Topic-Based Social Network Analysis for Virtual Communities of Interests in the Dark Web

Gaston L'Huillier
University of Chile
Republica 701
Santiago, Chile
glhuilli@dcc.uchile.cl

Hector Alvarez
University of Chile
Republica 701
Santiago, Chile
halvarez@ing.uchile.cl

Sebastián A. Ríos
University of Chile
Republica 701
Santiago, Chile
srios@dii.uchile.cl

Felipe Aguilera
University of Chile
Blanco Encalada 2120
Santiago, Chile
faguiler@dcc.uchile.cl

ABSTRACT

The study of extremist groups and their interaction is a crucial task in order to maintain homeland security and peace. Tools such as social networks analysis and text mining have contributed to their understanding in order to develop counter-terrorism applications. This work addresses the *topic-based community key-members extraction problem*, for which our method combines both text mining and social network analysis techniques. This is achieved by first applying latent Dirichlet allocation to build two topic-based social networks in online forums: one social network oriented towards the thread creator point-of-view, and the other is oriented towards the repliers of the overall forum. Then, by using different network analysis measures, topic-based key members are evaluated using as benchmark a social network built a plain representation of the network of posts. Experiments were successfully performed using an English language based forum available in the Dark Web portal.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; K.4.1 [**Computing Milieux**]: Computers and Society—*Public Policy Issues*

General Terms

Dark Web, Social Network Analysis, Text Mining

Keywords

Terrorism Informatics, Terrorism knowledge portals, Latent Dirichlet Allocation, Virtual Communities of Interest

ISI-KDD 2010, Extended Paper

ISI-KDD 2010 , July 25, 2010, Washington, D.C., USA.

1. INTRODUCTION

"The process of radicalization continues in a hostile physical environment, but it is enabled by the Internet, resulting in a disconnected, decentralized social structure".
– Marc Sageman [20]

Today, it is no mystery that terrorists and cyber-criminals are looking for uncovered means to seek peers with the same interests such as internet forums, blogs, and social networks, where they can share and comment their feelings and interests with other that support their cause. In this sense, Dark Web[1] has brought endless potential to achieve the coordination, propaganda delivery, and other unwanted interactions between extremists groups. However, web information can be retrieved for further analysis. Many researchers have been wondering whether the leaders of the Dark Web can be identified by their interaction with other members.

The Dark Web is conformed by Virtual Communities of Interests (VCoI) [9, 16] are communities which members are reunited by their shared interests in a topic, like fan clubs of artists. Therefore, a key aspects when studying VCoI is to achieve a complete understanding on what are the main interests of the community. Only then, it will be possible to obtain a better insight of community's social aspects, leading to a better identification of the key members (i.e. Opinion Leaders), density reduction, among other benefits.

Different approaches have been previously developed for Social Networks, Virtual Communities [11], and Virtual Communities of Practices [4, 17]. To the best of our knowledge, none Latent Semantic Analysis (LSA) for enhancing the Social Network Analysis (SNA) has ever been developed in Dark Web portals before. However, we think the application of these techniques may enhance greatly SNA results. This work's main contribution is the hybrid approach using topic-models and SNA for enhancing the extraction of key-members in a VCoI.

This paper is structured as follows: In section 2, previous work on SNA and text mining for key members identification is presented, as well as previous work on the Dark Web forums analysis. Then, in section 3 the proposed methodology

[1]Internet-based forums or platforms for terrorists and cyber-criminals.

to solve the *topic-based community key members extraction problem* and main contribution of this work is described. In section 4 the experimental setup is detailed, and its results are discussed. Finally in section 5, the main conclusions of this paper and future work are presented.

2. PREVIOUS WORK

Since Dark Web is a VCoI, there are different goals associated to their members' objectives [7]. In this case, the support of the community with an online forum where its anonymity, ubiquitous, and free-of-speech makes the perfect environment to share fundamentalism and terrorism propaganda. Therefore, a method to recognize the underlying members' objectives is needed in order to identify threats or security issues. Topic-based SNA is a way to track, in some degree, topics on members' thread, which can be related to members' goals.

In the following, previous work on topic-based SNA is presented, as well as related work on Dark Web Social Network Analysis applications.

2.1 Topic-Based Social Network Analysis

SNA [22] helps to understand relationships in a given community analyzing its graph representation. Users are seen as nodes and relations among users are seen as arcs. This way, several techniques have been proposed to extract key members [12], classify users according his relevance within the community [14], discovering and describing resulting subcommunities [10], amongst other applications. However, all these approaches leave aside the meaning of relationships among users. Therefore, analysis based only on reply of mails or posts to measure relationships' strongness or weakness it is not a good indicator.

McCallum et al. in [11] described how to determine roles and topics in a text-based social networks by building Author-Recipient-Topic (ART) and Role-Author-Recipient-Topic (RART) models. Furthermore, in Pathak et al. [13], a community based topic-Model integrated social network analysis technique (Community-Author-Recipient-Topic model or CART) is proposed to extract communities from a emails corpus based on the topics covered by different members of the overall network. These approaches novelty is the use of data mining on text from the social network to perform SNA to study Roles or Sub-groups extraction.

2.2 Social Network Analysis on the Dark Web

As described by Xu & Chen in [25], the topology of dark networks share different properties with other types of networks, where small-world structures are determined by the information flow properties, characterized by short average path and a high clustering coefficient. In this sense, social network analysis tools used in other virtual communities applications [19], could be useful to analyze this kind of structures. They combine concepts obtained from communities' administrators into a concept-based text mining, which allow to obtain interesting results in terms of purpose accomplishment in a Virtual Community of Practice (VCoP).

In Reid et al. [18], Web forums where analyzed in order to determine whether a given community has been involved with terrorist presence by using automated and semi-automated procedures to gather information and analyze it. Other applications, such as the proposed by Zhou et al. in [29], aims towards the capture of domestic Web forums and the creation of a social network mapping to identify their structure and cluster affinities. Finally, in [28] a complete framework on how to build a portal of Dark Web forums is presented, where data collection and integration, as well as its visualization and open access, are the main contributions of the authors.

Previous work on key-members identification on the Dark Web [24], describes how several centrality measures can be used to identify different key-members of a social network. Here, the degree, betweenness, and closeness measures [6] have been described as tools to characterize key-members of a given social network. It is important to highlight that this approach has been traditionally used in previous work.

Latent semantic analysis has been previously applied in different Dark Web applications, such as [5] where Latent Semantic Indexing was used to link nodes to certain topics in the network construction. Furthermore, in [26], an latent Dirichlet allocation (LDA) based web crawling framework is proposed to discover different topics from Dark Web forum cites, but not further social network analysis applications with these findings. Thus, leaving without measuring many social aspects from the Dark Web social structure.

In terms of text mining in Dark Web forums [2], applications have been oriented to model assessment and feature selection in order to improve the classification of messages that contains sensitive information on extremists' opinions and sentiments. Also, authorship analysis [1] has been previously applied in order to analyze the groups' authorship tendencies, and more important, tackling the anonymity problem associated to this type of virtual communities.

3. PROPOSED METHOD

The main question of this work is how to enhance the key-members discovery based on a topic-based social network structure. The first step is to obtain a reduced/filtered representation of the inner social community. However, this representation must be created in such way that information contained is better to discover key-members (aligned to community's goals and members who produce interaction). The second step is to apply a core members' algorithm to obtain the key-members of a virtual community based on their respective topics of interest. As a result, we obtain the rank of the complete community members where for each community interesting topic, we are able to reveal members that can be considered as experts or key on that topics.

We must mention the difference between a key-member and a highly-radical member, since key members have several characteristics that define them. Firstly, a key member may be highly-radical member in a given topic or not. Secondly, he/she may increase the interaction in the community because he/she points out interesting messages, which produce replies from different level of members. We will focus on key-members of any kind, afterwards, depending on the topic it would be possible to classify him/her as a radical member or not. Of course, it is possible to automate the process of radical member discovery, however, more work is needed to avoid subjectivity. Furthermore, accusing or pointing such label on a person is not a simple matter and implications are far beyond this work.

We defined a key-member as a person totally aligned with the VCoIs' goals and topics. Thus, producing contents which are very relevant to satisfy other members' interests. The

only way to measure a key-member as defined, is using an hybrid approach of SNA combined with latent semantic-based text mining. This way, we can discover threatening topics, and perform specific analysis on each one.

3.1 Basic Notation

Let us introduce some concepts. In the following, let $\mathcal{V}$ a vector of words that defines the vocabulary to be used. We will refer to a word w, as a basic unit of discrete data, indexed by $\{1, ..., |\mathcal{V}|\}$. A post message is a sequence of S words defined by $\mathbf{w} = (w^1, ..., w^S)$, where w^s represents the s^{th} word in the message. Finally, a corpus is defined by a collection of $\mathcal{P}$ post messages denoted by $\mathcal{C} = (\mathbf{w}_1, ..., \mathbf{w}_{|\mathcal{P}|})$.

A vectorial representation of the posts corpus is given by TF-IDF$= (m_{ij}), i \in \{1, \ldots, |\mathcal{V}|\}$ and $j \in \{1, \ldots, |\mathcal{P}|\}$, where m_{ij} is the weight associated to whether a given word is more important than another one in a document. The m_{ij} weights considered in this research is defined as an improvement of the *tf-idf* term [21] (*term frequency times inverse document frequency*), defined by

$$m_{ij} = \frac{n_{ij}}{\sum_{k=1}^{|\mathcal{V}|} n_{kj}} \times log\left(\frac{|\mathcal{C}|}{n_i}\right) \quad (1)$$

where n_{ij} is the frequency of the i^{th} word in the j^{th} document and n_i is the number of documents containing word i. The *tf-idf* term is a weighted representation of the importance of a given word in a document that belongs to a collection of documents. The *term frequency* (TF) indicates the weight of each word in a document, while the *inverse document frequency* (IDF) states whether the word is frequent or uncommon in the document, setting a lower or higher weight respectively.

3.2 Topic Modeling

A topic model can be considered as a probabilistic model that relates documents and words through variables which represent the main topics inferred from the text itself. In this context, a document can be considered as a mixture of topics, represented by probability distributions which can generate the words in a document given these topics. The inferring process of the latent variables, or topics, is the key component of this model, whose main objective is to learn from text data the distribution of the underlying topics in a given corpus of text documents.

A main topic model is the Latent Dirichlet Allocation (LDA) [3]. LDA is a Bayesian model where latent topics of documents are inferred from estimated probability distributions over the training dataset. The key idea of LDA, is that every topic is modeled as a probability distribution over the set of words represented by the vocabulary ($w \in \mathcal{V}$), and every document as a probability distribution over a set of topics ($\mathcal{T}$). These distributions are sampled from multinomial Dirichlet distributions.

For LDA, given the smoothing parameters β and α, and a joint distribution of a topic mixture θ, the idea is to determine the probability distribution to generate from a set of topics $\mathcal{T}$, a message composed by a set of S words w ($\mathbf{w} = (w^1, ..., w^S)$),

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p(w^s|z_s, \beta) \quad (2)$$

where $p(z_s|\theta)$ can be represented by the random variable θ_i, such that topic z_s is presented in document i ($z_s^i = 1$). A final expression can be deduced by integrating equation 2 over the random variable θ and summing over topics $z \in \mathcal{T}$. Given this, the marginal distribution of a message can be defined as follows:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{s=1}^{S} \sum_{z_s \in \mathcal{T}} p(z_s|\theta) p(w^s|z_s, \beta) \right) d\theta \quad (3)$$

The final goal of LDA is to estimate previously described distributions to build a generative model for a given corpus of messages. There are several methods developed for making inference over these probability distributions such as variational expectation-maximization [3], a variational discrete approximation of equation 3 empirically used by [23], and by a Gibbs sampling Markov chain Monte Carlo model efficiently implemented and applied by [15].

3.3 Network Configuration

To build the social network, the members' interaction must be taken into consideration. In general, members' activity is followed according to its participation on the forum. Likewise, participation appears when a member post in the community. Because the activity of the VCoI is described according members' participation, the network will be configured according to the following: Nodes will be the VCoI members, and arcs will represent interaction between them. How to link the members and how to measure their interactions to complete the network is our main concern.

In this work, will be describe two VCoIs' network representation according the following replying schema of members:

1. **Creator-oriented Network**: When a member create a thread, every reply will be related to him/her. This network representation is the less dense network (density is measured in terms of the number of arcs that the network have).

2. **Last Reply-oriented Network**: Every reply of a thread will be a response of the last post. This network representation has a middle density.

In figure 1 the latter two approaches of network conversion of the forum is presented. In figure 1, arcs represents members' replies and nodes represent the users who made the posts. In our first approach, the weight of arcs will be a counter of how many times a given member replies to other one.

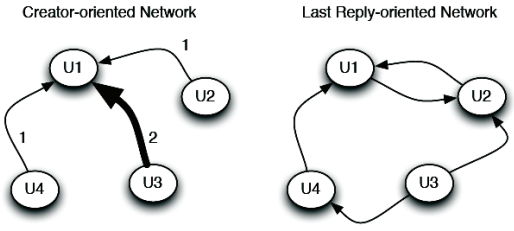

Figure 1: Two different network models to represent a given forum interaction.

In order to consider the reply of members according to the community purpose (for any of these configurations), and to filter noisy posts, a topic-based message reduction is performed. According to previous network configurations, the topic-based filtering method is used in order to remove all replies that are not according to the posts' topic. Here, by using LDA, a list of keywords for each topic is determined, and later, used to build the final version of the social network.

3.4 Topic-based Network Filtering

Previous work [19] brings a method to evaluate community goals accomplishment. In this work we will use this approach to classify the members' posts according VCoIs' goals. These goals are defined as a set of terms, which are composed by a set of keywords or statements in natural language.

The idea is to compare with euclidean distance two members' posts. If the distance is over a certain threshold θ, an interaction will be considered between them. We support the idea that this will help us to avoid irrelevant interactions. For example, in a VCoI with k goals (or topics), let TB_j a post of user j that is a reply to post of user i (TB_i). The distance between them will be calculated with equation 4.

$$d_m(TB_i, TB_j) = \frac{\sum_k g_{ik} g_{jk}}{\sqrt{\sum_k g_{ik}^2 \sum_k g_{jk}^2}} \quad (4)$$

Where g_{ik} is the score of topic k in post of user i. It is clear that the distance exists only if TB_j is a reply to TB_i. After that, the weight of arc $a_{i,j}$ is calculated according to equation 5.

$$a_{i,j} = \sum_{\substack{i,j \\ d_m(TB_i, TB_j) \geq \theta}} d(TB_i, TB_j) \quad (5)$$

We used this weight in both configurations previously described (Creator-oriented and Last Reply-oriented). Afterwards, we applied HITS [8] to find the key members on the different network configurations.

3.5 Network Construction

First, as a baseline proceedure, Algorithm 1 shows how to determine the Semantic Weights Matrix, that assigns an score to every post according to all topics considered for the network construction. In this case, the algorithm intially determines the `TF-IDF` matrix according to Equation 1, the a semantic matrix `SM` according to topic-based text mining described in Section 3.2 and Section 3.1 respectively. Finally, the matrix multiplication between `TF-IDF` and `SM` defines the `SWM` matrix.

Algorithm 2 presents the pseudo-code on the graph $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$ is build by using the Creator-oriented network. First, the SWM matrix is build according to Algorithm 1. Then, considering all of the posts $\mathcal{P}$, the network is built following the structure presented in Figure 1 (a). This is, for each post creator i, the arc weight $a_{i,j}$ is increased according to the number of repliers j, that their messages' distance greater or equal than the threshold θ.

Algorithm 3, as well as the latter, defines in the first place the `SWM` matrix. Then, according to in Figure 1 (b), for each post username i, the arc weight $a_{i,j}$ is increased according to its number of all repliers j whose messages' distance are greater or equal than a threshold θ.

Algorithm 1 Initialize Semantic Weights Matrix

Input: $\mathcal{V}$ (Vocabulary)
Input: $\mathcal{P}$ (Posts)
Input: k (Number of Topics)
Output: Semantic Weights Matrix `SWM`$[|\mathcal{P}|, k]$
1: `TF-IDF`$[|\mathcal{P}|, |\mathcal{V}|]$ (Eq. 1)
2: `SM`$[k\,\mathcal{V}] \leftarrow$ Build `SM` (semantic matrix) according to `Topics` (Sec. 3.2)
3: `SWM`$[|\mathcal{P}|, k] \leftarrow$ `TF-IDF` $\times$ `SM`$^{\mathrm{T}}$

Algorithm 2 Creator-oriented Network

Input: $\mathcal{P}$ (Posts)
Output: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$
1: Build `SWM` according to Algorithm 1
2: Initialize $\mathcal{N} = \{\}, \mathcal{A} = \{\}$
3: **for** each $i \in \mathcal{P}$ **do**
4: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup i$
5: **end for**
6: **for** each $i \in \mathcal{P}.creator$ **do**
7: $\quad$ **for** each $j \in \{i.replies\}, i \neq j$ **do**
8: $\quad\quad$ **if** $d_m(P_i, P_j) \geq \theta$ **then**
9: $\quad\quad\quad a_{i,j} \leftarrow a_{i,j} + 1$
10: $\quad\quad\quad \mathcal{A} \leftarrow \mathcal{A} \cup a_{i,j}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **end for**

Algorithm 3 Reply-oriented Network

Input: $\{\mathcal{V}, \mathcal{P}, k\}$
Output: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A})$
1: ... (as presented in algorithm 2)
2: **for** each $i \in \mathcal{P}$ **do**
3: $\quad$ **for** each $j \in \{i.reply\}, i \neq j$ **do**
4: $\quad\quad$ **if** $d_m(P_i, P_j) \geq \theta$ **then**
5: $\quad\quad\quad a_{i,j} \leftarrow a_{i,j} + 1$
6: $\quad\quad\quad \mathcal{A} \leftarrow \mathcal{A} \cup a_{i,j}$
7: $\quad\quad$ **end if**
8: $\quad$ **end for**
9: **end for**

4. EXPERIMENTAL SETUP AND RESULTS

The proposed methodology was applied in the `Ansar1` English language based forum, available on the Dark Web portal[2], for which examples of the topic extraction methodology, network construction, and key-members using the HITS algorithm were determined.

In the following, an analysis of topics extracted using LDA (described in Section 3.2) is presented. Then, the network topology construction by using both Reply-oriented and Creator-oriented structures for the whole period is described. Finally, key-members where determined using HITS (both authority and hub scores), whose results where compared against different topologies of the social network.

[2] `http://128.196.40.222:8080/CRI_Indexed_new/login.jsp` [last accessed 21-10-2010]

4.1 Dark Web evaluation results and discussions

In this section, results obtained for the Dark Web `Ansar1` forum are presented and discussed accordingly. Firstly, the definition of topics using graphical models are presented as well as brief analysis on main extracted topics. Secondly, different network representations are displayed as well as benchmark network evaluations. Finally, HITS algorithm results for authority and hub scores and discussion are presented.

4.1.1 Topic Extraction

There are 14 months (Dec. 2008 - Jan. 2010) of data available. Posts were created by 376 members and extracted topics where realized over 29.057 posts $\mathcal{P}$ and 103.791 words in the vocabulary $\mathcal{V}$ by using a `C++` Gibbs sampling-based implementation of LDA[3] previously described in Section 3.2.

Topics extracted where evaluated according to the compactness of each topic, by the average shortest distance (ASD) among the top 30 words in the topic [27]. The ASD between two keywords w_i and w_j is determined by,

$$\text{ASD}(w_i, w_j) = \frac{max\{log(N_i), log(N_j)\} - log(N_{i,j})}{log(|\mathcal{P}| - min\{log(N_i), log(N_j)\})} \quad (6)$$

where N_i is the number of posts that contains w_i, $N_{i,j}$ is the number of posts that contain both of them, and $|\mathcal{P}|$ is the total number of posts in the corpus. By using equation 6 [27], a representative number of topics was determined, by evaluating the ASD number for $k \in [10, 50]$, where 47 was the number of topics with lowest ASD. Then, topics were manually grouped into concepts categories, used for social network analysis.

In Table 1, the overall concept proposed is "Local Conflicts" as for its main topics are related to terrorist and counter-terrorism activities over different localities, such as Russia, Irak, Somalia, India, and Afghanistan.

Table 2 is "War on Terror", as its main topics are related to U.S. activities and relations towards terrorism activities, where the proposed topic names are "War", "American Soldier", "Imprisonment", "Obama", and "Military Operations".

Finally, as shown in Table 3, the overall concept proposed is "Recruiting" where topics such as "Religious Propaganda", "Religious Conflicts", "*Ansar*", "Family", and "Ideology" are included.

4.1.2 Topic-Based Social Network Visualization

In Figure 2 (a) the whole corpus social network is presented without any filtering methods. This can be interpreted as the complete network built by using the complete repliers structure, where the edge $a_{i,j}$ between user i and j is defined for every interaction between users. Then in Figure 2 (b) and (c) we present the same graphs but filtered using topic-based methods. In this case, it is possible to see a great density reduction. This suggests that visalization techniques, such as graphs, can provide better visual information to analysts. Other great benefit, is that, since we have lesser dense graph, SNA algorithms, such as HITS, PageRank, among others, run in shorter times.

Furthermore, in Figure 2 (a), it is possible to observe very clearly three centers of interaction (or participation). If we use HITS to analyze the information on this graph, the results are that key-members correspond to the three administrators of the community which can also be seen in the graph. Therefore, the huge amount of posts from these members makes all other interactions look rather small, making useless the traditional way for key-members detecton. We can correct this situation by using eather concept-based or Topic-based filtering.

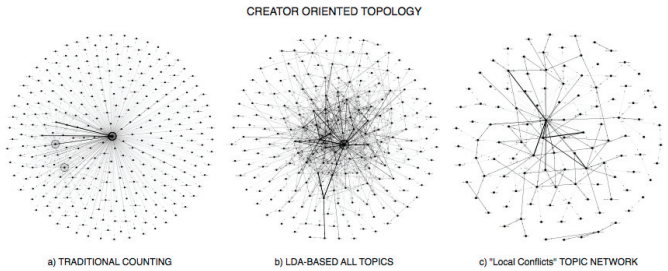

Figure 2: Social network visualization (a) without topic-based filtering and (b) topic-based filtering, and (c) topic-based filtering for "Local Conflict" (using creator network representation)

In Figure 2 (b) and (c) we can observe that previously stated three centers are gone, and it is much more diffuse to observe centers of interaction from the graphs. However, making an analysis to discover which are the key members now is faster and provide better results. In fact, when appliying HITS, it is possible to figure out other members, rather than administrators, as key-members of the community. The comparison may be seen in Table 4.

More specifically, for "Local Conflicts" topics, in Figure 3 and Figure 4, the visualization for both Reply-oriented and Creator-oriented social networks are presented respectively with their top 8 users using HITS hub score.

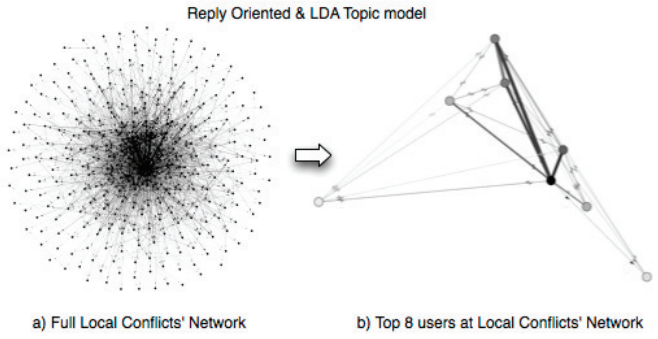

Figure 3: Visualization of the complete "local conflict" network and their top 8 users in the Reply-oriented network for the complete period of the forum activity, using the hub score from HITS.

In Figure 3 (a) all interactions presented for the "Local Conflicts" topic for Reply-oriented network configuration are depicted, where there is no relevant information that could be infered or deduced directly. However, by listing top key-

[3] `http://gibbslda.sourceforge.net/` [Last accessed 21-10-2010]

Table 1: Ten most relevant words with their respective conditional probabilities for five topics associated between them to the "Local Conflicts" concept from the Ansar1 forum.

Topic 1 "**Eastern Europe**"	Topic 8 "**Irak Conflicts**"	Topic 13 "**Somalia Conflicts**"	Topic 17 "**India Conflicts**"	Topic 19 "**Afghanistan Conflicts**"
russian (0.0289)	iraq (0.0885)	somalia (0.0517)	india (0.0334)	afghanistan (0.1023)
russia (0.0272)	iraqi (0.0566)	somali (0.0302)	indian (0.0274)	afghan (0.0760)
chechnya (0.0211)	baghdad (0.0457)	govern (0.0269)	kashmir (0.0224)	taliban (0.0676)
caucasus (0.0180)	mosul (0.0197)	mogadishu (0.0217)	pakistan (0.0180)	nato (0.0327)
kill (0.0145)	sunni (0.0151)	shabaab (0.0164)	isi (0.0147)	troop (0.0321)
chechen (0.0137)	citi (0.0143)	fight (0.0158)	mumbai (0.0121)	forc (0.0241)
oper (0.0127)	forc (0.0125)	islamist (0.0152)	attack (0.0106)	kabul (0.0197)
ingushetia (0.0104)	sourc (0.0115)	islam (0.0134)	let (0.0095)	insurg (0.0145)
north (0.0096)	aswat (0.0115)	shabab (0.0124)	arrest (0.0088)	karzai (0.0133)
moscow (0.0093)	shiit (0.0086)	town (0.0124)	lashkar (0.0087)	elect (0.0130)

Table 2: Ten most relevant words with their respective conditional probabilities for five topics associated between them to the "War on Terror" concept from the Ansar1 forum.

Topic 3 "**War**"	Topic 21 "**American Soldier**"	Topic 24 "**Imprisonment**"	Topic 25 "**Obama**"	Topic 31 "**Military Operations**"
peopl (0.0114)	islam (0.1238)	prison (0.0194)	presid (0.0146)	cia (0.0130)
war (0.0096)	emir (0.0964)	court (0.0137)	obama (0.0143)	report (0.0112)
countri (0.0092)	afghanistan (0.0811)	case (0.0104)	govern (0.0113)	million (0.0107)
american (0.0091)	provinc (0.0524)	charg (0.0099)	countri (0.0103)	work (0.0103)
world (0.0062)	destruct (0.0248)	releas (0.0092)	offici (0.0094)	compani (0.0100)
forc (0.0060)	american (0.0158)	detaine (0.0083)	secur (0.0092)	money (0.0095)
america (0.0058)	kill (0.0145)	tortur (0.0080)	militari (0.0082)	oper (0.0093)
one (0.0055)	soldier (0.0129)	alleg (0.0072)	year (0.0080)	blackwat (0.0082)
govern (0.0055)	tank (0.0104)	investig (0.0069)	minist (0.0079)	use (0.0077)
militari (0.0050)	enemi (0.0096)	guantanamo (0.0061)	state (0.0077)	state (0.0071)

Table 3: Ten most relevant words with their respective conditional probabilities for five topics associated between them to the "Recruiting" concept from the Ansar1 forum.

Topic 0 "**Religious Propaganda**"	Topic 36 "**Religious Conflicts**"	Topic 39 "***Ansar* propaganda**"	Topic 40 "**Family**"	Topic 37 "**Ideology**"
islam (0.0270)	mosqu (0.0267)	ansar (0.0940)	women (0.0272)	muslim (0.0475)
movement (0.0247)	muslim (0.0254)	wmv (0.0419)	school (0.0160)	islam (0.0471)
allah (0.0217)	protest (0.0163)	info (0.0204)	children (0.0157)	law (0.0116)
youth (0.0217)	anti (0.0109)	jihad (0.0188)	year (0.0149)	rule (0.0088)
media (0.0211)	peopl (0.0096)	final (0.0184)	old (0.0149)	scholar (0.0087)
apost (0.0184)	bodi (0.0089)	ansarnet (0.0140)	woman (0.0126)	say (0.0086)
mujahideen (0.0172)	prayer (0.0085)	showthread (0.0132)	men (0.0124)	issu (0.0070)
crusad (0.0169)	ramadan (0.0077)	issu (0.0120)	man (0.0121)	group (0.0069)
god (0.0149)	christian (0.0072)	video (0.0115)	girl (0.0109)	state (0.0068)
state (0.0133)	grave (0.0069)	media (0.0086)	student (0.0103)	call (0.0065)

Table 4: Top 9 members ordered by HITS's hub score over the social network for the complete topics evaluation for Creator-oriented network.

Complete Network		Topic-based (All topics)		Topic-based ("Local Conflicts")	
User	Hub	User	Hub	User	Hub
user1	0.6959	user2	0.0725	user2	0.0785
user2	0.1544	user3	0.0383	user3	0.0331
user3	0.1496	user4	0.0342	user4	0.0331
user24	0.0000	user21	0.0280	user20	0.0289
user237	0.0000	user20	0.0259	user5	0.0248
user53	0.0000	user8	0.0238	user40	0.0247
user36	0.0000	user5	0.0217	user6	0.0208
user231	0.0000	user40	0.0217	user13	0.0207
user14	0.0000	user13	0.0197	user21	0.0207

members using computed HITS hub score, it is possible to make decissions towards who's who in this topic-based network. This can be visualized in Figure 3 (b) top 8 hub score are presented from previously described network configuration, where members with higher score have stronger arcs between each other.

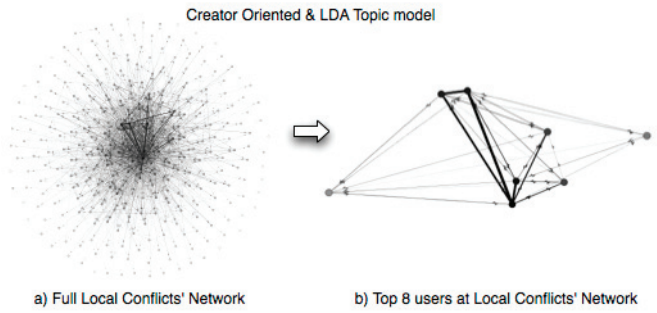

Figure 4: Visualization of the complete "local conflict" network and their top 8 users in the Creator-oriented network for the complete period of the forum activity, using the hub score from HITS.

4.1.3 Hub Based Social Network Analysis

In Table 5, the top 5 users are listed (with their respective hub score) for the Creator-oriented network and last Reply-oriented networks for all topics. In this case, it can be seen that in none of both cases, the users were repeated, which states that the topic-based network structure has completely different properties.

Table 5: Top 5 members ordered by HITS's hub score over the social network for the complete topics evaluation for both Creator-oriented network (*type 1*) and last Reply-oriented network (*type 2*).

All topics (*type 1*)	All topics (*type 2*)
user39 (0.07985)	user2 (0.2666)
user70 (0.07985)	user4 (0.2067)
user40 (0.07985)	user43 (0.1965)
user16 (0.07985)	user13 (0.1781)
user53 (0.07985)	user3 (0.1753)

Table 6: Top 5 members ordered by HITS's hub score over the social network for the "Local Conflict" concept for both Creator-oriented network (*type 1*) and Replier-oriented network (*type 2*).

"Local Conflicts" (*type 1*)	"Local Conflicts" (*type 2*)
user2 (0.3757)	user2 (0.3001)
user13 (0.2333)	user43 (0.2221)
user43 (0.2125)	user3 (0.2203)
user25 (0.2055)	user4 (0.2142)
user24 (0.1937)	user13 (0.1955)

In Table 6, top 5 members are listed, but for the "Local Conflicts" topic-filtered social network. In this case, the information that Creator-oriented and last Reply-oriented top lists highlights, is that forum members are repeated, showing that the "Local Conflicts" social network is specially built by posts of specific users (like "user2" and "user13").

5. CONCLUSION

We propose to combine traditional SNA with data mining techniques in order to produce results which enable to measure social aspects which are not considered by applying SNA alone. This way, we can obtain results closer to reality when performing further analysis like experts detection, sub-groups detection, centrality measures, among other measures, on any social network, virtual community, VCoP, VCoI, and many other human-based interactions.

We used our approach to study a VCoI called the Dark Web, for which the `Ansar1` forum was used. This forums was collected from Dec 2008 to Jan 2010, and was created by 376 members in 29057 posts. By applying latent Dirichlet allocation (LDA) and using average shortest distances (ASD), we were capable to obtain 47 topics using the closes 30 words on each topic. Afterwards, 11 topics were discarded by inspection. Finally, we used two different topology representations of the forum: a creator-oriented and last reply-oriented networks.

We showed in which ways SNA alone could be very poor in order to detect key-members which are specifically talking about certain subjects. However, when combining a topic-based text mining approach with SNA we outperforms SNA alone to discover VCoIs' key members, for which a better analysis of social networks can be performed.

In our experiments, we draw a complete 14 months social network which is quite dense and performed SNA. The information obtained was useless as its visual representation does not helps to identify any patterns. However, using our method, we were able to focus on a specific group of topics to create a topic-based network, which provides specific information since it has a lesser density. Then, by using HITS, key members can be extracted from both network configurations, and results also are radically different.

We can say that our proposal can lead to much better results that applying SNA alone. However, as future work the overall evaluation of the top ranked members by the HITS algorithm can be potentiated by experts interviews. Also, by combining other SNA based measures and techniques, such as betweenness centrality scores, weighted-HITS, and HITS authoritative scores, the key-members could be identified with a higher performance. Besides, simple visual inspection of topic-based SNA allow to view a perfect sub-group with key members.

We can say that we have successfully tested our approach and we have shown how to apply it into the Dark Web to discover potential groups of topics (e.g. potential homeland security threats), and key-members that potentiate these topics.

6. ACKNOWLEDGEMENTS

Authors would like to thank the continuous support of "Instituto Sistemas Complejos de Ingenieria" (ICM: P-05-004- F, CONICYT: FBO16); Initiation into Research Funding, project code 11090188, entitled "Semantic Web Mining Techniques to Study Enhancements of Virtual Communities"; The Social Network Analysis Research Group (`sna.dii.uchile.cl`); and the Web Intelligence Research Group (`wi.dii.uchile.cl`).

7. REFERENCES

[1] A. Abbasi and H. Chen. Applying authorship analysis

to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[2] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26(3):1–34, 2008.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning ...*, Jan 2003.

[4] A. Bourhis, L. Dubé, and R. Jacob.... The success of virtual communities of practice: The leadership factor. *The Electronic Journal of Knowledge ...*, Jan 2005.

[5] R. B. Bradford. Application of latent semantic indexing in generating graphs of terrorist networks. pages 674–675, 2006.

[6] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 1978.

[7] W. Kim, O.-R. Jeong, and S.-W. Lee. On social web sites. *Information Systems*, 35(2):215–236, 2010.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[9] M. Kosonen. Knowledge sharing in virtual communities – a review of the empirical research. *Int. J. Web Based Communities*, 5(2):144–163, 2009.

[10] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon. Mining communities in networks: a solution for consistency and its evaluation. pages 301–314, 2009.

[11] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial ...*, Jan 2007.

[12] R. D. Nolker and L. Zhou. Social computing and weighting to identify member roles in online communities. *Web Intelligence, IEEE / WIC / ACM International Conference on*, 0:87–93, 2005.

[13] N. Pathak, C. Delong, A. Banerjee, and K. Erickson. Social topic models for community extraction. Aug 2008.

[14] U. Pfeil and P. Zaphiris. Investigating social network patterns within an empathic online community for older people. *Computers in Human Behavior*, 25(5):1139–1155, 2009.

[15] X. H. Phang and C. Nguyen. Gibbslda++, 2008.

[16] C. E. Porter. A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of Computer-Mediated Communication*, 10(1):00, 2004.

[17] G. Probst and S. Borzillo. Why communities of practice succeed and why they fail. *European Management Journal*, 26(5):335–347, 2008.

[18] E. Reid, J. Qin, Y. Zhou, G. Lai, M. Sageman, G. Weimann, and H. Chen. Collecting and analyzing the presence of terrorists on the web: A case study of jihad websites. pages 402–411, 2005.

[19] S. A. Ríos, F. Aguilera, and L. Guerrero. Virtual communities of practice's purpose evolution analysis using a concept-based mining approach. *Knowledge-Based and Intelligent Information and Engineering Systems*, 2:480–489, 2009.

[20] M. Sageman. A strategy for fighting international islamist terrorists. *ANNALS of the American Academy of Political and Social Science*, 618(1):223–231, 2008.

[21] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, Vol. 18(11):613–620, 1975.

[22] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. 1994.

[23] D. Xing and M. Girolami. Employing latent dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, Vol. 28(13):1727–1734, 2007.

[24] J. Xu and H. Chen. Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems (TOIS)*, Jan 2005. Aqui sale bien el blockmodeling.

[25] J. Xu and H. Chen. The topology of dark networks. *Commun. ACM*, 51(10):58–65, 2008.

[26] L. Yang, F. Liu, J. Kizza, and R. Ege. Discovering topics from dark websites. In *CICS '09: IEEE Symposium on Computational Intelligence in Cyber Security.*, pages 175–179. IEEE, 2009.

[27] L. Yang, F. Liu, J. Kizza, and R. Ege. Discovering topics from dark websites. *Computational Intelligence in Cyber Security, 2009. CICS '09. IEEE Symposium on*, pages 175 – 179, 2009.

[28] Y. Zhang, S. Zeng, L. Fan, Y. Dang, C. A. Larson, and H. Chen. Dark web forums portal: searching and analyzing jihadist forums. pages 71–76, 2009.

[29] Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai. Us domestic extremist groups on the web: Link and content analysis. *IEEE Intelligent Systems*, 20(5):44–51, 2005.