusters. In other words, structural properties of the graphs may be affected by contextual properties and the actual conversations of people in social media. Thus, consideration of this point may assist to detect more realistic clusters. This is actually what online media and real world networks are starving for; a content-based Overlapping Community Detection (OCD) algorithm. The research questions tackled in this work, can be listed as follows:

- We are inspired to figure out how structural properties like number of overlapping nodes, modularity and average community size are affected by contextual similarities among users in forums.
- Can adding of content improve the performance of structural based algorithms?

To proceed and investigate the above research questions, we employ information retrieval techniques such as Term Frequency and Inverse Document Frequency (TF-IDF) to devise OCD algorithms. In this regard, posts related to each user are extracted and converted to TF-IDF vectors. Afterwards, by defining an optimization problem together with K-means clustering algorithm, communities are detected. Overlapping members are identified by applying a threshold value based on node distances to the centroids; it is named Cost Function Optimization Clustering Algorithm (CFOCA). As for a baseline for content-based methods, a simple algorithm based on merging of communities is implemented. Considering each term as a cluster and merging of features based on an overlapping threshold, result in overlapping communities; this algorithm is called Term Community Merging Algorithm (TCMA). Moreover, the content similarity of users are added as an extra weight. In other words, weights induced by implicit number of communications among members, are combined with content weights. Afterwards, fully structural based techniques are applied on the new graphs such as SLPA [20], DMID [18], SSK [19] and CLIZZ [14]. Results indicate the better performance of structural algorithms when content is added up to the links. Moreover, the CFOCA and TCMA competitively detect communities. These algorithms are compared and contrasted regarding number of overlapping nodes, modularity and average community sizes. In addition for CFOCA, similarity costs versus some structural properties are plotted and analysed. To summarize, we make the following contributions:

- We crawled a new data from interactions and activities in an open source developer forum named Jmol. It is quite interesting for us to figure out dynamics of detection and analysis of communities in open source software development projects and thus we make this data publicly available for further analysis [1].

- We develop two simple content-based OCD algorithms based on simple information retrieval and optimization techniques. Results indicate competitiveness of these approaches. We further indicate that adding of content improves structural OCD algorithms.

- We compare content-based and structural based algorithms in terms of number of overlapping nodes, average community size and modularity and their correlation with content similarity of forum members. Results indicate reverse correlation between content similarity and modularity.

2. RELATED WORKS

One may ponder regarding OCD algorithms from different aspects; local versus global, dynamic versus static, node versus edge oriented techniques, structural versus context-based and behaviour of the algorithms are prevalent indicators to impart them to different categories. Structural methods refer to those approaches only considering the connections among members in a social network [17] [14]. In contrast, context-based methods apply node, edge attributes and actual context of the media to decide about communities [21] [6] [8].

Zhou et. al. apply a clustering method, more specifically a version of k-medoid clustering, to determine the distances between the nodes of an executed random-walk model on an attributed augmented graph. This graph is generated by adding dummy nodes for each of the attribute values and dummy links between nodes and associated attributes of the nodes [22]. Fisher proposes another method also using some kind of clustering which is named the Structure-Attribute Clustering (SAC). At the beginning each node represents one community, then a combination of the structure modularity and attribute modularity, called composite modularity gain, is computed and used to merge the communities [6].
Martin et. al. devised the connected k-center problem which uses intrinsic node characteristics to compute features. First, center nodes are determined, therefore the pairwise node distance is computed. Afterwards all non-center nodes are assigned to these centers, as a result, the nodes have to be within a certain radius of the center and the connectedness within clusters has to be respected [8]. Another possibility to determine the overlapping communities is using network decomposition as proposed in [7], called NCOCD. First the center clique is identified using a greedy polynomial algorithm and a local optimization strategy is used to expand this clique. Then the network decomposition is performed, so all links in the derived link communities in the current network are removed. This is repeated until no more new center cliques are found [7].
Liu et al. introduced the concept of content propagation to determine communities in networks, using content and structure. In this regard, community structure will be computed according to the interaction of the nodes. These interactions are modelled using two principles of content propagation. One based on influence propagation, which is approximated by a linear model, and the second principle employs a random walk to directly model the interactions. Basically both methods simply calculate the probability that the content of a node propagates to another node [11]. Yu et. al. propose two feature integration strategies, to regulate the effect of linkage structure and edge content in OCD process. In both methods TF-IDF is used to transform the content, then one approach combines the content vector with the vector representing the set of neighboring edges and corresponding weights. The other approach first applies Mahalanobis distance to calculate the distance between two nodes based on content [12]. Although some previous works have been proposed in this direction, they are not suitable to directly apply on QAFs or they may need huge tuning. Moreover, most of the methods do not deal with actual content rather with attributes of nodes and edges, therefore we approach these issues by introducing CFOCA and TCMA.

3. PROPOSED OCD ALGORITHMS

In this section we propose two content-based algorithms to detect overlapping communities. The second algorithm is very simple and we developed as for a simple baseline. The algorithms apply content of QAFs as a means to identify meaningful communities. Often communities form while people have some thoughts to discuss and to share some knowledge [1].

3.1 Constructing Term Matrix

To map the problem to vector and matrix spaces, we consider a vector named *vocabulary* for each user. This vocabulary vector is calculated based on all threads that user has posted and commented. Via applying TF-IDF, one may constitute the term matrix, therefore rows are corresponding to users and columns representing term frequencies. The

[1] `https://github.com/rwth-acis/REST-OCD-Services/wiki/Jmol-Dataset`

TF-IDF is computed as follows:

$$\begin{aligned} tf_t &= freq(t, v_u) \\ idf &= log(\frac{N}{|d \in D : t \in d|}) \qquad (1) \\ tf - idf(t, d_u, D) &= tf(t, v_u) \times idf(t, D). \end{aligned}$$

where v_u is the vocabulary related to user u, tf is the term frequency, idf is the inverse document frequency and N is the number of users. It worth noticing that one document (a vocabulary vector) is computed for each user u. $|d \in D : t \in d|$ is the number of all documents containing the term t. D is the whole set of documents/vocabularies of all the users.

3.2 Cost Function Optimization Clustering Algorithm (CFOCA)

In this algorithm, positions of nodes are employed to discover overlapping communities. Each node is representative by its corresponding row in the Term matrix T. The idea is that nodes with close positioning reside in the same communities and nodes with far positions locate in different ones. To consider positioning of nodes, K-means clustering is employed. Some nodes are selected as community representatives. The position of these selective nodes are updated until they relativistically resembles the community kernels. Often, to find the best centroids and the optimal distances, a cost function J is employed. To optimize the cost in CFOCA, the gradient descent method is applied. The centroids can be updated based on following:

$$c_j^{t+1} = c_j^t - \alpha \cdot \frac{\partial}{\partial c_j} J(c_1, ..., c_k). \qquad (2)$$

An optimization objective helps to find the best number of communities and the optimal distances among the centroids. As the cost function needs to be minimized, searching determine the k producing the lowest costs. Each of the data points is a vector and thus their distances can be calculated based on cosine similarity as follows:

$$\begin{aligned} J &= \frac{1}{n}\sum_{i=1}^{n} 1 - cosSim(u_i, c^{(u_i)}) \\ &= \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{u_i \cdot c^{(u_i)}}{\| u_i \|\| c^{(u_i)} \|} \qquad (3) \\ &= \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{\sum_{j=1}^{l} u_{i,j} \times c_j^{(u_i)}}{\sqrt{\sum_{j=1}^{l}(u_{i,j})^2} \times \sqrt{\sum_{j=1}^{l}(c_j^{(u_i)})^2}}. \end{aligned}$$

Here node u_i is assigned to $c^{(u_i)}$ and l is the number of words in T.

To compute the $\frac{\partial}{\partial c_j} J$, we proceed as follows:

$$\begin{aligned} &\frac{\partial}{\partial c}\frac{1}{n}\sum_{i=1}^{n} 1 - cosSim(u_i, c^{(u_i)}) \\ =&\frac{\partial}{\partial c}\frac{1}{n}\sum_{i=1}^{n} 1 - \frac{u_i \cdot c^{(u_i)}}{\| u_i \|\| c^{(u_i)} \|} \qquad (4) \\ =&\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial c} 1 - \frac{u_i \cdot c^{(u_i)}}{\| u_i \|\| c^{(u_i)} \|}. \end{aligned}$$

To compute the above derivation, we require the gradient of $g = u_i \cdot c^{(u_i)}$ and $h = \| c^{(u_i)} \|$ that can be computed as follows:

$$\begin{aligned} \nabla g &= u_i \\ \nabla h &= \frac{1}{2}\frac{2 \cdot c^{(u_i)}}{\sqrt{\sum_{j=1}^{l}(c_j^{(u_i)})^2}} \qquad (5) \\ &= \frac{c^{(u_i)}}{\| c^{(u_i)} \|}. \end{aligned}$$

By replacing the above values, gradient can be computed as follows:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial c} 1 - \frac{u_i \cdot c^{(u_i)}}{\| u_i \|\| c^{(u_i)} \|} \\ =&\frac{1}{n}\sum_{i=1}^{n} \frac{\nabla g \cdot \| u_i \|\| c^{(u_i)} \| - u_i \cdot c^{(u_i)} \cdot \nabla h \cdot \| u_i \|}{(\| u_i \|\| c^{(u_i)} \|)^2} \qquad (6) \\ =&\frac{1}{n}\sum_{i=1}^{n} \frac{u_i \cdot \| u_i \|\| c^{(u_i)} \| - u_i \cdot \frac{(c^{(u_i)})^2}{\| c^{(u_i)} \|} \cdot \| u_i \|}{(\| u_i \|\| c^{(u_i)} \|)^2}. \end{aligned}$$

The centroid c_j will be updated for step $t + 1$ as follows.

$$c_j^{t+1} = c_j^t - \alpha \cdot \frac{\partial}{\partial c_j} J(c_1, ..., c_k). \qquad (7)$$

Number of communities (k) is not known and it is needed to run the algorithm with different k values. The pseudo code of CFOCA is shown in Algorithm 1.

Algorithm 1 CFOCA based on k-means clustering

```
centroids ← centroid initialization
tempCent ← centroids
clustering ← membership matrix
i ← 0
j ← 0
while centroids ≠ tempCent do
    for all nodes do
        for all centroids do
            distCent1 ← dist(node.pos(i), centroid.pos(j))
            distCent2 ← dist(node.pos(i), centroid.pos(j+1))
            if distCent1 >= distCent2 then
                clustering.pos(i) ← centroid.pos(j+1)
            else clustering.pos(i) ← centroid.pos(j)
            j ← j + 1
        i ← i + 1
    tempCent ← centroids
    centroids ← updateCentroids(centroids)
return clustering
```

At this stage CFOCA yields disjoint communities that are not realistic. To assign nodes to overlapping communities a suitable threshold value ε is employed. Hence, a node is a member of different communities if its distance to centroids is less than ε.

3.3 Term Community Merging Algorithm (TCMA)

To consider the term vectors and content of forum with a much simpler approach, an algorithm called TCMA is proposed that works based on naive merging of overlapping communities and the idea is based on the approach proposed by Wang et. al. [3]. First a term matrix is constituted out of the existing terms of the vocabulary. Each column of this term matrix T indicates a term and the users will be incident to these word columns if they have used the term in her posts. In the continuation, the clusters are merged using the following overlapping coefficient:

$$overlapping\ coefficient = \frac{|C_i \cap C_j|}{min\{|C_i|, |C_j|\}}. \quad (8)$$

in which C_i and C_j are intermediate clusters identified in each step. We need to choose a suitable threshold β for the comparison of overlapping coefficient. We demonstrate the TCMA with a pseudo code shown in Algorithm 2.

Algorithm 2 Term Community Merging Algorithm

```
beta ← Overlapping Coefficient
features ← Term Matrix computed using tf-idf
clustering ← resulting clustering matrix
for each column in features do
    if features.position(currColumn,i) $\neq 0$ then
        $addToCluster(clustering.position(currColumn), i)$
refine(clustering)
    return clustering
```

3.4 Structural and Content-based Weighting

To better understand the effect of content on structural based OCD approaches, context of posts as similarities among users are also combined with the structural weight. $e_{u,v}$ is a tuple $\langle r, s \rangle$, where r is the structural weight and s shows the weight relating to contextual similarity of users u and v. The structural weight r can be computed from the number of links existing between the two users. Links are generated, for example when user u answers to a thread of user v. The content-based weight s is the cosine similarity between two users which its value ranges between 0 and 1. To detect the community structures, any structural-based OCD algorithms from the literature can be employed.

4. BASELINE METHODS

Here we compare CFOCA and TCMA with other structural based OCD algorithms. The baseline approaches are applied on structural-based information extracted from the QAFs. The structural information are extracted based on the communication traces of users in the same thread. Moreover, the baselines are also runned on the graphs with embedded content-based weights.

4.1 SLPA

SLPA is an extension of label propagation algorithm that each node holds several labels in its memory. Each node can play the role of either listener or speaker that receives or signals the information. Firstly, memory of nodes are initialized with multiple labels. Afterwards, nodes propagate the labels based on some speaking and listening rules. Finally, one can extract the communities based on the labels stored in the memory of nodes [20].

4.2 DMID

It is a two-phase leader-base algorithm which works based on two simple social dynamics named disassortative degree mixing and information diffusion. In the first phase, disassortative hubs are identified by performing a random walk. In other words, dissimilarity among nodes together with their degrees, are combined to identify the hierarchy of the network and the influence of members. In the second phase, a network coordination game is applied to compute degree membership of nodes to influential members. Second phase employs diffusion of information and innovations [18].

4.3 CLIZZ

The CLIZZ algorithm starts by constituting an influence range for each node. Influence ranges are made based on the distances as follows:

$$LS_i = \sum_{j=1\ ;\ d_{ij} \leq = \lfloor \frac{3\delta}{\sqrt{2}} \rfloor}^{n} e^{-\frac{d_{ij}}{\delta}}. \quad (9)$$

Leaders are identified as nodes with high linkage values and as big proxies to other nodes. To compute the membership of nodes to leaders, a membership vector is initilized with a uniform distribution and then updated as follows:

$$M_i(t+1) = \frac{1}{1 + \sum_{j=1}^{n} A_{ij}} \left[M_i(t) + \sum_{j=1}^{n} A_{ij} M_j(t) \right], \quad (10)$$

where M is the membership, A is the adjacency matrix and d indicates the shortest path between two nodes.

4.4 SSK

SSK starts with identifying most influential nodes. It is accomplished by constructing a transitive weight matrix as follows:

$$w'_{ji} = w_{ji} + \sum_{k} c^{k}_{ji}, \quad (11)$$

that $c^{k}_{ji} = min\{A_{ki}, A_{jk}\}$ is the transitive link weight. This matrix is normalized to be applied for the random walk process in order to obtain the steady state influence vectors. By comparing the influence of a node with its neighbours the leaders are identified. Membership of nodes to leaders are computed with another random walk on the transition matrix induced from the adjacency matrix [19].

5. DATASETS AND METRICS

The investigated data is based on the forum discussions about a Java project called Jmol. The Jmol[2] project resulted in an open-source Java-Tool for molecular modeling of chemical structures in 3D. The reason behind crawling this dataset is to explore and investigate the open source developer communities. We crawled the data over a period of eleven years (2002 - 2012) and made it publicly available for research purposes. We investigated this data set based on years, months of the year 2004 and the project releases. Some information about this dataset is shown in Table 1. In this paper, analysis over the release periods are demonstrated. Figure 1 indicates the degree distribution which is plotted for several release periods of the data. As it can be observed the revived communications in the threads more or less follow a power-law degree distribution which confirms the data is felicitous to be further evaluated.

For the evaluation we used a metric that takes into account the content of the forums and the structure of the constructed graph equally; in the following we define the

[2] www.jmol.org

Periodnr	# nodes	# edges	# terms
1	21	0	1705
2	19	0	2523
3	25	0	3316
4	21	0	2246
5	30	0	3766
6	352	615	32081
7	170	529	15832
8	109	285	8220
9	80	250	7841
10	127	514	12483
11	164	573	13528
12	160	567	16900
13	183	651	17818
14	111	364	8962

Table 1: Each release period, number of nodes, number of implicit created edges and number of edges in Jmol

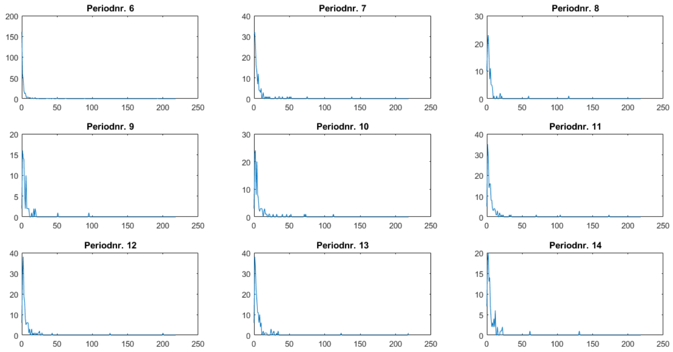

Figure 1: Degree distribution of the Jmol data for several releases; As it can be observed the data follows a power law degree distribution.

Combined Modularity [5]. It combines the Newman modularity with a similarity measure, in this case we use the cosine similarity defined in Equation 3. So we can compute the values of the Combined Modularity as follows:

$$Q_{comb} = \alpha * \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{deg(i)deg(j)}{2m}] \delta(i,j) \\ + (1-\alpha) * \sum_{ij} cosSim(v_i, v_j) \delta(i,j) \quad (12)$$

,

Where m is the number of edges, A is the adjacency matrix and $\delta(i,j)$ is equal to 1 if node i and j have one community in common, otherwise 0. Furthermore v_i corresponds to the row vector in T representing node i and α is a threshold to determine the importance of the structural and the content measure. If $\alpha = 1$, this measure will behave as the normal Newman modularity. It causes equal importance for both content and structure if the parameter is set to 0.5.

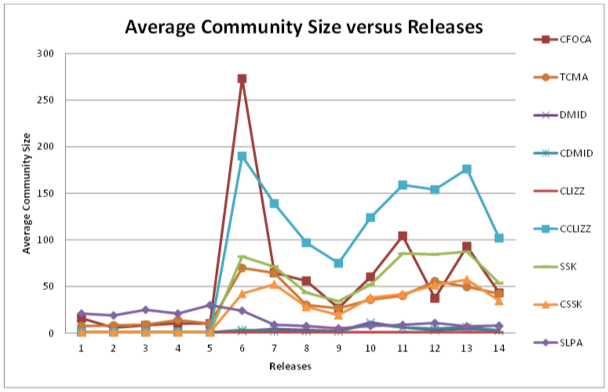

Figure 2: Average community sizes by different applied algorithms on Jmol dataset. The horizontal axis indicates the releases in Jmol forum.

6. RESULTS

We apply different algorithms on Jmol data and compare them from different aspects. The algorithms include CFOCA, DMID, Content-based DMID (CDMID), SLPA, ClIZZ, Content-based ClIZZ (CCLIZZ), SSK, Content-based SSK (CSSK) and TCMA. While SLPA is suitable for unweighted networks, we could not apply it on weighted graphs. To analyse the algorithms, several different diagrams are plotted; number of overlapping nodes, average community sizes, modularity, correlation of modularity, average community size versus similarity costs are plotted. Jmol is a temporal data that its corresponding timely networks together with its context encounter transformations over time. In addition to the monthly and yearly analysis of the dataset, we applied a release-based investigation on it that we demonstrate the release-based results in this paper.

Release-based analysis reveals more meaningful analysis of detected covers while major alterations in the communities happen due to releases of software. Usually close to releases, people may communicate with each other more than other times. Or right after a release the contextual and structural communications among members face massive changes [10]. In Figure 2, one can observe the average community sizes by applied OCD algorithms versus the release times. Almost structural based algorithms were unable to detect overlapping communities while there were no explicit thread communications among members during the first 5 releases. This can be justified while the developers were working structurally more isolated than cooperative thus leading to separated content generations. Usually at the beginning of the project, people try to understand the elements of the project and get familiarize with it and thus causing a cold start situation. However, their generated contents could reveal some communities with the applied CFOCA and TCMA. This indicates that when explicit structural information of the network is missing, content-based algorithms dealing with actual context of social media may be more informative. In other words, when cold start problem exists, content-based methods are still able to detect overlapping communities.

Furthermore, DMID and CDMID detect low average community sizes. One may spot less fluctuations among detected communities while applying structural methods such

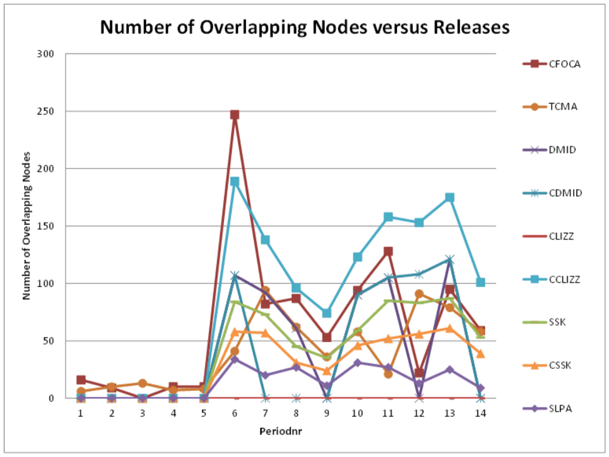

Figure 3: Number of overlapping nodes versus the release dates in Jmol dataset.

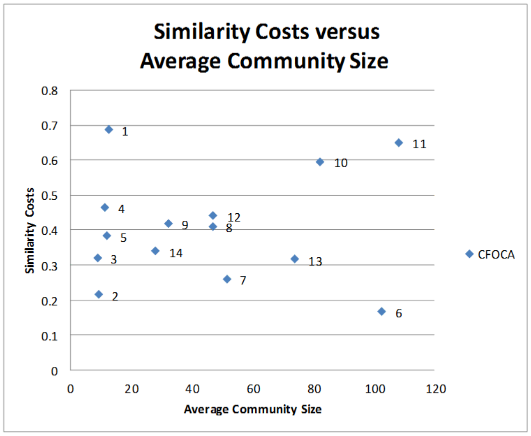

Figure 4: Similarity costs versus average community size for CFOCA algorithm.

as SLPA, DMID, SSK and CLIZZ. This can be justified as structural methods only track the communication threads, however forums are more context dependent and conversations' topic and content change over the period of releases. Content-based algorithms such as CFOCA and TCMA further reflect this issue. Other interesting findings can be observed, for instance, CLIZZ is unable to detect community structures when there is no content but content-based weighting version, can detect communities. Additionally, in release 6, highest average community sizes belong to CFOCA (273), CCLIZZ (190), SSK (82.1), TCMA (69.8) and CSSK (42.33). Other algorithms like DMID (2.28), CDMID (3.53), SLPA (24) detect smaller average community sizes. If we also consider release 9, we notice CLIZZ (75), SSK (34.26), CFOCA (26.8), TCMA (26.6), CSSK (19.44), DMID (2.22), CDMID (2.22) which indicates the clear difference in community resolutions for different algorithms.

Figure 3 indicates number of overlapping nodes generated by different algorithms. Similar to Figure 2, structural methods were unable to detect overlapping nodes when there are lack of edges for the first five releases. As for DMID, adding of content negligibly affects on the number of overlapping nodes. For instance, in periods 10 and 11 both DMID and CDMID detected respectively 105 overlapping members. Regarding CFOCA, number of overlapping nodes are also higher than other algorithms. For instance, CFOCA detected 247, 82, 87 and 53 overlapping nodes for periods 6-9, however TCMA as a content-based method discovered 41, 94, 62 and 36.

Although DMID detects low average community sizes, the number of overlapping nodes are as high as CFOCA and TCMA. Other interesting issue which can be perceived is that, adding of content increase number of overlapping nodes in almost all of the structural based approaches except SSK. This can be justified while content and context reveal broader boundary overlaps among the covers. As for SLPA, number of overlapping nodes are much lower than other algorithms. Additionally, number of overlapping nodes are at levels of average community sizes. This may be because of availability of content that cause identification of most of the nodes as overlapping; members participate in different contexts and disciplines of the project.

To investigate the relationship between the contents generated in forum communications and the community structure, we plotted the similarity costs versus the average community size, modularity and number of overlapping nodes. Higher similarity costs indicate less similarities among people. On the contrary, lower similarity costs are signs of higher similarity. As Figure 4 indicates the content similarity versus average community size, it can be observed that periods 1, 10 and 11 have the highest similarity costs (lowest content similitude). Release 1 may indicate small communities of people at the beginning of the project. In contrast, releases 10 and 11 show bigger communities and still lower similarities which might indicate lower collaboration at the end of the project. Although observing small communities at releases 2 , 3 , 4 and 5, content similarities enhance for them. The highest content similarity can be observed for release 6 that increment of members from 30 to 352 members, can be a reason for this. Furthermore in Jmol dataset, one may observe that average community sizes between 10 to 50 have the highest similarity of content. This has an exception with release 6 with more 100 than average community size property.

Regarding similarities of content and overlapping nodes, one may take a look at Figure 5. As it can be perceived, highest content similarities can be observed for releases 2, 3, 4 and 5, and this may be due to few number of members in the beginning of the project. When there are few number of members (around 20 members), overlapping nodes are limited to 2 to 3 members which seems realistic. There is an increase in number of nodes at release 6 and correspondingly in the number of overlapping nodes, that may indicate communities having high mixing tendency. As for releases 7, 9, 14 and 10 there are somehow high content similarity and around 36 to 68 overlapping members. At these release periods, number of members are around 100 which indicates half of the nodes are overlapping among communities. While there is high content similarity among members, they have some ideas to share and boundary spanners can further propagate information over the network.

Figure 6 indicates a reverse relation between the similarities and the modularity values of the detected communities by CFOCA algorithm. For most of the releases, for instance

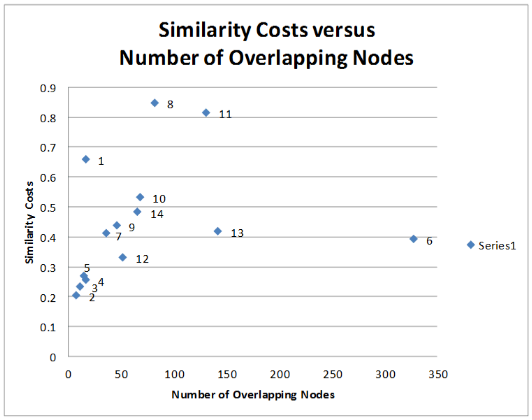

Figure 5: This figure indicates similarity costs versus number of overlapping nodes for CFOCA algorithm.

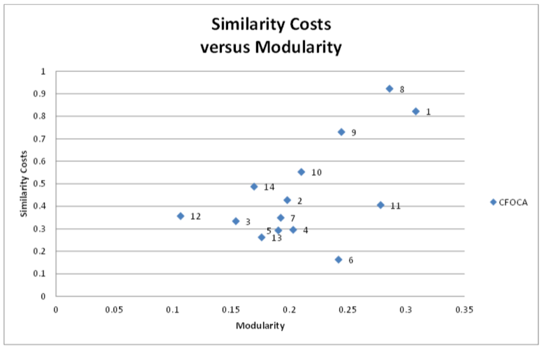

Figure 6: This figure indicates similarity costs versus modularity for the CFOCA algorithm.

3, 5, 13, 7, 2 and so on, the similarity costs are low which indicate high similarity of content among members that lead to low modularity in detected communities. Other releases including 1, 8 and 9 have high similarity costs and high modularity structures that indicate lower similarity contents; the detected communities tend to be more modular.

Regarding the modularity of the detected communities by different applied algorithms, Figure 7 is suitable to catch information. As it is observable, for the first 5 releases structural methods generate 0 modularity due to lack of communication threads in the network. In five snapshots, SLPA generates approximately high values for modularity but it does not have the proper resolution while each node is assigned to a single community which is not realistic. TCMA obtains higher modularity values in comparison to CFOCA. However, as results for SLPA indicates, modularity is not the only prevailing factor. Because considering nodes as single communities even generate high modularity values but they are not meaningful communities. If one consider the release 6, modularity values of the algorithms are respectively, CCLIZZ (0.40), TCMA (0.31), SLPA (0.27), CFOCA (0.243), CSSK (0.11), DMID (0.05), CDMID (0.05), CLIZZ (0). This indicates that content-based algorithms such as CCLIZZ, TCMA and CFOCA also reach satisfactory values for modularity. However, we need to recognize that modularity is not the only prevailing factor and other factors such as meaningfulness of the communications should be taken into account.

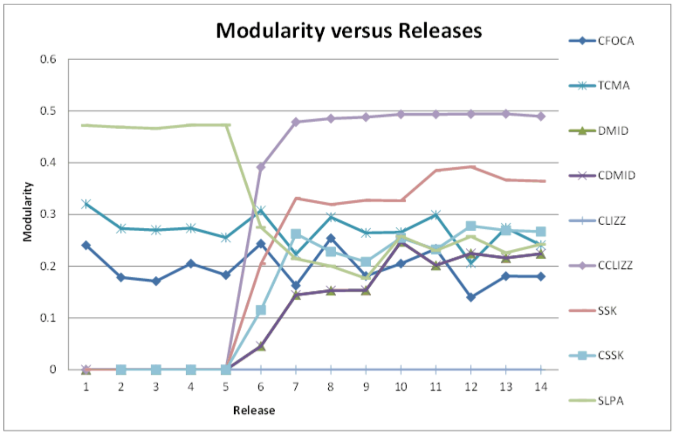

Figure 7: This figure indicates modularity of detected communities for different releases for Jmol dataset.

7. DISCUSSION AND FUTURE WORK

In this paper, we innovated two simple OCD algorithms employing content to identify overlapping community structures. We also enriched the structural methods with content information as new weights and applied structural techniques. Algorithms are evaluated on an open source developer communities named Jmol. A temporal version of this dataset is considered for the analysis of results. Number of overlapping nodes, average community sizes and modularity are investigated to extract informative corollaries. Furthermore, number of overlapping nodes, community sizes and modularities are sketched versus modularity. Our experiments suffer from several shortcomings that we need to address them in our future works. Our cost function generates a global similarity cost value that may not reflect the actual similarities among members. To gain more realistic and fine grain dynamics, we need to investigate and define some local similarity values.

In some cases, assigning each node to single communities yielded the highest values for modularities. Although it does not dissatisfy the definition of overlapping communities, it is extremely unfair to get the highest values of modularity. Even applying of a combined version of modularity using of both content and structure, did not resolve the issue and increases our curiosity to investigate a better way to evaluate the goodness of the algorithms. Furthermore for some algorithms, adding of content were beneficial for the structural based techniques and resulted in better performance of the algorithm. In this regard, we would like to investigate what structural-based OCD techniques may be improved via content adage.

8. CONCLUSION

Veritable increment in using peer production systems such as forums, has created massive amount of connections and contextual data which put a challenge on researchers for their analysis. One instance of the platforms is the open source developer networks that members of a project collaborate together to develop a product. In this article, we crawled one of those networks named Jmol. We indicated that Jmol has power law degree distribution and demonstrated basic properties of this network. Moreover, we developed two simple OCD algorithms suitable for QAFs where there does not exist explicit connections among members. Term frequency of posts together with a k-means algorithm equipped with a suitable cost function optimization are employed to identify communities.

Furthermore in the second approach, communities are detected by merging and finding the overlaps among the term frequency vectors. As a third strategy, content of the communications are added to the implicit structural weight extracted from forums and several structural methods of community detection are investigated on them. Results indicate the positive effect of content on structural OCD methods. Moreover, content-based algorithm versions are able to identify meaningful and competitive communities. Result section of the paper manifests some informative correlation findings in our study.

9. ACKNOWLEDGEMENT

The work has received funding from the European Commission's FP7 IP Learning Layers under grant agreement no 318209.

10. REFERENCES

[1] S. C. Baek, S. Kang, H. Noh, and S. W. Kim. Contents-based analysis of community formation and evolution in blogspace. In *2009 IEEE 25th International Conference on Data Engineering*, pages 1607–1610, 2009.

[2] M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 866–874, New York, NY, USA, 2008. ACM.

[3] B. Cai, H. Wang, and H. Zheng. An improved random walk based clustering algorithm for community detection in complex networks. In *2011 IEEE International Conference on Systems, Man and Cybernetics - SMC*, pages 2162–2167, 2011.

[4] R. Cazabet, F. Amblard, and C. Hanachi. Detection of overlapping communities in dynamical social networks. In *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pages 309–314.

[5] T. A. Dang and E. Viennet. Community detection in social networks with attribute and relationship. In *Extraction et gestion des connaissances (EGC'2012)*, pages 563–564, 2012.

[6] David Fisher. Principles of trust for embedded systems, 2012.

[7] Z. Ding, X. Zhang, D. Sun, and B. Luo. Overlapping community detection based on network decomposition. *Sci Rep*, 6(24115), 2016.

[8] R. Ge, M. Ester, B. J. Gao, Z. Hu, B. Bhattacharya, and B. Ben-Moshe. Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications. *ACM Transactions on Knowledge Discovery from Data*, 2(2):7:1–7:35, 2008.

[9] D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in dynamic social networks. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183, 2010.

[10] A. Hannemann and R. Klamma. Community dynamics in open source software projects: Aging and social reshaping. In *Open Source Software: Quality Verification*, volume 404 of *IFIP Advances in Information and Communication Technology*, pages 80–96. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[11] L. Liu, L. Xu, Z. Wangy, and E. Chen. Community detection based on structure and content: A content propagation perspective. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 271–280, 2015.

[12] L. Yu, B. Wu, S. Zhao, and B. Wang. Overlapping community detection in large networks from a data fusion view. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 118–121, 2014.

[13] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[14] H. J. Li, J. Zhang, Z. P. Liu, L. Chen, and X. S. Zhang. Identifying overlapping communities in social networks using multi-scale local information expansion. *The European Physical Journal B*, 85(6), 2012.

[15] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[16] G. Palla, P. Pollner, A.-L. Barabási, and T. Vicsek. Social group dynamics in networks. In *Adaptive networks: NECSI. Thilo Gross ; Hiroki Sayama ed*, New England Complex Systems Institute book series, pages 11–38. Springer, Berlin and Heidelberg, 2009.

[17] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *Phys. J. Spec. Top*, 178:13–23, 2009.

[18] M. Shahriari, S. Krott, and R. Klamma. Disassortative degree mixing and information diffusion for overlapping community detection in social networks (dmid). In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1369–1374, 2015.

[19] A. Stanoev, D. Smilkov, and L. Kocarev. Identifying communities by influence dynamics in social networks. *Physical Review*, 84(4), 2011.

[20] J. Xie, B. K. Szymanski, and X. Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *CoRR*, abs/1109.5720, 2011.

[21] J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*, pages 1151–1156.

[22] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729, 2009.