

# Identifying Cliques in Dark Web Forums – An Agglomerative Clustering Approach

Tarique Anwar

Center of Excellence in Information Assurance  
King Saud University, Riyadh, KSA  
Email: tAnwar.c@ksu.edu.sa

Muhammad Abulaish, *SMIEEE*

Center of Excellence in Information Assurance  
King Saud University, Riyadh, KSA  
Email: mAbulaish@ksu.edu.sa

**Abstract**—In this paper, we present a novel agglomerative clustering method to identify cliques in dark Web forums. Considering each post as an individual entity accompanying all the information about its thread, author, time-stamp, etc., we have defined a similarity function to identify similarity between each pair of posts as a blend of their contextual and temporal coherence. The similarity function is employed in the proposed clustering algorithm to group threads into different clusters that are finally presented as individual cliques. The identified cliques are characterized using the homogeneity of posts therein, which also establishes the homogeneity of their authors and threads as well.

**Index Terms**—Cyber security, Clique discovery, Dark Web forums, Agglomerative clustering.

## I. INTRODUCTION

The present day Web forums are a result of technological evolution of traditional bulletin boards (or notice boards) intended to manage *user-generated contents* (UGC), inspired by the rapid growth of Web applications. Web forums provide a platform for formal, vivid and dynamic discussions among an unrestricted number of participants. In this folksonomy, discussions are started by its members in the form of a discussion thread with a title and an entry message post, and viewers of this thread annotate their own opinions or replies. Thus the system keeps on evolving as the number of posts grow in the thread. During the course of discussions, the interactions in the form of replies and responses stir to establish a social relationship of trust and faith among the unknown users, and this nature of Web forums promotes them to be a part of online social media. The group of elusive relationships established thus among the participants usually evolve to develop an online society where people sharing similar interests gradually move closer to each other. Not varying from the real world societies, the inherent evils have found its place even in the online societies in the form of racism, extremism, offensive or disruptive online behavior, and Cyber-bullying.

Research on analyzing Web forums for tracking and dealing with the grievous threats posed by extremist and hate groups being active in them, has gained considerable attention of the research community. The portion of the World Wide Web (WWW), circumscribing the sinister objectives of terrorist and extremist groups is said as the “Dark Web”, and specifically

the Web forums with substantial prevalence of activities supporting terrorism or extremism are said as “Dark Web Forums” [1]. It is usually seen that the dark Web forums prevalent in our society are not marked with hate and violence in each thread and post of discussions. Rather, there remains only a small and subtle portion inside the dark circle. Due to its covert and obscured nature, carving out this dark clique from the complete Web forum is a challenging task. In this paper, we present an agglomerative clustering method to deduce different cliques of a dark Web forum. For this, we consider each post as an individual independent entity, accompanying all the information about its thread, author, and time-stamp. The intuition behind considering posts as individual entities instead of threads is that, a thread is participated by multiple users, and each of them may not be promoting the hate and violence in their posts. Similarity between each pair of existing posts is identified following a novel similarity measure, designed in the context of Web forums.

## II. RELATED WORK

Since last few years, research on dark web forums has received substantial attention from the research community working in the domain of *intelligence and security informatics*. In a recent work [2], Qin *et al.* performed an empirical study of different global extremist organizations on the Web and presented how sophisticatedly they propagate their ideologies. Several studies have focused on sentiment analysis, opinion mining and affect analysis of user posts in Web forums [3], and the discovery of user roles and their ties have been appraised [4].

In [5], the authors analyzed Web opinions to cluster them into groups of major themes of discussions. They found it as a challenging task because of its unstructured nature, unrestricted growth, noise, and evolving topics. In this work, a distance-based algorithm is applied to avoid the requirement of predefined number of clusters, that ensures a required density for initial clusters, and uses scalable distances to expand them. Recently, Yang *et al.* [6] came up with a spectral coherence based clustering approach to identify dark Web clusters, which considers the temporal coherence of user activeness rather than contents or links as the primary information. They represented a group of users as a  $m$ -dimensional multivariate process

which is used to derive the spectral density matrix and finally spectral coherence score is computed to identify the clusters.

### III. PROPOSED METHOD

Prior research works show that a similarity comparison of Web forum posts is not as trivial as usual content similarity [7]. Liu *et al.* [7] defined this measure as a function of body text, thread title and author of the post. However, on analysis, we found that time plays a substantial role in deciding the topics of discussion and its deviation with respect to the daily happenings in one's personal life. For example, immediately after the tsunami outbreak in Japan in March 2011, all social media got flooded with this hot discussion all over the world. To find overall similarity between a pair of posts, we calculate four different similarity measures – *content similarity*, *time similarity*, *author similarity* and *title similarity*. Let  $D = \{d_1, d_2, \dots, d_n\}$  be the set of discussion threads and  $P^i = \{p_1^i, p_2^i, \dots, p_m^i\}$  be the set of ordered posts in thread  $d_i$  in a forum  $F$ . After being cleaned and chunked as a pre-processing step, each post  $p_j^i$  is converted into bag of unigrams, bigrams and trigrams. Thereafter, vector space model (VSM) is used to transform each post into vectors of unigrams,  $\overrightarrow{Un}_{i,j}$ , bigrams,  $\overrightarrow{Bi}_{i,j}$ , and trigrams,  $\overrightarrow{Tr}_{i,j}$ , using their *tf-idf* values. The content similarity  $CSim(p_j^i, p_l^k)$  between a pair of posts,  $p_j^i$  and  $p_l^k$  is calculated using equation 1, where  $\alpha_1 \leq \alpha_2 \leq \alpha_3$  are weight controlling parameters such that  $\alpha_1 + \alpha_2 + \alpha_3 = 1$ .

$$CSim(p_j^i, p_l^k) = \alpha_1 \times \frac{\overrightarrow{Un}_{i,j} \cdot \overrightarrow{Un}_{i,l}^k}{\|\overrightarrow{Un}_{i,j}\| \|\overrightarrow{Un}_{i,l}^k\|} + \alpha_2 \times \frac{\overrightarrow{Bi}_{i,j} \cdot \overrightarrow{Bi}_{i,l}^k}{\|\overrightarrow{Bi}_{i,j}\| \|\overrightarrow{Bi}_{i,l}^k\|} + \alpha_3 \times \frac{\overrightarrow{Tr}_{i,j} \cdot \overrightarrow{Tr}_{i,l}^k}{\|\overrightarrow{Tr}_{i,j}\| \|\overrightarrow{Tr}_{i,l}^k\|} \quad (1)$$

Time similarity,  $TSim(p_j^i, p_l^k)$ , is calculated using equation 2, where  $ts(p_j^i)$  and  $ts(p_l^k)$  are times-tamps of posts  $p_j^i$  and  $p_l^k$ , respectively and  $\beta_1 \in [0, 1]$  is a constant.

$$TSim(p_j^i, p_l^k) = \beta_1^{|ts(p_j^i) - ts(p_l^k)|} \quad (2)$$

Author similarity,  $ASim(p_j^i, p_l^k)$ , is calculated using equation 3, whereas thread title similarity,  $LSim(p_j^i, p_l^k)$  is calculated in the same way as content similarity. The only difference lies in the text content which in this case is the text of thread title, as shown in equation 4. Finally, the overall similarity,  $Sim(p_j^i, p_l^k) \in [0, 1]$ , is calculated by aggregating all the above-mentioned measures using equation 5, where  $\alpha, \beta, \gamma$  and  $\delta$  are constants such that  $\alpha + \beta + \gamma + \delta = 1$ .

$$ASim(p_j^i, p_l^k) = I_{[author(p_j^i) == author(p_l^k)]} \quad (3)$$

$$LSim(p_j^i, p_l^k) = CSim(title(p_j^i), title(p_l^k)) \quad (4)$$

$$Sim(p_j^i, p_l^k) = \alpha \times CSim(p_j^i, p_l^k) + \beta \times TSim(p_j^i, p_l^k) + \gamma \times ASim(p_j^i, p_l^k) + \delta \times LSim(p_j^i, p_l^k) \quad (5)$$

For clustering process, we start with assigning each post into a separate cluster. Let us suppose there are  $n_0$  number of total posts in the forum and at time  $t = 0$  it starts with  $C^0 = \{c_1^0, c_2^0, \dots, c_{n_0}^0\}$  as the set of clusters assuming that every post is dissimilar from others. At each iteration,  $t$ , in the clustering process, a similarity matrix  $\Phi_{n_t \times n_t}^t$  is maintained to contain the similarity information between each pair of clusters. For the initial similarity matrix,  $\Phi_{n_0 \times n_0}^0$ , at  $t = 0$  its values are calculated as a similarity measure between each pair of posts as shown in equation 6, where  $p_i \in c_i^0$  and  $p_j \in c_j^0$ .

$$\Phi_{ij}^0 = Sim(p_i, p_j) \quad (6)$$

At time,  $t$ , each value in the matrix,  $\Phi_{n_t \times n_t}^t$ , is compared with the similarity threshold value,  $\epsilon$ . The pair of clusters for whom this value is found to be greater are added to the set of pairs,  $\Lambda^t$ , that need to be merged. After collecting all the cluster pairs that show a sign to get merged, they are ranked by their corresponding matrix values. Starting with the top ranking pair, the two clusters are merged to form a unified cluster and all those pairs in  $\Lambda^t$  containing either of the two sub-clusters are removed from the set. The merging process is continued until  $\Lambda^t$  becomes empty. After the completion of merging, it proceeds to next iteration,  $t + 1$ , the new set of clusters becomes  $C^{(t+1)}$  with number of clusters as  $n_{(t+1)} < n_t$ , and the new matrix becomes  $\Phi_{n_{(t+1)} \times n_{(t+1)}}^{(t+1)}$ .

Each cluster,  $c_i^t$ , at time,  $t$ , keeps information about all its posts grouped into two sub-clusters,  $c_k^{(t-1)}$  and  $c_l^{(t-1)}$ , if  $c_i^t$  is a result of merging  $c_k^{(t-1)}$  and  $c_l^{(t-1)}$ , else  $c_i^t$  contains a single cluster of posts,  $c_k^{(t-1)}$ , the same as it was in last iteration. Each value,  $\Phi_{ij}^t$ , in the new matrix is calculated using equation 7, where  $|c_i^t|$  and  $|c_j^t|$  denote the number of sub-clusters in  $c_i^t$  and  $c_j^t$ , respectively. After  $t$  iterations, when there remains no  $\Phi_{ij}^t$  value greater than the  $\epsilon$ , the terminating condition becomes true and the final clusters are returned as grouped posts.

$$\Phi_{ij}^t = \frac{\sum_{c_k^{(t-1)} \in c_i^t, c_l^{(t-1)} \in c_j^t} \Phi^{(t-1)}(k, l)}{|c_i^t| \cdot |c_j^t|} \quad (7)$$

After applying the above-mentioned clustering process, the posts of a forum are grouped into unpredictable number of clusters. Each of the obtained clusters comprise a set of uniquely focalized comments to a certain level decided by the  $\epsilon$ , which binds their authors together to form a like-minded network of people. The richness of information in these clusters make us to present them as cliques that are susceptible to discover a plethora of undiscovered facts by going through a network analysis inside each clique.

### IV. EXPERIMENTAL SETUP AND RESULTS

For experiment, we have considered the popular neo-Nazi Web forums of stormfront<sup>1</sup>, which has been identified as the first major hate-site on the Web [8]. As of 4<sup>th</sup> Feb. 2012, it contained 58 different forums with a total of 619,634

<sup>1</sup><http://www.stormfront.org>

threads and 8,376,678 posts. We have selected 10 threads from the “*eActivism and Stormfront Webmasters*” forum under the category of *Activism*, which consists of 1,698 threads including 11,210 replies and 2,271,494 views.

Firstly, our crawler module based on `craler4j` and parser module developed for the `vBulletin` platform crawls and parses the forum webpages to extract all meaningful pieces of information from them. We extract all unigrams, bigrams and trigrams from the body-texts of a total of 934 posts. The values of the constants,  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  are determined as 0.167, 0.333 and 0.5, respectively. For computing time similarity, the time difference is calculated in unit of hours, and value of  $\beta_1$  is set to 0.995. All similarity measures are computed for each pair of posts which are then integrated to calculate the overall similarity using values of the constants  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  that are set to 0.7, 0.1, 0.1 and 0.1, respectively. The dominating nature of *content similarity*, observed experimentally, guide us to fix such a high value for the constant  $\alpha$ . Thereafter, the clustering algorithm is executed varying the threshold  $\epsilon$  from 0.20 to 0.50 at an interval of 0.05. On the other hand, manual grouping of posts is done in a strict manner, i.e., posts are grouped together only if they seemed to be exactly in the same context or substitutable to each other, and thus the set of 934 posts is grouped into 207 cliques. However, generating the gold standard set in this way may vary from person to person and also with the level of similarity assumed to merge two posts. The line chart shown in Figure 1 presents the trend of increasing number of generated clusters as the value of  $\epsilon$  increases. We found that as we move away from the line  $\epsilon = 0.3$  in either side, the difference between the number of automatically generated and manually identified clusters goes on broadening, which leads to a fall in accuracy of the algorithm.

Table I presents the summary of results obtained in terms of the evaluation metrics. We can see that due to minimum difference between the number of manually-generated and system-generated cliques at  $\epsilon = 0.3$ , the system has shown its best results with  $F_P$  as 0.825 and  $F_B$  as 0.804.

For comparison task, we have used the metrics  $F_{\alpha=0.5}$  (or  $F_P$  at  $\alpha = 0.5$ ) and  $F_{B-Cubed}$  (or  $F_B$ ).  $F_P$  is defined as the harmonic mean of *purity* and *inverse purity* measures as defined in equations and , respectively.  $F_B$  is defined as the harmonic mean of overall *B-Cubed precision* and *B-Cubed recall* values that are computed as the mean of the precision and recall values, respectively for each element (or post).

$$Purity = \sum_i \frac{|C_i|}{n} \maxPrecision(C_i, L_j) \quad (8)$$

$$InversePurity = \sum_i \frac{|L_i|}{n} \maxPrecision(L_i, C_j) \quad (9)$$

## V. CONCLUSION

In this paper, we have presented a novel agglomerative clustering method to identify cliques in dark Web forums. A similarity function based on both contextual and temporal

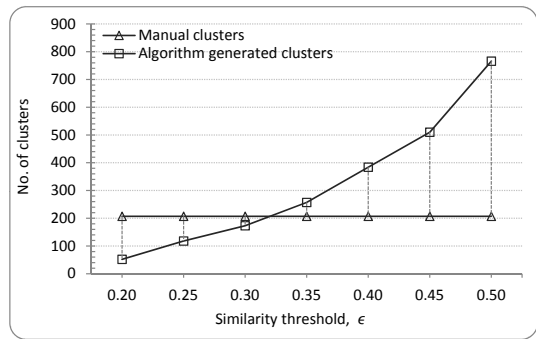


Fig. 1. Number of clusters generated by the proposed clustering method at different values of  $\epsilon$

TABLE I  
EVALUATION RESULTS OF THE PROPOSED CLUSTERING METHOD AT DIFFERENT VALUES OF  $\epsilon$

Evaluation Metric	Threshold Value ( $\epsilon$ )						
	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Manual Cliques	207	207	<b>207</b>	207	207	207	207
Auto-generated Cliques	52	118	<b>173</b>	257	384	510	766
$F_P$	0.639	0.771	<b>0.825</b>	0.807	0.740	0.664	0.598
$F_B$	0.534	0.685	<b>0.804</b>	0.794	0.707	0.619	0.486

coherence is proposed and applied to group posts into clusters. The identified clusters are presented in the form of cliques characterized by the homogeneity of posts in them, which in turn establish the homogeneity of authors and threads as well.

## ACKNOWLEDGMENT

The authors would like to thank King Abdulaziz City for Science and Technology (KACST) and King Saud University for their support. This work has been funded by KACST under the NPST project number 11-INF1594-02.

## REFERENCES

- [1] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, “Uncovering the dark web: A case study of jihad on the web,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 8, pp. 1347–1359, Jun 2008.
- [2] J. Qin, Y. Zhou, and H. Chen, “A multi-region empirical study on the internet presence of global extremist organizations,” *Information Systems Frontiers*, vol. 13, no. 1, pp. 75–88, Mar 2011.
- [3] A. Abbasi, H. Chen, S. Thoms, and T. Fu, “Affect analysis of web forums and blogs using correlation ensembles,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 20, no. 9, pp. 1168–1180, Sep 2008.
- [4] C. C. Yang, X. Tang, and B. M. Thuraisingham, “An analysis of user influence ranking algorithms on dark web forums,” in *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ser. ISI-KDD ’10, New York, NY, USA: ACM, 2010, pp. 10:1–10:7.
- [5] C. C. Yang and T. D. Ng, “Web opinions analysis with scalable distance-based clustering,” in *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, ser. ISI’09, Piscataway, NJ, USA: IEEE Press, 2009, pp. 65–70.
- [6] C. C. Yang, X. Tang, and X. Gong, “Identifying dark web clusters with temporal coherence analysis,” in *Proceedings of the 2011 IEEE international conference on Intelligence and security informatics*, ser. ISI’11, IEEE Press, 2011, pp. 167–172.
- [7] D. Liu, D. Percival, and S. E. Fienberg, “User interest and interaction structure in online forums,” in *Proc. of the 4th Int’l AAAI Conf. on Weblogs and Social Media*, ser. ICWSM ’10, The AAAI Press, 2010, pp. 283–286.
- [8] J. Kaplan and L. Weinberg, *The Emergence of a Euro-American Radical Right*. New Brunswick, N.J.: Rutgers University Press, 1998.