Final version of the accepted paper. Cite as: *T. Anwar and M. Abulaish, Ranking Radically Influential Web Forum Users, IEEE Transactions on Information Forensics and Security, Vol. 10, Issue 6, pp. 1289-1298, 2015.*

Ranking Radically Influential Web Forum Users

Tarique Anwar and Muhammad Abulaish, *Senior Member, IEEE*

***Abstract*—The growing popularity of online social media is leading to its widespread use among the online community for various purposes. In the recent past, it has been found that the Web is also being used as a tool by radical or extremist groups and users to practice several kinds of mischievous acts with concealed agendas and promote ideologies in a sophisticated manner. Some of the Web forums are predominantly being used for open discussions on critical issues influenced by radical thoughts. The influential users dominate and influence the newly joined innocent users through their radical thoughts. This paper presents an application of collocation theory to identify radically influential users in Web forums. The radicalness of a user is captured by a measure based on the degree of match of the commented posts with a threat list. Eleven different collocation metrics are formulated to identify the association among users, and they are finally embedded in a customized PageRank algorithm to generate a ranked list of radically influential users. The experiments are conducted on a standard data set provided for a challenge at ISI-KDD'12 workshop to find radical and infectious threads, members, postings, ideas, and ideologies. Experimental results show that our proposed method outperforms the existing UserRank algorithm. We also found that the collocation theory is more effective to deal with such ranking problem than the textual and temporal similarity based measures studied earlier.**

***Index Terms*—Social media analysis, Security informatics, Radical user identification, Users collocation analysis.**

[1]

I. Introduction

IN the recent past, it has been found that the Web is being used as a tool to practice several kinds of mischievous acts with concealed agendas and promote ideologies in a sophisticated manner [1]. Infiltration of extremist groups, hate groups, racial supremacy groups, and terrorist organizations on the Web with hundreds of multimedia websites, online chat rooms and Web forums is posing grievous threats to our societies as well as the national security. The multimedia websites provide support for their psychological warfare, fund-raising, recruitment, and propagation of their agendas, whereas chat rooms and Web forums promote their strategies and ideologies through discussions with naive users. Often the public discussions among differently minded extremist groups lead to irascible talks accompanied with abusive languages, and promote online hate and violence. Web forums are recognized for their exhaustive, vivid and non-spontaneous nature of discussions that are archived for later reference [2]. Previous studies have found Web forums as the most active medium being used for this purpose [3]. Research on identifying radical and infectious threads, members, postings, ideas and ideologies in Web forums for tracking the grievous threats posed by the active extremist and hate groups has gained considerable attention of the research community. The portion of the Web circumscribing the sinister objectives of extremist groups is said as the *Dark Web*, and specifically the Web forums with substantial prevalence of activities supporting extremism are said as *Dark Web Forums* [4]. Another class called *Gray Web Forums* [5] refer to the forums in which the discussions focus on topics that might potentially encourage biased, offensive, or disruptive behaviors and may disturb the society or threaten public safety. They include topics like pirated CDs, gambling, spiritualism, bullying, and online-pedophilia.

The global extremist groups, ranging from US domestic racist and militia groups to Latin American guerilla groups and radically motivated Islamic military groups, have created thousands of websites that support psychological warfare, fund-raising, recruitment, and distribution of propaganda materials [1]. To keep their agenda alive and attract more supporters or sympathizers, they always maintain certain level of publicity and influence in the community for their causes and activities [6]. Prior to the Internet and social media era, they used to maintain their influence through the mainstream traditional media, but as the Internet and social media flourished, their intent of getting influence found a sophisticated way to promote their ideology. They predominantly use the Dark Web forums for expression and dissemination of their ideologies [7], [3].

Role of influential users: Due to enormous and rapid growth of user-generated content on social media sites, a significant portion of such data remains just a noise, and users generally avoid going through every comment posted by others. There always exist some users who develop some relationship of trust with other members by their activeness and quality of comments, and their comments always receive significant attention of a large community [8]. These are the *influential users*, sometimes also called *community leaders*, who play a leading and dominating role in the community, and their activities and comments greatly affect the sentiments of others [9]. For example, the popularity of a personal blog is completely dependent on the owner's influence, where a majority of users remain silent spectators following the few influential leaders. As a result, be it a political campaign or a product marketing or

T. Anwar is currently with SUCCESS, Swinburne University of Technology, Australia. This work was done when he was with CoEIA, King Saud University, Saudi Arabia. E-mail: tAnwar@swin.edu.au

M. Abulaish (corresponding author) is currently with Dept. of Computer Science, Jamia Millia Islamia, New Delhi, India. This work was done when he was with CoEIA, King Saud University, Saudi Arabia. E-mail: mAbulaish@jmi.ac.in

an extremist ideology propagation, influential users most of the time find it very easy to convince the silent spectators and promote their ideologies. *Influential hypothesis* [10] comprises two fundamental claims about inter-personal influence: i) some people are more influential than others, ii) the same people are very important because of their direct influence on their peers as well as a disproportionate indirect influence on the much larger community of which both they and their immediate influences are a part. In Dark Web forums, the leaders of extremist groups maintain their own influence strategically to win over the sentiments of silent spectators by their convincing approach. Previous studies have found that it is an important problem and a challenging task to identify such influential leaders of radical groups propagating through the Dark Web forums [11]. Some factors that characterize influential members in a network are *high connectivity* in the network, *interest* on the network domain, *leadership* or *asymmetric influence* over the network, and higher level of *cascading influence*.

Our contribution: We make the following key contributions in this paper. i) An application of collocation theory to rank radically influential Web forum users who are persuaded by fanatics of hate, extremism, and war. ii) A measure to compute the degree of radicalness of a user based on the degree of match her posts with a manually crafted threat list. iii) A contingency table generation method for a pair of users based on their interaction and collocation in different threads, which is used to define eleven different collocation-based association metrics. The association measures along with radicalness measure are embedded in a customized PageRank algorithm to generate ranked list of radically influential users. iv) A manual analysis of a standard Web forum data set (provided for a challenge at ISI-KDD'12 workshop), and establishment of five different criteria to define users' radicalness and to calculate radicalness score for each users.

The rest of the paper is organized as follows. Section II presents a review of the related works, followed by definition of radically influential users in Section III. Section IV presents the proposed method, and Section V presents experimental results and their evaluation. Finally, Section VI concludes the paper with few important future research directions.

II. Related Work

With the rapid growth of user-generated contents, the study of information propagation and influential users in social networks has become crucial to a plethora of related analysis problems. This section presents some of the important previous works on influential user identification and Dark Web research.

Influential user identification: A majority of previously studied works on the problem of influential user identification have been done in a business intelligence orientation for marketing products through targeted influential users or viral marketing [12], [8]. Some other objectives are information dissemination [13], community leader identification [14], and expertise discovery [15].

[12] worked on the social network formed from collaborative ratings, and modeled it as Markov random fields, considering each customer's product buying probability as a function of both its intrinsic desirability for the customer and the influence of others. [16] utilized the dynamics of voting on *digg* posts to rank influential users. They defined an empirical measure of influence based on the number of in-network votes that the post of a user receives. [17] devised a greedy approach based discrete-optimization model to maximize the spread of influence through a social network. However, [13] found that the computational cost of a conventional greedy approach to identify influential nodes in a network is very high, and consequently they proposed a method of estimating marginal gains on the basis of bond percolation and graph theory. [18] performed a statistical analysis on *email* network-based marketing and established a hypothesis for a direct affect of network linkages on product/service adoption. [19] applied the influence models proposed by [17], in addition to applying algorithms like PageRank, in blogosphere. They also discussed the importance of splog removal and its implications on influence models. [9] came up with a comprehensive definition of influential bloggers and the challenges associated with their identification. Using an influence graph of blog posts, they defined some measures to find influential blog-posts and bolggers. [15] proposed ExpertiseRank to rank the Java expertise using forum threads and posts in the popular `Java Forum`. [20] contributed towards online healthcare social networks, specifically the Swine Flu online forum which is a sub-community of `MedHelp`. Based on the concepts of PageRank algorithm, they proposed UserRank to identify the influential users using content similarity and response immediacy. It is shown as out-performing PageRank, in-degree and out-degree rankings. In [11], they also showed the application of UserRank algorithm in the domain of Dark Web forums.

Dark Web research: A recent work [1] described how all major extremist organizations in the world, ranging from the US domestic racist and militia group to Latin American guerrilla groups and Islamic military groups, show their presence on the Internet. They also performed a multi-region empirical study on these organizations' Internet presence. Set up in 1995 by Don Black, the `Stormfront` (http://www.stormfront.org), a White nationalist and supremacist neo-nazi Web forum, was identified as the first major hate-site on the Web [21]. AI Lab of the university of Arizona started to automatize the complete monitoring system and came up with their Dark Web Portal with several functionalities for data collection as well as analysis [22]. The research on the Dark Web starts from the automatic accumulation of extremist websites and all related Web data in a repository [23], [24], on which the data mining techniques are applied. It includes content analysis [25], [26], [27], [3] and user interaction analysis [28], [29], [11] as the main research area to analyze the sentiments and affects on the whole community. Ranging from automatic to semi-automatic processes, several attempts have been made in the past for crawling and downloading of webpages from the surface Web as well as hidden Web [23], [24]. [24] being the most recent is a language-independent incremental crawler focussed on extremist groups from three specific regions – US Domestic, Middle East, and Latin America/Spain. [25] differentiate affect analysis from sentiment analysis by characterizing it as assigning text with emotive intensities across a set of mutually inclusive and possibly correlated affect classes. [26] performed a content analysis of `Ansar` forum for topic-based ranking of posts. Clustering of posts and threads has also been attempted in several studies to get communities with overlapping interests [3]. [27] analyzed `Ansar` forum for a clustering-based unsupervised anomaly detection with an objective to provide a robust, focus-of-attention mechanism to identify emerging threats in time-dependent, unlabeled datasets. In [28], the authors present a hybrid approach to generate a social network from the interactions in threaded discussions of a forum. [29] consider a Dark Web forum as virtual communities of interests (VCoI) and performed a topic-based social network analysis of the `Ansar` community with an objective to discover key members. Based on the concept of page rank algorithm, [11] devised the UserRank algorithm to rank influential users using content similarity and response immediacy.

Although this algorithm is proposed for dark Web forums, it lacks domain-specific properties. To the best of our knowledge, no such work has been done till date to identify radically influential users in a Web forum.

III. Radically Influential Users

Radicalization is defined as galvanization of people by fanatic thoughts beyond the norm to an extreme antagonistic political, religious, racial, nationalist or any other ideology. The people undergoing this galvanization usually have no personal values for ethics and rationalism, and are characterized by the term *radical*. This kind of thoughts arouse in minds when they feel of some unjust or discrimination happened with them either directly or indirectly, though it actually may be false. These thoughts are sometimes triggered by their personal involvement (e.g., death of a close relative or friend), political involvement (e.g., being a follower of a political or religious belief), and social involvement (e.g., racism, nationalism). Thus, their hostility may be against a race, or a political party, or a religion, or a nation, or any organization with a mass of followers. These are the most committed followers of a cause who commit such ill-willing acts of terrorism.

Cha et al. [30] contend that *influence* is very hard to define concretely or measure tangibly, despite the large number of existing theories of sociology. In fact, formulation of the exact definition remains critical to the focus in mind for which it needs to be defined. In the very first step, it can be approximated as something attained by the *activeness* of a person. However, [9] differentiated them very clearly. Just being active in communication does not make someone influential in a social network. Rather, an influential person can remain inactive and maintain her

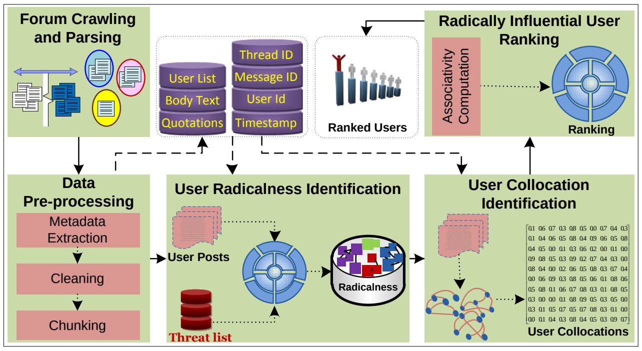


Fig. 1. Work-flow of the proposed ranking method

own dignity, whereas a person participating actively in discussions may be non-influential (e.g. because of her repeated non-sense replies or suggestions that is of no interest to others). Influential users generally get a very good response from others in their comments, and it differentiates them from the spammers, who in spite of being active do not receive much attention. In their study to identify influential bloggers, [9] came up with four major factors that make a blog post influential, which are *recognition*, *activity generation*, *novelty*, and *eloquence*. Trusov et al. [8] define *influential users* as members whose increased (or decreased) usage or activeness in social media sites reflect the same trend in other connected members.

It can now be established that the radically influential users are characterized by two key properties – *radicalness* and *influence*. There can be two different approaches to tackle the problem of radically influential user identification. The first one is to consider it as a *two-stage sequential* problem, in which each stage identifies the users' measure for one of the two properties. The two stages remain completely independent of each other where the first stage is followed by the second, and the output of the first is fed into to the second to get the final result. There can be two possible ordering for this approach. The final output with these different orderings will differ from each other depending on the nature of data. This introduces another problem as which ranking to consider as more promising. A solution to this problem lies in an intelligent integration of the two properties into a single property and then following a *one-stage parallel* ranking approach to identify radically influential users. We follow this parallel approach.

IV. Proposed Ranking Method

The proposed method starts with crawling and preprocessing the forum data, followed by user radicalness identification, user collocation identification, and finally ranking the users based on a customized PageRank algorithm, as shown in figure 1.

A. Forum Crawling and Preprocessing

The process starts with a data crawling and preprocessing step in which the URL of the forum home page is passed to the forum crawler, which crawls all relevant webpages and eliminates the duplicates heuristically. A platform-specific parser module is employed to extract the meaningful snippets from the crawled webpages, which are then passed to the data preprocessing module. The metadata extraction task works in close coordination with the parser module to extract all relevant metadata. The obtained data is organized as a collection of threads having a unique id and title; each thread containing one or more posts having a post id, time-stamp, body text, author, and quotations. The body text is additionally processed through some cleaning and chunking mechanisms to remove the noise and crystalize into individual meaningful pieces of information.

B. Measuring Radicalness

A few previous works attempted to identify the radical elements based on discussion contents [26], [27]. However, the foundation of their automatic radical identification process is laid on a set of manually crafted list of threat words that are typically found in radical texts. In [27], the author manually crafted the list of threat words as a subset of the pruned list of words from the `Ansar` forum, which consists of 370 English and Arabic words. The forum is believed by many people as representing radical Jihadi ideology. We noticed that the threat list is quite long, and most of the words in the list are also used in general situations. For example, *honor*, *hard*, *puppet*, and *movement* are general terms and these are very likely to mark a non-radical message as a radical. Because the list is manually crafted, there needs to be strong rationality to use the words for characterizing radicalness. We reduced the list to a set of 23 highly focused words based on our observation and perception, and added two new words – *shaheed* and *taliban*, shown in Table I. All the words in the list except a few like *support* and *victory*, clearly express the sense of radicalness, and the exceptions, although pose a non-radical sense in usual cases, but in the context of radicalization they stand for a specific meaning. In real situations, it is very likely that the potentially radical members avoid using the obvious radical terms and prefer using some disguise of words. Also the terms could be acronyms or synonyms or in different languages. To handle these real scenarios, the list needs to be updated regularly with time. Incremental learning based on Naive Bayes classification can be used to learn and introduce such new terms. Shorter lists may give some radical members a chance to evade, whereas longer lists (including some general terms that are perhaps also radical in a sense) may mark even innocents as radicals. Therefore one needs to be extreme careful while preparing or updating the threat list.

TABLE I
THREAT LIST FOR RADICAL JIHADI IDEOLOGY

Terrorism	Blast	Killing	Bombing	War
Missile	Explosive	Insurgent	Al-Qaeda	Mujahideen
Destruction	Murder	Clash	Jihad	Attack
Crime	Violence	Detonate	Suicide	Operation
Martyrdom	Support	Shaheed	Taliban	Victory

Let Ω denotes the set of words in the threat list. A radicalness measure ρ is assigned to each user u_i of the forum being studied, based on the existence of each word Ω_j in each message post p_k^i of u_i using equation 1, where $\text{exists}(\Omega_j, p_k^i)$ is a binary function which returns 1 if Ω_j exists in p_k^i, otherwise 0.

$$\rho(u_i) = \frac{\sum_{p_k^i \in posts(u_i)} \sum_j \text{exists}(\Omega_j, p_k^i)}{\max\left\{\sum_{p_k^i \in posts(u_i)} \sum_j \text{exists}(\Omega_j, p_k^i)\right\}} \quad (1)$$

C. Identifying Collocations

It has been found that there exists an intimate relationship between the users interacting in same thread, and in the context of Web forums the term *collocation* can be defined as the association of users co-interacting in same threads. Therefore we apply the collocation theory to study the associativity of different users, and estimate their influence while propagating an ideology through their interactions. To capture this information, a *contingency* table, shown in Table II, is constructed for each pair of users, where U is the set of users, and u_i and u_j represent two individual users. In this table, a denotes the number of instances (or threads) in which u_i and u_j have co-occurred, b denotes the number of instances (or threads) in which u_i has co-occurred with all other users in a thread, $(b - a)$ denotes the number of instances (or threads) in which u_i has co-occurred with all other users except u_j in a thread. Similarly, all other values in this table denote the number of instances (or threads) in which interactions have taken place between the corresponding users.

TABLE II
CONTINGENCY TABLE FOR A PAIR OF FORUM USERS (u_i, u_j)

	u_j	$U - u_j$	U
u_i	a	$(b - a)$	b
$U - u_i$	$(c - a)$	$(d - c - b + a)$	$(d - b)$
U	c	$(d - c)$	d

D. Defining Association Metrics

This subsection defines 11 statistical association metrics based on user collocation measures that determine the associativity between a pair of users using Table II in different statistical ways.

Co-occurrence Frequency (μ_1): For a pair of users u_i and u_j, the co-occurrence frequency, $\mu_1(u_i, u_j)$, is defined as the number of instances or threads in which both of them participate, i.e., $\mu_1(u_i, u_j) = a$. The intuition behind this feature is that the more a pair of users' comments co-occur in threads the higher their associativity. The active users in a forum comment frequently to respond to most of the threads and they are likely to co-occur with most of the users in the forum. The limitation of this metric lies in its biasness towards such kind of active users. It does not look into any other information, like total comments or the portion of co-occurrences with a specific user out of the total co-occurrences.

CF-ITF (μ_2): In the field of information retrieval, there exists an immense contribution of TF-IDF (term frequency-inverse document frequency) [31] for various text processing tasks. For a given term, it multiplies its frequency with the logarithm of the inverse of the portion of documents in which the term appears. Its composition makes it to reflect the importance of the terms in a document collection. In a Web forum, several users participate in threaded discussions and each of them co-occur with others through their message posts in the discussions. Therefore, along the lines of TF-IDF formulation, CF-ITF (co-occurrence frequency-inverse thread frequency) between a pair of users u_i and u_j is defined as their co-occurrence frequency a multiplied by the logarithm of the inverse of the portion of threads in which u_i co-occurs with others. Using Table II, the CF-ITF of a pair of users u_i and u_j is calculated using equation 2.

$$\mu_2(u_i, u_j) = a \times \log\left(\frac{d}{b+1}\right) \tag{2}$$

PMI (μ_3): PMI (point-wise mutual information) [31] is a standard measure which is used in the fields of information theory and statistics to determine the association or dependence of two probabilistic events. For a pair of discrete random variables x and y, it is defined as the discrepancy between their co-occurrence probability given their joint distribution and their co-occurrence probability given only their individual distributions, assuming independence, and formulated as $\mathrm{PMI}(x,y) = \log_2 \frac{\mathrm{prob}(x,y)}{\mathrm{prob}(x)\times\mathrm{prob}(y)}$. Using Table II, we define the PMI-based association metric for a pair of users u_i and u_j using equation 3. In this equation, 1 is added to the numerator to avoid the case of $\log_2 0$, which generally happens due to no interaction between the respective users.

$$\mu_3(u_i, u_j) = \log_2 \frac{(a \times d) + 1}{b \times c} \tag{3}$$

Cosine (μ_4): Cosine similarity [31] is used to measure the strength of association between a pair of objects having feature vectors. It is formulated as $\mathrm{cosine}(X,Y) = \frac{|X \bigcap Y|}{\sqrt{|X| \times \sqrt{|Y|}}}$, where X and Y represent the feature vectors of same dimension. We define this metric based on the contingency table to compute the association between two users u_i and u_j using equation 4.

$$\mu_4(u_i, u_j) = \frac{a}{\sqrt{b} \times \sqrt{c}} \tag{4}$$

Overlap (μ_5): Overlap [31] is also used for the same purpose as cosine measure, but with slight difference in its formulation, $\mathrm{overlap}(X,Y) = \frac{|X \bigcap Y|}{\min(|X|,|Y|)}$.

Using the contingency table, we define the overlap-based association metric for two users u_i and u_j using equation 5.

$$\mu_5(u_i, u_j) = \frac{a}{\min(b, c)} \tag{5}$$

Dice (μ_6): Dice coefficient [31] is another association measure formulated as $\mathrm{Dice}(X,Y) = \frac{2 \times |X \bigcap Y|}{|X|+|Y|}$.

Using the contingency table, we define the dice-based association metric for two users u_i and u_j using equation 6.

$$\mu_6(u_i, u_j) = \frac{2 \times a}{b + c} \tag{6}$$

Jaccard (μ_7): For a given pair of sets, say X and Y, the Jaccard similarity coefficient [31] is measured as the ratio of their intersection to their union, $\mathrm{Jaccard}(X,Y) = \frac{|X \bigcap Y|}{|X \bigcup Y|}$.

With the help of the contingency table, we define the Jaccard-based association metric for two users u_i and u_j using equation 7.

$$\mu_7(u_i, u_j) = \frac{a}{b + c - a} \tag{7}$$

Chi-square (μ_8): The chi-square (χ^2) measure [31] is generally used as a test to determine the difference between the distribution of an actually observed sample and another hypothetical or previously established distribution that is normally expected. It always tests the *null hypothesis*, which states that there is no significant difference between the expected and observed result, and the deviation of observed outcome from the expected distribution is used by the investigator to conclude that whether the reason of deviation is just by chance or something else. It is calculated as the sum of the squared differences between observed and expected values scaled by the magnitude of the expected values, $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. In this work, this measure is used to determine the dependency of a pair of users established by their interactions in the threaded discussions. Using the contingency table, we define the chi-square-based association metric using equation 8.

$$\mu_8(u_i, u_j) = \frac{d \times \{a \times (d - b - c + a) - (b - a) \times (c - a)\}^2}{b \times c \times (d - c) \times (d - b)} \tag{8}$$

LLR (μ_9): Similar to chi-square, the LLR (log likelihood ratio) [31] is another approach for hypothesis testing, which is considered more appropriate for sparse data. It provides a means to compare the likelihood of two alternate hypotheses and defined as the ratio of two likelihoods. Using the contingency table, the LLR-based association metric for two users u_i and u_j is defined using equation 9.

$$\begin{aligned} \mu_9(u_i, u_j) = a \times \log_2 \frac{(a \times d) + 1}{b \times c} + (b - a) \times \log_2 \frac{((b - a) \times d) + 1}{b \times (d - c)} + \\ (c - a) \times \log_2 \frac{(d \times (c - a)) + 1}{c \times (d - b)} + \\ (d - b - c + a) \times \log_2 \frac{(d \times (d - b - c + a)) + 1}{(d - b) \times (d - c)} \end{aligned} \tag{9}$$

Phi Coefficient (μ_{10}): The phi coefficient [31] is a measure of association between two variables, which is derived from their previously mentioned chi-square measures. With the help of contingency table, the phi coefficient-based association metric for two users u_i and u_j is defined using equation 10, where χ^2 is the chi-square value.

$$\mu_{10}(u_i, u_j) = \sqrt{\frac{\chi^2}{d}} \tag{10}$$

Contingency Coefficient (μ_{11}): Contingency coefficient [31] is another association measure, which is defined using equation 11.

$$\mu_{11}(u_i, u_j) = \sqrt{\frac{\chi^2}{d + \chi^2}} \tag{11}$$

E. Ranking

It is generally not practical that a subset of users exist as radically influential and others not; rather it is like a property that exists in every user with varying intensities. Therefore, we consider the problem of identifying radically influential users as a ranking problem. Both the individual properties of radicalness and influence in a user are very much regulated by the other users with whom the former interacts, in addition to one's own default properties. Therefore, the interaction linkages act crucially to determine the overall magnitude. For this nature of the influence ranking problem, some previous works found the concept of PageRank algorithm as much suitable to establish its foundation [19], [15], [11].

The PageRank algorithm computes a ranking of webpages to find their probable importance to Web navigators and page authors [32]. Authors of webpages generally hyperlink important terms in them to refer to a further detail in other webpages. It considers these Web hyperlinks as recommendations made by the directing page for the page to which the former is linking. To compute the ranking score of webpages, each of them is initialized with a small value as their page rank score ($\mathrm{PR}(p_i)$), and the linkages (L) among them are iteratively used to compute their new page rank score ($\mathrm{PR}(p_j)$) using equation 12, where $d \in [0,1]$ is the damping factor typically set to 0.85 [32], $\mathrm{prob}(p_j|p_i) = \frac{1}{out-degree(p_i)}$ is the transition probability from webpage p_i to webpage p_j, and $l_{ij} \in L$ is the hyperlink from page p_i to p_j. The iteration process is continued until a convergence is achieved and the scores at that instance are accepted as their final page rank scores.

$$\mathrm{PR}(p_j) = (1-d) + d \times \sum_{\forall p_i : l_{ij} \in L} \mathrm{prob}(p_j|p_i) \times \mathrm{PR}(p_i) \qquad (12)$$

The proposed radically influential user ranking method is based on the concept of PageRank algorithm. Threaded discussions among users in a Web forum are used to construct a directed graph by adding each user in the forum as a node, and each user interaction as a directed link. Unidirectional links from all commenters to the thread initiator and bi-directional links between each pair of commenters are established for each thread in the graph Each user node is initialized with a small value as its page-rank score, and just like the PageRank algorithm, the directed linkages among them are used iteratively to keep on updating their rank scores, until a convergence is achieved. Equation 13 is used to compute updated user rank scores, $rank(u_j)$, iteratively, where $d \in [0,1]$ is the damping factor set to 0.85 as in [32], $\mathrm{R}(u_i,u_j)$ is the radicalness measure of interactions between u_i and u_j, $\mathrm{I}(u_j|u_i)$ is the influence transmission probability from u_i to u_j, and $l_{ij} \in L$ is the directed link from u_i to u_j.

$$\mathrm{rank}(u_j) = (1-d) + d \times \sum_{\forall u_i : l_{ij} \in L} \{\log_2\left(\mathrm{R}(u_i,u_j) \times \mathrm{I}(u_j|u_i) + 1\right) \times \mathrm{rank}(u_i)\} \qquad (13)$$

One of the two major information components, the radicalness measure $\mathrm{R}(u_i,u_j)$, is computed as the summation of the individual radicalness values $\rho(u_i)$ and $\rho(u_j)$ (Equation 1), as shown in Equation 14.

$$\mathrm{R}(u_i,u_j) = \log_2\left(\rho(u_i) \times \rho(u_j) + 1\right) \qquad (14)$$

The other major information component, i.e., influence transmission probability, $\mathrm{I}(u_j|u_i)$, from u_i to u_j is computed using equation 15, where $\mu(u_i,u_j)$ is the value for one of the association metrics defined between u_i and u_j in section IV-D, and $l_{ik} \in L$ is the directed link from u_i to u_k.

$$\mathrm{I}(u_j|u_i) = \frac{\mu(u_i,u_j)}{\sum_{\forall u_k : l_{ik} \in L} \mu(u_i,u_k)} \qquad (15)$$

In equations 13 and 14, we apply the logarithm transformation as $\log_2(x \times y + 1)$ to get the combined effect of two quantities x and y. The reason is that when quantities having values less than 1 are multiplied, the result tends to go lower and decrease the overall effect. The lower the values are, severe is the effect. Logarithm function transforms the relative spacing between the different values to normalize this effect. Furthermore, as $x \times y \in [0,1]$, 1 is added to make its range as $[1,2]$, so that $\log_2() \in [0,1]$.

V. Experiments and Evaluation

To evaluate the soundness and accuracy, we made a significant effort in generating a benchmark through manually ranking radically influential users in the experimental data set[2] explained in Section V-B.

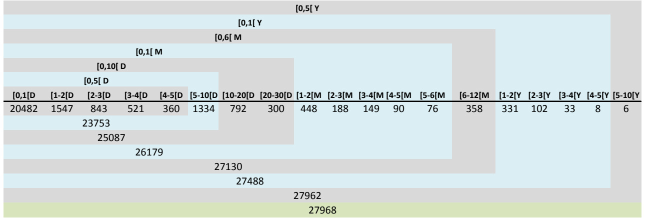
A. Data Set and its Lifespan

The experimental data set[3] is a set of threads provided for a challenge[4] at the ISI-KDD'12 workshop to find radical and infectious threads, members, postings, ideas and ideologies. It is generated by a panel of terrorism study experts by crawling the `Islamic Awakening` Web forum, considered by many as a dark Web forum, where participants are radically motivated for terror related causes. It is composed of a total of 1,29,425 message posts commented as response to a total of 27,968 threads by 2803 users. As per our knowledge, it includes all discussions carried on in the forum from April 28, 2004 to May 20, 2010. Figure 2 visualizes the lifespan of threads with the help of a *span-line*, where the upper-half is the span-line comprising different spans of time, and the lower-half shows the number of threads having the corresponding lifespan. Lifespans are denoted using open and closed intervals followed by a character D, M, or Y, where D stands for Day, M stands for Month, and Y stands for Year, e.g., `[0,1)D` stands for a lifespan of greater than or equal to 0 day but less than 1 day. A vast majority of threads (i.e., 20482 or 73.23 %) ended up in less than a day, and 26179 or 93.6 % of threads ended up in less than a month. However, the longest thread continued up to about 7 years.

[2] A complete set of our experimental results is available at http://abulaish.com/data/ISIKDD12ChallengeResults.zip.

[3] Can be downloaded from ftp://128.196.239.164/

[4] http://www.ischool.drexel.edu/ISI-KDD2012/challenge.html



[0,1[D	[1-2[D	[2-3[D	[3-4[D	[4-5[D	[5-10[D	[10-20[D	[20-30[D	[1-2[M	[2-3[M	[3-4[M	[4-5[M	[5-6[M	[6-12[M	[1-2[Y	[2-3[Y	[3-4[Y	[4-5[Y	[5-10[Y
20482	1547	843	521	360	1334	792	300	448	188	149	90	76	358	331	102	33	8	6

[0,5[D	23753
[0,10[D	25087
[0,1[M	26179
[0,6[M	27130
[0,1[Y	27488
[0,5[Y	27962
	27968

Fig. 2. Lifespan of threads in the experimental data set

B. Manual Analysis

A team of three members performed a thorough manual analysis of the data set by navigating through all the posts commented by 2803 forum users. This analysis is based on five different criteria (given below) that generally convince a layman to conclude about the radicalness of a person. A score assigning methodology is followed for each criteria based on a user's behavior and the nature of participated discussions. For each criterion, a binary score (0 or 1) is assigned to each user by the team members, where the conflicts between the members are resolved using a voting scheme.

Explicit declaration (C_1): The first step towards radical user identification is to look for claims and declarations made by users in support of radical acts. We found users who claim to be a part of radical organizations and explicitly claim their support for radical ideas. For example, a user named *abu-abdallah-al-bulghari* stated: *It is incorrect to criticize any martyrdom operation*. Our review of the forum shows that radicals use the term *martyrdom operation* for suicide bombings, and in the above statement the user is clearly supporting the radical idea of suicide bombing. If any such post by a user is found in the data set, the user is assigned a score of 1 for this criterion, otherwise 0.

Explicit reply (C_2): The second step is to identify users claiming radicalness in the next level of the forum's hierarchy, i.e., in the form of replies to posts. The original post may or may not support a radical idea, but users show their agreement or disagreement clearly in replies. We found several discussions on the topic of suicide bombing. For example, a user named *suhaib-jobst* replied: *I was talking about his article about martyrdom operations. He declared it permissible. . . . . I (as a layman) believe that he is correct*. This reply clearly supports a radical thought. Users commenting such kind of posts are assigned a score of 1 for this criterion, otherwise 0.

Hint in declaration (C_3): In case of ambiguous posts in which there is no clear declaration, the user's radicalness can be identified to some extent by analyzing the nature of the posts. A user may not declare its association with a radical group or may not clearly support a radical idea, but the user's sentiment towards a topic of discussion and the choice of words provide hints on radicalness. For example a user named *abukhalid* states, *So when they say things such as 'these are suicide' it is much better if we can refute them with evidence from Al Albani or Uthaymeen*. Users with high radicalness support the idea of suicide bombing by using the term *martyrdom operation*, which reduces the negative impact of the repulsive word *suicide* and convince innocents in a better way. This kind of users are assigned a score of 1, otherwise 0.

TABLE III

A RANDOM SAMPLE OF *dead members*

talha-bin-ahmad	humble-slave-of-allah	abu-ibrahim2	strangetraveler
ibrahim-al-qubrusee	fatia	salinas	arabiclanguageacademy
iftihar	bb_aisha	al-hajeji	qad_aflahal_mominun
solaiman	adilmalik	umm-fulaan	alomgir
taahirah	sabbar	alislaam	aboo-abdillah

Hint in reply (C_4): Similar to the second criterion, this one is also related to the replies of users to an existing thread. The users' sentiment toward a radical post provides hints about being supportive to a radical ideology. For example, a user named *hussain* states: *Let's see: 'deviant methods such as suicide bombing...' Yep, sounds like Amrika lackey speak to me*. The user first quotes another person and then states his own sentiment towards the quotation. Similar to other users quoted above, this user has a supportive tone towards the radical idea of suicide bombing. Such users are assigned a score of 1 for this criterion, otherwise 0.

Sharing supporting information (C_5): In order to increase the number of supporters, radical users share faked and fabricated information with innocent users. We found several users sharing documents and videos containing fabricated emotive contents to persuade and influence others. For example, a user named *aboo-ayat-al-hindee* shared archives of a radically influential person. Thus, users exhibiting such property are scored as 1 for this criterion, otherwise 0.

C. Experimental Results

In order to establish the efficacy of the proposed method, we have considered three standard metrics that compare the closeness of two different rankings – MRR (Mean Reciprocal Rank) [33], Kendall's tau measure [34], and Spearman's footrule measure [34].

We start with applying some level of preprocessing for smoothing and proper organization of the data set. The radicalness measure $\rho(u_i)$ is computed for each user u_i. The user *Daniel* came out to be the most radical user in the entire forum. According to our manual analysis, this user has commented very lengthy posts which are nothing but the news articles related to terrorism and radical activities copied from some authentic sources. He has commented a total of 2770 posts, which made him to rank third in terms of post frequency, after *Umm Ahmed* with 2800 posts and *Abuz Zubair* with 2792 posts. Table IV shows the top-10 users in the forum in terms of post frequency and radicalness along with other ranking measures.

Through manual analysis, we found that a majority of users do not involve much in the discussions and remain as silent spectators. There exist a class of users who have started a thread and never got any response from others, due to which they could not establish any interaction relationship with others. Also, there are users who never participated in any kind of radical discussions. We define this kind of completely non-radical and non-influential users as *dead members* in the context of a dark Web forum, and filter them out to reduce the problem size. To identify them, a matrix $\Psi_{n\times n}$ is generated where n is the number of users in the forum and the corresponding matrix values are calculated using equation 16. $\mathrm{R}(u_i, u_j)$ and $\mathrm{I}(u_j|u_i)$, defined earlier, use $\mu(u_i, u_j) \leftarrow \mu_1(u_i, u_j)$. For any row i in Ψ, if there is no non-zero value in the entire row, then the corresponding user u_i is marked as a *dead member*. We thus found 896 *dead* members out of the total of 2803 users. A random sample of dead members is shown in Table III.

$$\Psi(i, j) = \log_2\left(\mathrm{R}(u_i, u_j) \times \mathrm{I}(u_j|u_i) + 1\right) \tag{16}$$

The proposed ranking algorithm is applied individually for each association metric on the remaining 1907 users. Table IV shows the 10 top-ranked users based on post frequency, radicalness (ρ measure), and the proposed method with each association metric, μ_i. All of them resulted into the same ranking for top three radically influential users, *Daniel*, followed by *AbuUsama* and *Mustafa al-Muhaajir*. The fourth place is occupied by one of *ahaneefah*, *Rakan* and *tayfah_mansurah* in the different association metrics based rankings. As we move on to lower ranks, the difference goes on increasing.

Unlike the radicalness property, it is sometimes hard for a human to say that one user is more influential than the other, or one user is influential and the other is not. Though our manual analysis was intended to establish a gold standard that could be used to compare with the automatically generated rankings, due to high complexity in the perception of *influence* and limitations of the human brain, we focused more on the radicalness of users. We are able to find a total of 70 radical users with varying intensities based on different criterion. The binary values for the five criterion are aggregated using a weighting scheme as $\mathrm{aggregate}(u_i) = \sum_{C_j \in \mathrm{Criterion}} \mathrm{weight}(C_j) \times C_j(u_i)$, where the weights considered for C_1 to C_5 are 0.5, 0.25, 0.10, 0.05, and 0.10, respectively. These weights are decided upon mutual agreement of the manual analysis team considering the prominence of different criterions in signifying their radicalness. Table V shows the 10 most radical users thus found.

TABLE V
10 MOST RADICAL USERS BASED ON MANUAL ANALYSIS

User	C_1	C_2	C_3	C_4	C_5	Aggregate
abu-abdallah-al-bulghari	1	0	1	0	0	0.60
suhaib-jobst	0	1	0	1	1	0.40
abumuwahid	0	1	1	1	0	0.40
shaheed666	0	1	0	0	1	0.35
leo	0	1	0	0	0	0.25
hussain	0	0	1	1	1	0.25
abu-ayoub-al-ansari	0	0	1	1	1	0.25
mustafa al-muhaajir	0	0	1	0	1	0.20
tayfah_mansurah	0	0	1	0	1	0.20
rakan	0	0	1	0	1	0.20

Considering this set of 70 radical users as gold standard, MRR values are computed for rankings obtained by applying the proposed method with different association metrics, as shown in Table VI. It includes two additional rankings; PF and ρ indicate the sorting based on frequency of posts and radicalness of corresponding users, respectively. We observe that, for top-10 radical users, the best performance is shown by μ_2 (CF-ITF) with MRR value as 12.126%, and at all other levels from top-20 to top-70, μ_{11} (Contingency Coefficient) performs the best with MRR values as 10.193%, 07.730%, 06.087%, 05.327%, 04.695%, and 04.091%, respectively. Thus it can be said that most of the times the proposed method gives the best results with contingency coefficient. The existing methods for identifying influential users in Dark Web forums have not been able to successfully capture the user radicalness. UserRank [11] is one such recent algorithm. To compare UserRank with our method, we applied it on our data set. The second last row in Table VI shows the MRR values obtained by UserRank. It can be observed from this table that for all levels from top-10 to top-70, all proposed association metrics outperform this existing state-of-the-art method.

TABLE VI
COMPARISON WITH THE GOLD STANDARD USING MRR

	Top 10	Top 20	Top 30	Top 40	Top 50	Top 60	Top 70
Proposed							
PF	0.04271	0.03986	0.03324	0.02704	0.02826	0.02509	0.02211
ρ	0.10656	0.09983	0.07413	0.05796	0.05059	0.04420	0.03851
μ_1	0.12102	0.10072	0.07590	0.05938	0.05181	0.04542	0.03958
μ_2	**0.12126**	0.10083	0.07569	0.05929	0.05141	0.04515	0.03936
μ_3	0.10464	0.09882	0.07557	0.05921	0.05128	0.04506	0.03929
μ_4	0.11902	0.09972	0.07488	0.05851	0.05090	0.04459	0.03886
μ_5	0.10693	0.09809	0.07513	0.05865	0.05102	0.04457	0.03885
μ_6	0.11895	0.09998	0.07531	0.05893	0.05120	0.04473	0.03899
μ_7	0.11895	0.09999	0.07531	0.05893	0.05120	0.04473	0.03899
μ_8	0.11525	0.10069	0.07628	0.06010	0.05206	0.04602	0.04015
μ_9	0.10694	0.10017	0.07610	0.05955	0.05194	0.04557	0.03971
μ_{10}	0.11298	0.10124	0.07684	0.06052	0.05299	0.04672	0.04071
μ_{11}	0.11437	**0.10193**	**0.07730**	**0.06087**	**0.05327**	**0.04695**	**0.04091**
UserRank [11]	0.04365	0.03496	0.02872	0.02375	0.02633	0.02368	0.02105
UserRank+Rad	0.08458	0.07920	0.06018	0.04775	0.04245	0.03773	0.03303

While it is clear that the proposed method outperforms UserRank, one question arises for the relatively poor performance of UserRank. Is it only because there is no radicalness measure in this method? Would UserRank perform similar to our method, if it is integrated with our radicalness measure? To study this, we generated results by replacing $\mathrm{I}(u_j|u_i)$ in Equation 13 with $\mathrm{P}(v_j|v_i)$ defined in [11] for UserRank. Table VI shows the MRR values for the ranking obtained using this approach under the name `UserRank+Rad`. On comparing the last two rows of this table, it can be seen that incorporating our radicalness measure improves the results of UserRank up to some extent, but still lower than the proposed method. Thus, in a broader perspective, it can be said that the collocation-based metrics (used in the proposed method) can deal with such ranking problem more effectively than the textual and temporal similarity based metrics (used in UserRank). Another interesting observation is that even the ranking directly based on the radicalness measure (row 2) outperforms `UserRank+Rad` in our results. However, this actually may not be true. One reason for such biased behavior towards radicalness measure may be due to focusing on users'

TABLE IV
10 TOP-RANKED MEMBERS ACCORDING TO DIFFERENT RANKING STRATEGIES

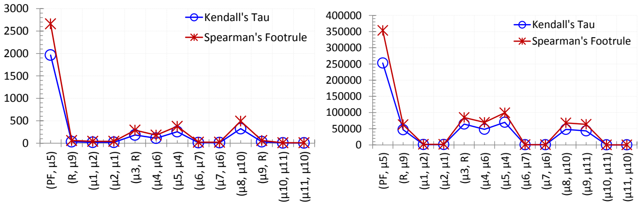
Post Frequency	**Radicalness** (ρ)	**Proposed**$_{\mu_1}$	**Proposed** $_{\mu_2}$	**Proposed**$_{\mu_3}$	**Proposed**$_{\mu_4}$	**Proposed**$_{\mu_5}$	**Proposed**$_{\mu_6}$	**Proposed**$_{\mu_7}$	**Proposed**$_{\mu_8}$	**Proposed**$_{\mu_9}$	**Proposed**$_{\mu_{10}}$	**Proposed**$_{\mu_{11}}$
umm-ahmed	daniel	daniel	daniel	daniel	daniel	daniel	daniel	daniel	daniel	daniel	daniel	daniel
abuz-zubair	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama	abuusama
daniel	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir	Mustafa al-Muhaajir
abuhannah	ahaneefah	tayfah_mansurah	tayfah_mansurah	ahaneefah	tayfah_mansurah	ahaneefah	tayfah_mansurah	tayfah_mansurah	rakan	ahaneefah	rakan	rakan
abuusama	rakan	rakan	rakan	rakan	rakan	rakan	rakan	rakan	tayfah_mansurah	rakan	ahaneefah	ahaneefah
isma-eel	tayfah_mansurah	abumuwahid	abumuwahid	tayfah_mansurah	abumuwahid	abumuwahid	abumuwahid	abumuwahid	ahaneefah	tayfah_mansurah	tayfah_mansurah	tayfah_mansurah
abumuwahid	abumuwahid	ahaneefah	ahaneefah	abumuwahid	ahaneefah	tayfah_mansurah	ahaneefah	ahaneefah	cageprisoners-com	abumuwahid	hajjaj	hajjaj
abu-abdallah-al-bulghari	hajjaj	abuz-zubair	abuz-zubair	abuz-zubair	umm-ahmed	umm-ahmed	umm-ahmed	umm-ahmed	hajjaj	hajjaj	cageprisoners-com	abumuwahid
abu-treika	abuz-zubair	gag-order	gag-order	hajjaj	abuz-zubair	abuz-zubair	abuz-zubair	abuz-zubair	abu_salmah	abuz-zubair	abumuwahid	cageprisoners-com
waziri	cageprisoners-com	umm-ahmed	umm-ahmed	umm-ahmed	abuhannah	abuhannah	abuhannah	abuhannah	abumuwahid	cageprisoners-com	abu_salmah	abu_salmah

radicalness more than their influence while preparing gold standard data set. As a result, the ranking produced by the radicalness measure resembles the gold standard more than that by `UserRank+Rad` (radicalness and influence).

TABLE VII
PAIR-WISE DISTANCE MEASURES FOR $k = 100$

Kendall's tau measure (K^p) **/ Spearman's footrule measure** (F^{k+1})

	PF	ρ	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	$\cdots$	2674/3580	2294/3058	2286/3058	2670/3510	2038/2726	**1964/2660**	2083/2780	2091/2792	3387/4506	2680/3588	3144/4186	3140/4178
ρ	2674/3580	$\cdots$	769/1100	767/1098	183/290	939/1338	1134/1586	881/1250	872/1242	959/1366	**31/60**	661/968	657/960
μ_1	2294/3058	769/1100	$\cdots$	**20/42**	804/1126	316/468	503/730	296/438	301/440	1325/1882	772/1102	1106/1594	1103/1590
μ_2	2286/3058	767/1098	**20/42**	$\cdots$	804/1130	323/482	511/738	295/430	300/434	1321/1876	770/1110	1104/1590	1101/1586
μ_3	2670/3510	**183/290**	804/1126	804/1130	$\cdots$	923/1320	1120/1568	876/1240	858/1230	1114/1568	214/338	832/1198	828/1190
μ_4	2038/2726	939/1338	316/468	323/482	923/1320	$\cdots$	256/376	**110/180**	115/188	1584/2208	950/1352	1362/1938	1359/1934
μ_5	1964/2660	1134/1586	503/730	511/738	1120/1568	**256/376**	$\cdots$	369/520	372/530	1776/2412	1142/1602	1556/2138	1553/2134
μ_6	2083/2780	881/1250	296/438	295/430	876/1240	110/180	369/520	$\cdots$	**11/22**	1526/2122	892/1262	1308/1860	1305/1856
μ_7	2091/2792	872/1242	301/440	300/434	858/1230	115/188	372/530	**11/22**	$\cdots$	1519/2108	883/1254	1305/1850	1302/1846
μ_8	3387/4506	959/1366	1325/1882	1321/1876	1114/1568	1584/2208	1776/2412	1526/2122	1519/2108	$\cdots$	953/1330	**324/490**	329/500
μ_9	2680/3588	**31/60**	772/1102	770/1110	214/338	950/1352	1142/1602	892/1262	883/1254	953/1330	$\cdots$	635/930	631/922
μ_{10}	3144/4186	661/968	1106/1594	1104/1590	832/1198	1362/1938	1556/2138	1308/1860	1305/1850	324/490	635/930	$\cdots$	**5/10**
μ_{11}	3140/4178	657/960	1103/1590	1101/1586	828/1190	1359/1934	1553/2134	1305/1856	1302/1846	329/500	631/922	**5/10**	$\cdots$



(a) $k = 100$ (b) $k = 1907$

Fig. 3. Pair-wise closest rankings

We also analyze the closeness of rankings generated by the different association metrics in the proposed method. We use Kendall's tau measure and Spearman's footrule measure to find the distance between them. Table VII shows the distance measures for each pair of association metrics used in the proposed approach when k is set to 100 (top 100 users). The values for each ranking in the left-hand side is intersected by the ranking on the top to form the pair. Each row has the lowest value in bold face, which indicates the pair as the closest ranking. The first row having (PF, μ_5) value in bold shows that PF-based ranking is closest to μ_5 ranking. Among the others, ρ (radicalness) is closest to μ_9, μ_1 is closest to μ_2, μ_2 is closest to μ_1, μ_3 is closest to ρ, μ_4 is closest to μ_6, μ_5 is closest to μ_4, μ_6 is closest to μ_7, μ_7 is closest to μ_6, μ_8 is closest to μ_{10}, μ_9 is closest to μ_{11}, μ_{10} is closest to μ_{11}, and μ_{11} is closest to μ_{10}. Figure 3 shows the closest ranked pairs as line charts. It is very clear that the ranking generated by sorting users upon their post frequency is the most dissimilar of all. μ_{10} ranking is very close to μ_{11} ranking. The other pairs close to each other are (μ_6, μ_7), and (μ_1, μ_2) rankings.

VI. Conclusion and Future Work

In this paper, we have proposed an approach to identify a ranked list of radically influential users in Web forums. We have formulated a radicalness measure and a variety of collocation-based association measures, and designed an algorithm based on PageRank to rank the radically influential users. Among the proposed association measures, the contingency coefficient measure is found as the most promising measure, when embedded in the customized PageRank algorithm along with the radicalness measure. The experimental results on a standard data set are promising that outperforms the existing UserRank algorithm. It is also found that the collocation-based association measures deal with such ranking problem more effectively than textual and temporal similarity based measures.

This work opens several promising directions for future research. Considering social relations in addition to the threaded interactions, exploring semantic factors like discussion context and topic drift for radicalness identification, and applying sentiment analysis to differentiate between the users taking positive and negative sides of radicalness, are few important research problems. Analyzing the affect of radical influence on the forum community is also a promising research direction to study the radicalness propagation in different extremist and hate groups.

Acknowledgment

We thank Faraz Ahmad from Michigan State University for assisting us to generate the benchmark data set during his stay at CoEIA, King Saud University.

References

[1] J. Qin, Y. Zhou, and H. Chen, "A multi-region empirical study on the internet presence of global extremist organizations," *Inf. Sys. Front.*, vol. 13, no. 1, pp. 75–88, 2011.

[2] T. Anwar and M. Abulaish, "Modeling a web forum ecosystem into an enriched social graph," in *Ubiquitous Social Media Analysis*, 2013, pp. 152–172.

[3] ——, "Identifying cliques in dark web forums- an agglomerative clustering approach," in *Proc. of the IEEE ISI*, 2012, pp. 171–173.

[4] 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, 2008.

[5] J.-h. Wang, T. Fu, H.-m. Lin, and H. Chen, "A framework for exploring gray web forums: Analysis of forum-based communities in taiwan," in *Proc. of the IEEE ISI*, 2006, pp. 498–503.

[6] J. Qin, Y. Zhou, E. Reid, G. Lai, and H. Chen, "Analyzing terror campaigns on the internet: Technical sophistication, content richness, and web interactivity," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 1, pp. 71–84, 2007.

[7] J. Glaser, J. Dixit, and D. P. Green, "Studying hate crime with the internet: What makes racists advocate racial violence?" *Journal of Social Issues*, vol. 58, no. 1, pp. 177–193, 2002.

[8] M. Trusov, A. V. Bodapati, and R. E. Bucklin, "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, vol. 47, no. 4, pp. 643–658, Aug. 2010.

[9] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. of the WSDM*, 2008, pp. 207–218.

[10] D. J. Watts, "Challenging the influential hypothesis," *WOMMA Measuring Word of Mouth*, vol. 3, pp. 201–211, 2007.

[11] C. C. Yang, X. Tang, and B. M. Thuraisingham, "An analysis of user influence ranking algorithms on dark web forums," in *Proc. of the ISI-KDD*, 2010, pp. 10:1–10:7.

[12] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the SIGKDD*, 2002, pp. 61–70.

[13] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *Proc. of the AAAI*, 2007, pp. 1371–1376.

[14] I. Esslimani, A. Brun, and A. Boyer, "Detecting leaders in behavioral networks," in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 281–285.

[15] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proc. of the WWW*, 2007, pp. 221–230.

[16] R. Ghosh and K. Lerman, "Predicting influential users in online social networks," *CoRR*, vol. abs/1005.4882, 2010.

[17] D. Kempe, J. Kleinberg, and E. Tardos, "Influential nodes in a diffusion model for social networks," in *Proc. of the ICALP*, 2005, pp. 1127–1138.

[18] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Sciences*, vol. 21, no. 2, pp. 256–276, 2006.

[19] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," in *Proc. of the WWW workshop*, 2006.

[20] X. Tang and C. C. Yang, "Identifing influential users in an online healthcare social network," in *Proc. of the IEEE ISI*, 2010, pp. 43–48.

[21] J. Kaplan and L. Weinberg, *The Emergence of a Euro-American Radical Right*. New Brunswick, N.J.: Rutgers University Press, 1998.

[22] J. Qin, Y. Zhou, G. Lai, E. Reid, M. Sageman, and H. Chen, "The dark web portal project: collecting and analyzing the presence of terrorist groups on the web," in *Proc. of the IEEE ISI*, 2005, pp. 623–624.

[23] S. Sizov, J. Graupmann, and M. Theobald, "From focused crawling to expert information: an application framework for web exploration and portal generation," in *Proc. of the VLDB*, 2003.

[24] T. Fu, A. Abbasi, and H. Chen, "A focused crawler for dark web forums," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 6, pp. 1213–1231, Jun. 2010.

[25] 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, 2008.

[26] D. B. Skillicorn, "Applying interestingness measures to ansar forum texts," in *Proc. of the ISI-KDD*, 2010, pp. 7:1–7:9.

[27] S. Kramer, "Anomaly detection in extremist web forums using a dynamical systems approach," in *Proc. of the ISI-KDD*, 2010, pp. 8:1–8:10.

[28] T. Fu, A. Abbasi, and H. Chen, "A hybrid approach to web forum interactional coherence analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 8, pp. 1195–1209, Jun 2008.

[29] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in *Proc. of the ISI-KDD*, 2010, pp. 9:1–9:9.

[30] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. of ICWSM*, 2010.

[31] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[32] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, Apr. 1998.

[33] E. M. Voorhees, "The trec-8 question answering track report," in *In Proc. of the TREC-8*, 1999, pp. 77–82.

[34] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *Proc. of the SODA*, 2003, pp. 28–36.

Appendix
Supporting Material

The supporting material and additional results have been made available as a supplemental document to this article and uploaded in the system.

Tarique Anwar received his Masters degree in Computer Science and Applications from Department of Computer Science, Jamia Millia Islamia, India, in 2010. After finishing his Masters, he worked as a Researcher at Center of Excellence in Information Assurance (CoEIA), King Saud University, Saudi Arabia. He is currently a PhD candidate at Swinburne University of Technology, Australia. His current research interests span over the areas of spatial databases, data mining, road networks, and information retrieval.

Muhammad Abulaish received the Masters degree in Computer Science and Applications from the Motilal Nehru National Institute of Technology, India, and PhD degree from the Indian Institute of Technology Delhi in 1998 and 2007, respectively. He is currently an Associate Professor and Head of the Computer Science department at the Jamia Millia Islamia (A Central University), Delhi. His research interests span over the areas of data mining, web intelligence, and security informatics. He is a senior member of the IEEE, ACM, and CSI. He has published over 71 research papers in reputed conference proceedings and journals related to his area of interests.

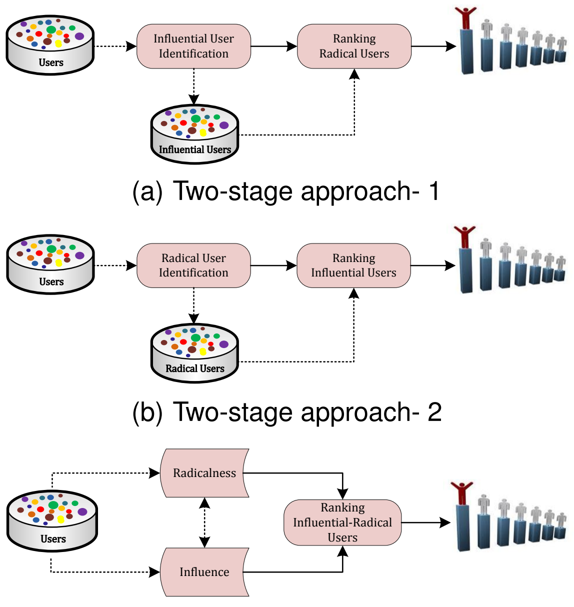
APPENDIX: SUPPORTING MATERIAL

I. Related Work

Table I presents a list of previous studies on the problem of influential user identification and the proposed core techniques.

II. Approaches for Ranking Radically Influential Users

As discussed in the paper, there can be two possible approaches to tackle the problem of radically influential user identification– *one-stage parallel* approach and *two-stage sequential* approach. Figures 1(a) and 1(b) shows their working mechanisms respectively.


Users
Influential User Identification
Ranking Radical Users
Influential Users
Users
Radical User Identification
Ranking Influential Users
Radical Users
Radicalness
Users
Ranking Influential-Radical Users
Influence

(a) Two-stage approach- 1

(b) Two-stage approach- 2

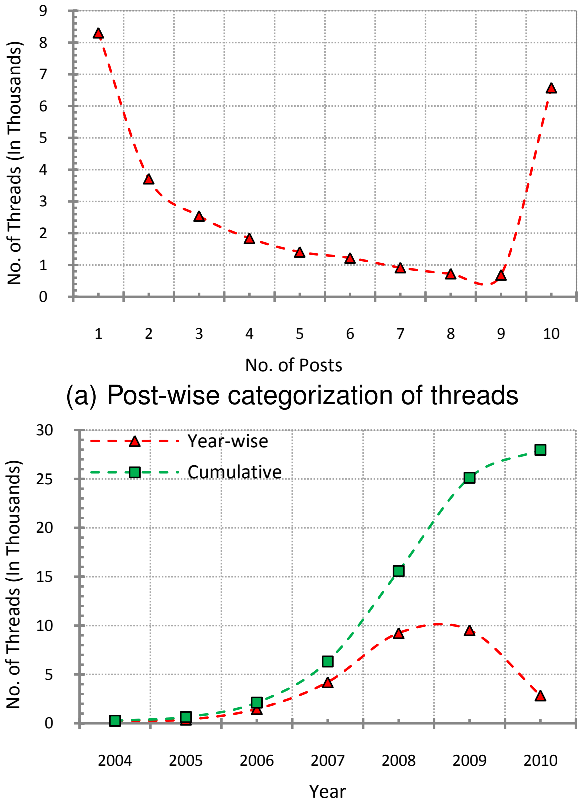
(c) One-stage approach

Fig. 1. Approaches for ranking radically influential users

III. Additional Dataset Statistics

Generally the *degree of effectiveness and intensiveness* of a Web forum can be estimated by some factual information derived from its discussions, like response to each thread, growth rate of the forum, and the length of time during which a thread remains alive. Looking further into the statistical composition, we found that a thread in the dataset has got response from a maximum of 10 posts, while a large section of threads (8311) ended up with just the first post of initiation.

Figure 2(a) shows the decreasing trend of number of threads as the number of posts they comprise, increases, with an exception at 10 posts. Going through a contemporary analysis, we found that it has remained most active in the years of 2008 and 2009, as the highest number of threads have been initiated in 2009 and next to it is 2008, with 9,540 and 9,238 threads initiated, respectively in these years. Figure 2(b) shows the yearly initiation and accumulation of threads in the forum.



(a) Post-wise categorization of threads

(b) Year-wise categorization of threads

Fig. 2. Dataset statistics

IV. Evaluation Metrics

The three evaluation metrics used in the paper are MRR (Mean Reciprocal Rank) [15], Kendall's tau measure [16], and Spearman's footrule measure [16]. This section presents the formulations of these measures.

The MRR measure focuses mainly on the rank of individual items in the gold standard ranking and compares it with the corresponding rank in the automatically generated ranking. It is computed using equation 1, where G is the set of gold standard set and $rank_i$ is the rank of i^{th} user of G in the ranked list of automatically generated ranking by the proposed approach. A higher value indicates a better accuracy.

$$MRR = \frac{1}{|G|} \times \sum_{i=1}^{|G|} \frac{1}{rank_i} \tag{1}$$

The remaining two metrics measure the distance between two different rankings generated by different approaches.

TABLE I
SELECTED PREVIOUS RESEARCH ON INFLUENTIAL USER IDENTIFICATION

Sl No.	Research	Platform	Testbed	Core Technique
1.	[1], [2]	Collaborative website	EachMovie database	Markov random fields
2.	[3]	Collaborative website	Epinions.com	PMI and RFM scores aggregated using ANN
3.	[4]	SNS	User activity logs in a major SN	Bayesian shrinkage approach implemented in a Poisson regression
4.	[5]	Collaborative website	Digg votes	Hypergeometric distribution, Normalized α-centrality measure
5.	[6]	Blogosphere	Intelliseek/ Blogpulse	PageRank
6.	[7]	Blogosphere	Digg, and The Unofficial Apple Weblog (TUAW)	InfluenceFlow
7.	[8]	Forum	Java Forum	Network structure, PageRank, HITTS, ExpertiseRank
8.	[9], [10]	Co-authorship network	arXiv database (High energy physics theory papers)	Discrete-optimization, Greedy approach, Decreasing cascade model
9.	[11]	Forum/Health care social network	MedHelp (Swine flu forum)	PageRank, UserRank
10.	[12]	E-mail network	Derived from a direct-mail marketing campaign	Statistical analysis
11.	[13]	Dark Web forum	AlJihad Network	PageRank, UserRank
12.	[14]	Blogosphere, Wiki	Japanese Wikipedia	Bond percolation

Kendall's tau measure considers just the relative ranking order of each pair of items in the two rankings, whereas Spearman's footrule measure provides an in-depth information by employing the absolute distance of each item in both rankings. Let τ_1 and τ_2 are two given top-k lists, $\tau_1(i)$ and $\tau_2(i)$ denote the rank of user i in τ_1 and τ_2, respectively, and D_{τ_1} and D_{τ_2} denote the domains of τ_1 and τ_2, respectively. Kendall's tau measure is computed using equation 2, where p is a penalty parameter constant with its value lying in between 0 and 1, and $P(\tau_1, \tau_2)$ is the set of all unordered item pairs in D_{τ_1} and D_{τ_2}. p is usually assumed as 0, unless there is additional supporting information to say about the ordering of i and j in the two top-k lists.

$$K^p(\tau_1, \tau_2) = \sum_{\forall (i,j) \in P(\tau_1, \tau_2)} \widehat{K}^p_{i,j}(\tau_1, \tau_2) \tag{2}$$

The value of $\widehat{K}^p_{i,j}(\tau_1, \tau_2)$ depends on the order of items i and j in τ_1 and τ_2, respectively. If they are in the same relative order in both the lists its value is 0, and if they are in opposite order its value is 1. Values assigned to it for every different situation is mentioned below:

$\widehat{K}^p_{i,j}(\tau_1, \tau_2) =$

1) **0**, if both i and j exist in both top k lists, and in same relative order;
2) **1**, if both i and j exist in both top k lists, but in opposite relative order;
3) **1**, if only i exists in one top k list, and only j exists in another top k list;
4) **1**, if both i and j exist in one top k list where i is ahead of j, and only j exists in another top k list;
5) **0**, if both i and j exist in one top k list where i is ahead of j, and only i exists in another top k list;
6) **1**, if both i and j exist in one top k list where j is ahead of i, and only i exists in another top k list;
7) **0**, if both i and j exist in one top k list where j is ahead of i, and only j exists in another top k list; and
8) **p**, if both i and j exist in one top k list, and neither i nor j exist in another top k list;

Spearman's footrule measure is computed using equation 3, where $\tau_1'(i) = \tau_1(i)$, if $i \in \tau_1$, and $\tau_1'(i) = k+1$, otherwise. Similarly, $\tau_2'(i) = \tau_2(i)$, if $i \in \tau_2$, and $\tau_2'(i) = k+1$, otherwise.

$$F^{k+1}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1} \cup D_{\tau_2}} |\tau_1'(i) - \tau_2'(i)| \tag{3}$$

V. ADDITIONAL EXPERIMENTAL RESULTS

In the paper, we used the Kendall's tau measure and Spearman's footrule measure to find the closeness of rankings generated by the different association metrics, and presented the results for 100 top ranking users. Here we present some additional results in Tables II, III, IV, V, and VI, showing the distance measures for k set to 200, 300, 400, 500, and 1907 (complete set) users, respectively. Figure 3 shows the closest ranked pairs as line charts for k set to 200, 300, 400, and 500, respectively.

REFERENCES

[1] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. of the SIGKDD*, 2001, pp. 57–66.

[2] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the SIGKDD*, 2002, pp. 61–70.

[3] Y.-M. Li, C.-H. Lin, and C.-Y. Lai, "Identifying influential reviewers for word-of-mouth marketing," *Electron. Commer. Rec. Appl.*, vol. 9, no. 4, pp. 294–304, Jul. 2010.

[4] M. Trusov, A. V. Bodapati, and R. E. Bucklin, "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, vol. 47, no. 4, pp. 643–658, Aug. 2010.

[5] R. Ghosh and K. Lerman, "Predicting influential users in online social networks," *CoRR*, vol. abs/1005.4882, 2010.

[6] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," in *Proc. of the WWW workshop*, 2006.

[7] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. of the WSDM*, 2008, pp. 207–218.

TABLE II
PAIR-WISE DISTANCE MEASURES FOR $k = 200$

	Kendall's tau measure (K^p) / Spearman's footrule measure (F^{k+1})												
	PF	R	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	$\cdots$	9101/12224	7751/10480	7751/10480	9038/12084	6827/9232	**6524/8868**	7048/9504	7056/9516	11075/14626	9128/12254	10275/13710	10269/13702
R	9101/12224	$\cdots$	2507/3482	2498/3476	519/790	3182/4432	4008/5516	2816/3916	2645/3746	**72/140**	1781/2598	1775/2590	
μ_1	7751/10480	2507/3482	$\cdots$	**57/106**	2624/3616	1110/1614	1927/2686	985/1456	990/1462	3952/5532	2505/3472	3316/4650	3312/4646
μ_2	7751/10480	2498/3476	**57/106**	$\cdots$	2621/3608	1137/1658	1953/2716	992/1454	999/1462	3926/5510	2496/3466	3291/4634	3287/4630
μ_3	9038/12084	**519/790**	2624/3616	2621/3608	$\cdots$	3188/4454	4020/5536	2828/3948	2823/3938	3085/4352	590/890	2281/3268	2275/3260
μ_4	6827/9232	3182/4432	1110/1614	1137/1658	3188/4454	$\cdots$	1025/1462	**458/698**	467/708	4879/6758	3201/4446	4207/5862	4203/5858
μ_5	6524/8868	4008/5516	1927/2686	1953/2716	4020/5536	**1025/1462**	$\cdots$	1469/2066	1473/2080	5740/7780	4027/5532	5029/6816	5025/6810
μ_6	7048/9504	2821/3924	985/1456	992/1454	2828/3948	458/698	1469/2066	$\cdots$	**21/40**	4554/6344	2842/3940	3888/5468	3884/5464
μ_7	7056/9516	2816/3916	990/1462	999/1462	2823/3938	467/708	1473/2080	**21/40**	$\cdots$	4549/6330	2837/3936	3888/5460	3884/5456
μ_8	11075/14626	2645/3746	3952/5532	3926/5510	3085/4352	4879/6758	5740/7780	4554/6344	4549/6330	$\cdots$	2584/3656	**965/1434**	971/1444
μ_9	9128/12254	**72/140**	2505/3472	2496/3466	590/890	3201/4446	4027/5532	2842/3940	2837/3936	2584/3656	$\cdots$	1716/2500	1710/2492
μ_{10}	10275/13710	1781/2598	3316/4650	3291/4634	2281/3268	4207/5862	5029/6816	3888/5468	3888/5460	965/1434	1716/2500	$\cdots$	**6/12**
μ_{11}	10269/13702	1775/2590	3312/4646	3287/4630	2275/3260	4203/5858	5025/6810	3884/5464	3884/5456	971/1444	1710/2492	**6/12**	$\cdots$

TABLE III
PAIR-WISE DISTANCE MEASURES FOR $k = 300$

	Kendall's tau measure (K^p) / Spearman's footrule measure (F^{k+1})												
	PF	R	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	$\cdots$	18215/24216	15253/20324	15263/20334	18182/24148	13200/17836	**12213/16444**	13954/18752	13963/18764	21005/27414	18300/24280	19888/26270	19883/26262
R	18215/24216	$\cdots$	5019/6798	5000/6782	1104/1650	6527/8884	8332/11366	5697/7790	5686/7776	4979/7160	**125/230**	3550/5194	3541/5184
μ_1	15253/20324	5019/6798	$\cdots$	**90/166**	5179/6998	2398/3394	4301/6002	1931/2774	1950/2800	7563/10406	5024/6814	6607/9122	6603/9118
μ_2	15263/20334	5000/6782	**90/166**	$\cdots$	5172/6984	2446/3456	4344/6050	1943/2780	1964/2810	7516/10364	5003/6796	6563/9088	6559/9084
μ_3	18182/24148	**1104/1650**	5179/6998	5172/6984	$\cdots$	6554/8934	8387/11418	5692/7834	5679/7818	5945/8514	1229/1792	4608/6656	4599/6646
μ_4	13200/17836	6527/8884	2398/3394	2446/3456	6554/8934	$\cdots$	2212/3182	**1136/1682**	1159/1708	9271/12614	6574/8934	8313/11364	8309/11360
μ_5	12213/16444	8332/11366	4301/6002	4344/6050	8387/11418	**2212/3182**	$\cdots$	3345/4734	3360/4760	10889/14466	8388/11420	9962/13302	9958/13296
μ_6	13954/18752	5697/7790	1931/2774	1943/2780	5692/7834	1136/1682	3345/4734	$\cdots$	**41/80**	8613/11870	5743/7840	7649/10616	7645/10612
μ_7	13963/18764	5686/7776	1950/2800	1964/2810	5679/7818	1159/1708	3360/4760	**41/80**	$\cdots$	8615/11866	5732/7830	7653/10606	7649/10602
μ_8	21005/27414	4979/7160	7563/10406	7516/10364	5945/8514	9271/12614	10889/14466	8613/11870	8615/11866	$\cdots$	4869/7016	**1709/2514**	1718/2526
μ_9	18300/24280	**125/230**	5024/6814	5003/6796	1229/1792	6574/8934	8388/11420	5743/7840	5732/7830	4869/7016	$\cdots$	3435/5046	3426/5036
μ_{10}	19888/26270	3550/5194	6607/9122	6563/9088	4608/6656	8313/11364	9962/13302	7649/10616	7653/10606	1709/2514	3435/5046	$\cdots$	**9/18**
μ_{11}	19883/26262	3541/5184	6603/9118	6559/9084	4599/6646	8309/11360	9958/13296	7645/10612	7649/10602	1718/2526	3426/5036	**9/18**	$\cdots$

TABLE IV
PAIR-WISE DISTANCE MEASURES FOR $k = 400$

	Kendall's tau measure (K^p) / Spearman's footrule measure (F^{k+1})												
	PF	R	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	···	29869/39994	24515/33238	24525/33248	29722/39796	22131/29850	**20430/27678**	22981/31004	22992/31024	34111/45540	29969/40098	32426/43428	32421/43420
R	29869/39994	···	88204/11736	8788/11724	1861/2734	10971/14920	14079/18976	9719/13270	9709/13260	8179/11684	**219/384**	5834/8328	5825/8318
μ_1	24515/33238	88204/11736	···	**122/228**	8926/11996	3701/5334	7031/9788	2844/4082	2868/4126	13025/17598	8871/11796	11526/15550	11523/15546
μ_2	24525/33248	8788/11724	**122/228**	···	8921/11986	3777/5424	7098/9866	2870/4104	2896/4152	12976/17548	8851/11782	11479/15512	11476/15508
μ_3	29722/39796	**1861/2734**	8926/11996	8921/11986	···	10983/14948	14137/19112	9637/13260	9623/13240	9882/13896	2082/2952	7688/10764	7679/10754
μ_4	22131/29850	10971/14920	3701/5334	3777/5424	10983/14948	···	3755/5280	**1828/2694**	1855/2726	15264/20896	11085/15036	13772/18872	13769/18868
μ_5	20430/27678	14079/18976	7031/9788	7098/9866	14137/19112	**3755/5280**	···	5586/7752	5605/7780	17845/23870	14197/19094	16460/22054	16456/22048
μ_6	22981/31004	9719/13270	2844/4082	2870/4104	9637/13260	1828/2694	5586/7752	···	**57/112**	14346/19814	9824/13388	12816/17728	12813/17724
μ_7	22992/31024	9709/13260	2868/4126	2896/4152	9623/13240	1855/2726	5605/7780	**57/112**	···	14353/19820	9816/13382	12826/17726	12823/17722
μ_8	34111/45540	8179/11684	13025/17598	12976/17548	9882/13896	15264/20896	17845/23870	14346/19814	14353/19820	···	8002/11458	**2858/4232**	2867/4244
μ_9	29969/40098	**219/384**	8871/11796	8851/11782	2082/2952	11085/15036	14197/19094	9824/13388	9816/13382	8002/11458	···	5642/8092	5633/8082
μ_{10}	32426/43428	5834/8328	11526/15550	11479/15512	7688/10764	13772/18872	16460/22054	12816/17728	12826/17726	2858/4232	5642/8092	···	**9/18**
μ_{11}	32421/43420	5825/8318	11523/15546	11476/15508	7679/10754	13769/18868	16456/22048	12813/17724	12823/17722	2867/4244	5633/8082	**9/18**	···

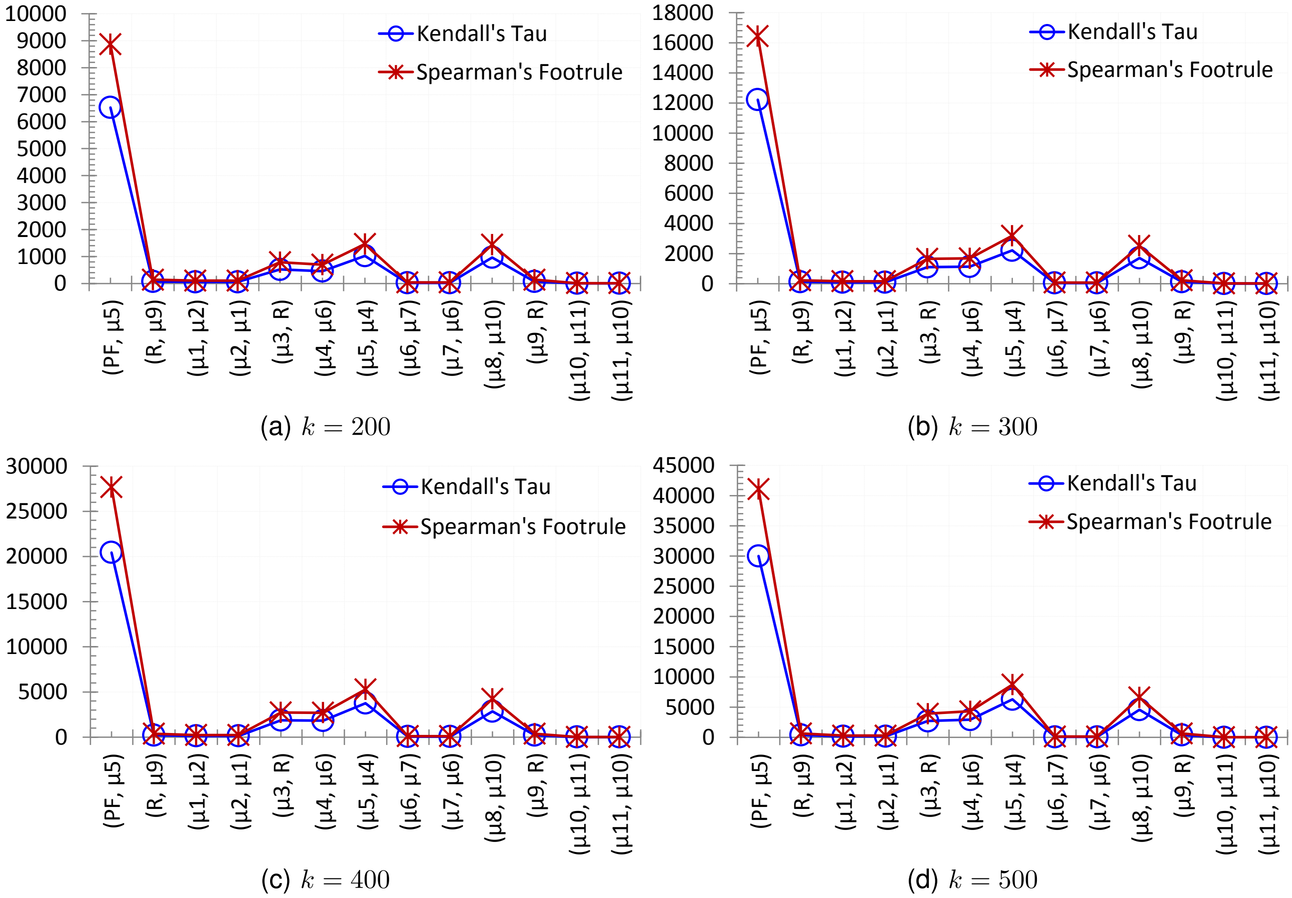
TABLE V
PAIR-WISE DISTANCE MEASURES FOR $k = 500$

	Kendall's tau measure (K^p) / Spearman's footrule measure (F^{k+1})												
	PF	R	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	···	43983/59496	36562/49314	36585/49338	44020/59296	33050/44648	**30009/41066**	34131/46236	34143/46264	49801/66890	44391/59708	47673/63880	47668/63872
R	43983/59496	···	13472/18088	13453/18074	2700/3918	16599/22528	21761/29194	14758/20144	14751/20134	12041/17140	**385/642**	8374/11998	8365/11988
μ_1	36562/49314	13472/18088	···	**160/298**	13747/18528	5573/7938	11244/15576	4274/6052	4296/6094	18964/25804	13622/18232	17179/23146	17176/23142
μ_2	36585/49338	13453/18074	**160/298**	···	13740/18514	5681/8050	11340/15678	4313/6098	4337/6144	18919/25754	13601/18216	17138/23106	17135/23102
μ_3	44020/59296	**2700/3918**	13747/18528	13740/18514	···	16718/22708	21949/29464	14705/20206	14693/20184	14444/20296	3063/4254	10993/15518	10984/15508
μ_4	33050/44648	16599/22528	5573/7938	5681/8050	16718/22708	···	6273/8738	**2914/4320**	2938/4352	22267/30488	16846/22764	20545/27874	20542/27870
μ_5	30009/41066	21761/29194	11244/15576	11340/15678	21949/29464	**6273/8738**	···	9115/12698	9137/12728	26631/35484	22031/29436	25211/33394	25208/33388
μ_6	34131/46236	14758/20144	4274/6052	4313/6098	14705/20206	2914/4320	9115/12698	···	**68/134**	21095/29050	15004/20378	19186/26252	19183/26248
μ_7	34143/46264	14751/20134	4296/6094	4337/6144	14693/20184	2938/4352	9137/12728	**68/134**	···	21106/29058	14999/20372	19199/26252	19196/26248
μ_8	49801/66890	12041/17140	18964/25804	18919/25754	14444/20296	22267/30488	26631/35484	21095/29050	21106/29058	···	11698/16764	**4562/6602**	4571/6614
μ_9	44391/59708	**385/642**	13622/18232	13601/18216	3063/4254	16846/22764	22031/29436	15004/20378	14999/20372	11698/16764	···	7988/11570	7979/11560
μ_{10}	47673/63880	8374/11998	17179/23146	17138/23106	10993/15518	20545/27874	25211/33394	19186/26252	19199/26252	4562/6602	7988/11570	···	**9/18**
μ_{11}	47668/63872	8365/11988	17176/23142	17135/23102	10984/15508	20542/27870	25208/33388	19183/26248	19196/26248	4571/6614	7979/11560	**9/18**	···

TABLE VI
PAIR-WISE DISTANCE MEASURES FOR THE COMPLETE LISTS ($k = 1907$)

	Kendall's tau measure (K^p) / Spearman's footrule measure (F^{k+1})						
	PF	ρ	μ_1	μ_2	μ_3	μ_4	μ_5
PF	···	334730/467120	285901/401208	285990/401236	353693/488988	267472/375386	**252829/353380**
ρ	334730/467120	···	177811/249974	177892/250036	64267/84568	191004/268172	220229/304996
μ_1	285901/401208	177811/249974	···	**1049/1794**	179960/254836	55323/77178	114810/160864
μ_2	285990/401236	177892/250036	**1049/1794**	···	179993/254854	55936/77974	115315/161510
μ_3	353693/488988	**64267/84568**	179960/254836	179993/254854	···	193969/272030	226676/314016
μ_4	267472/375386	191004/268172	55323/77178	55936/77974	193969/272030	···	69811/99240
μ_5	252829/353380	220229/304996	114810/160864	115315/161510	226676/314016	**69811/99240**	···
μ_6	275944/387302	184224/261714	47257/66004	47490/66300	184105/262538	48068/69634	113969/162488
μ_7	275976/387348	184298/261794	47429/66228	47666/66524	184131/262582	48262/69870	114161/162698
μ_8	355078/495084	107050/147386	189525/269998	189584/270028	159269/196070	207528/293548	233823/323106
μ_9	346233/479680	46839/62514	188548/263020	188609/263076	111106/114626	205542/286468	233612/321920
μ_{10}	350955/488582	88075/118984	203340/283562	203407/283580	151114/174642	220617/308280	244882/337964
μ_{11}	350948/488570	88072/118982	203345/283566	203412/283584	151113/174642	220620/308284	244885/337970

	μ_6	μ_7	μ_8	μ_9	μ_{10}	μ_{11}
PF	275944/387302	275976/387348	355078/495084	346233/479680	350955/488582	350948/488570
R	184224/261714	184298/261794	107050/147386	**46839/62514**	88075/118984	88072/118982
μ_1	47257/66004	47429/66228	189525/269998	188548/263020	203340/283562	203345/283566
μ_2	47490/66300	47666/66524	189584/270028	188609/263076	203407/283580	203412/283584
μ_3	184105/262538	184131/262582	159269/196070	111106/114626	151114/174642	151113/174642
μ_4	**48068/69634**	48262/69870	207528/293548	205542/286468	220617/308280	220620/308284
μ_5	113969/162488	114161/162698	233823/323106	233612/321920	244882/337964	244885/337970
μ_6	···	**376/724**	204202/291662	197649/279312	214811/302488	214816/302490
μ_7	**376/724**	···	204288/291784	197717/279418	214919/302628	214924/302630
μ_8	204202/291662	204288/291784	···	74901/106582	**47829/67866**	47846/67890
μ_9	197649/279312	197717/279418	74901/106582	···	43260/**63698**	**43257**/63700
μ_{10}	214811/302488	214919/302628	47829/67866	43260/63698	···	**17/34**
μ_{11}	214816/302490	214924/302630	47846/67890	43257/63700	**17/34**	···



(a) $k = 200$

(b) $k = 300$

(c) $k = 400$

(d) $k = 500$

Fig. 3. Pair-wise closest rankings

[8] J. Zhang, M. S. Ackerman, and L. Adamic, “Expertise networks in online communities: structure and algorithms,” in *Proc. of the WWW*, 2007, pp. 221–230.

[9] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. of the SIGKDD*, 2003, pp. 137–146.

[10] ——, “Influential nodes in a diffusion model for social networks,” in *Proc. of the ICALP*, 2005, pp. 1127–1138.

[11] X. Tang and C. C. Yang, “Identifing influential users in an online healthcare social network,” in *Proc. of the IEEE ISI*, 2010, pp. 43–48.

[12] S. Hill, F. Provost, and C. Volinsky, “Network-based marketing: Identifying likely adopters via consumer networks,” *Statistical Sciences*, vol. 21, no. 2, pp. 256–276, 2006.

[13] C. C. Yang, X. Tang, and B. M. Thuraisingham, “An analysis of user influence ranking algorithms on dark web forums,” in *Proc. of the ISI-KDD*, 2010, pp. 10:1–10:7.

[14] M. Kimura, K. Saito, and R. Nakano, “Extracting influential nodes for information diffusion on a social network,” in *Proc. of the AAAI*, 2007, pp. 1371–1376.

[15] E. M. Voorhees, “The trec-8 question answering track report,” in *In Proc. of the TREC-8*, 1999, pp. 77–82.

[16] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top k lists,” in *Proc. of the SODA*, 2003, pp. 28–36.