Anomaly Detection in Extremist Web Forums Using a Dynamical Systems Approach

Steve Kramer[1]
Paragon Science, Inc.
9104 Breeze Point Cove
Austin, TX 78759
512-569-9760

steve.kramer@paragonscience.com

ABSTRACT

In this paper, we present preliminary results of analyzing data from the Dark Web collection using a dynamical systems approach for unsupervised anomaly detection. The goal is to provide a robust, focus-of-attention mechanism to identify emerging threats in time-dependent, unlabelled data sets. In our method, finite-time Lyapunov exponents are used to characterize the time evolution of both the directed network structure and the distribution of text attributes in the forum messages. We provide a description of the technique and a summary of the initial anomaly detection results. We conclude with a summary of results and a brief discussion of promising avenues for future research.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence – *learning.*

H.2.8 [**Information Systems**]: Database Management – *database applications, data mining.*

General Terms

Algorithms, Security, Theory.

Keywords

Unsupervised anomaly detection, data mining, dynamical systems, finite-time Lyapunov exponents.

1. INTRODUCTION

In this paper, we present preliminary results of analyzing data from the Dark Web collection using a dynamical systems approach for unsupervised anomaly detection. The goal is to provide a robust, focus-of-attention mechanism to help identify emerging threats in time-dependent, unlabelled data sets. In our method, finite-time Lyapunov exponents are used to characterize the time evolution of both the directed network structure and the distribution of text attributes in the forum messages. Section 2 includes a description of our method for unsupervised or semi-supervised dynamic anomaly detection in time-dependent data sets. Section 3 presents initial research results based on the time evolution of the text attributes in the Ansar1 forum messages. Section 4 describes a subsequent analysis that includes a *k*-core decomposition of the Ansar1 network followed by anomaly detection using both the text attribute distribution and the local network structure. Section 5 includes a brief discussion of promising avenues for future research. Section 6 provides a summary of results.

2. DYNAMIC ANOMALY DETECTION USING FINITE-TIME LYAPUNOV EXPONENTS

The field of anomaly detection research spans work in a broad range of scientific disciplines, including applied mathematics, statistics, physics, computer science, data mining, engineering, and artificial intelligence. Some important sample applications of anomaly detection methods include financial fraud prevention (for example, in credit card payment processing), telephone fraud prevention, military battlespace awareness, surveillance for disease outbreaks or chemical or biological weapons, counterterrorism data mining programs, computer antivirus systems, anti-spam systems, and computer network intrusion detection and prevention systems.

We have developed a novel data-mining technique for anomaly detection in time-dependent data sets [1] based on concepts from dynamical systems theory. In our approach, finite-time Lyapunov exponents [2] are calculated from the source data and then are used to characterize the time evolution of the system. The aim is to provide a robust capability for unsupervised anomaly detection without requiring the user to specify normal vs. abnormal behavior *a priori*. The high-level steps in our method are shown below in Figure 1.

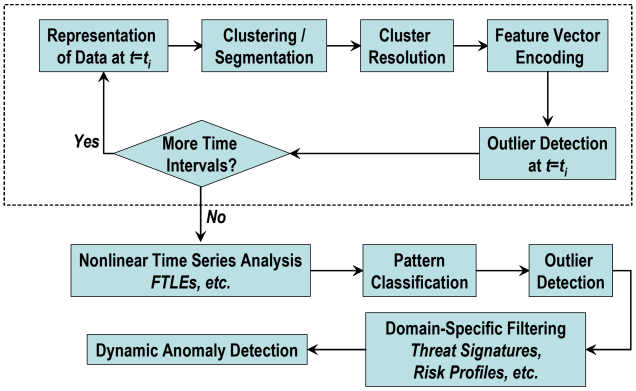
Representation of Data at $t=t_i$
Clustering / Segmentation
Cluster Resolution
Feature Vector Encoding
Outlier Detection at $t=t_i$
More Time Intervals?
Yes
No
Nonlinear Time Series Analysis FTLEs, etc.
Pattern Classification
Outlier Detection
Domain-Specific Filtering Threat Signatures, Risk Profiles, etc.
Dynamic Anomaly Detection

Figure 1

[1] Steve Kramer, Ph.D., is a computational physicist and the President and Chief Scientist of Paragon Science, Inc.

2.1 Clustering input data

Especially for large data sets, an important initial step is to apply a clustering algorithm to the incoming data set in order to separate it into smaller clusters for analysis. Clusters can consist of overlapping sets of nodes, or a cluster can comprise a single node for detailed examination. In Figure 1, the input data are subdivided into clusters and sequential time intervals for analysis. Standard clustering algorithms can be applied, such as spectral clustering [3], k-groups/GDA [4], or clique percolation [5], depending on the nature and size of the input data. If clustering is applied incrementally at the beginning of each time interval (rather than clustering once longitudinally over all time values), then cluster resolution might be required to track clusters that are newly born, recently expired, merged, or split.

2.2 Encoding feature vectors

Each cluster is then encoded into one or more feature vectors, which capture the state of each cluster at a given time. Multiple feature-vector-encoding methods can be applied simultaneously to "slice and dice" the evolving data set from different perspectives or using complementary analysis approaches. Many standard dimensionality-reduction methods, such singular value decomposition [6], could be used to encode feature vectors for each cluster and for each time interval.

One of Paragon Science's proprietary feature-vector-encoding methods can incorporate not only information about directed links (that is, which entity interacted with which other entity and when) but also any discrete- or continuous-valued attributes associated with time-dependent data sets. For example, this technique can be applied to text content from emails, websites, blogs, etc., where the individual textual terms are treated as discrete-valued attributes. As an example of a continuous attribute, time-stamped geospatial location data of insurgent attacks or sightings could likewise be analyzed.

2.3 Calculating finite-time Lyapunov exponents

The finite-time Lyapunov exponent (FTLE), σ, is a scalar value that characterizes the amount of stretching or contracting of infinitesimally close trajectories in a dynamical system, starting from an initial point x_0 in phase space during a time interval of finite length. The FTLE provides information about the local stability and predictability of a dynamical system.

$$
\begin{aligned}
\dot{\vec{x}}(t;t_0,\vec{x}_0) &= \vec{v}(\vec{x}(t;t_0,\vec{x}_0),t) \\
\vec{x}(t;t_0,\vec{x}_0) &= \vec{x}_0
\end{aligned}
$$

Equation 1

Within this analysis framework, the state of each cluster is defined by the feature vector x. The feature vector x can be interpreted as a pseudo-"position" vector within the state space spanned by all possible values of the components of the feature vectors. As each cluster changes over time, its feature vector x will be transported from its original position x_0 to position x by the flow map ϕ, which defines the time evolution of the system. Lekien, *et al.*, extended FTLEs to N-dimensional systems [7], and subsequently we developed a proprietary method of analyzing general time-dependent data sets for which the Jacobian of the flow map ϕ is not known analytically but can instead be estimated from measured trajectories in the feature vector state space. Once the Jacobian has been estimated, the FTLE σ can be calculated, according to Equation 2 below, in which $d\phi/dx$ is the Jacobian of the flow map, T is the finite time period over which the change is calculated, and $\lambda_{\max}$ is the largest eigenvalue of the matrix product Δ, which is analogous to the right Cauchy-Green deformation tensor in continuum mechanics.

$$
\begin{aligned}
\Delta &= \frac{d\phi_{t_0}^{t_0+T}(\vec{x}_0)^{*}}{d\vec{x}} \, \frac{d\phi_{t_0}^{t_0+T}(\vec{x}_0)}{d\vec{x}} \\
\sigma_{t_0}^{T}(\vec{x}) &= \frac{1}{T}\ln\sqrt{\lambda_{\mathbf{max}}(\Delta)}
\end{aligned}
$$

Equation 2

2.4 Identifying anomalies

After the FTLEs have been obtained for the requested feature-vector types, the next step is to identify potential anomalies. Because the FTLEs inherently quantify the rate of change of the clusters over time, the FTLEs can be used effectively to identify clusters that have shifted their behavior abruptly relative to their past and to identify clusters which are evolving significantly differently from other clusters in the data set.

One approach for anomaly detection in our framework is to define an adaptive threshold by scaling the deviation of each FTLE σ from the mean value for that time interval and that type of feature vector by the corresponding standard deviation. Values that exceed a user-specified threshold value can be flagged as anomalies.

A complementary approach that we have used is to calculate the time derivative of σ and then to identify the top N peaks of $d\sigma/dt$ as additional anomalies. Similarly, a scaled deviation of $d\sigma/dt$ from the mean for each time interval could be calculated for anomaly identification.

As an optional post-processing step, the results of the FTLE analysis can be supplied as input to a number of the standard machine learning tools, such as neural networks or support vector machines, for pattern classification, if desired.

2.5 Relating anomalies to source data

In order to help make sense of the identified anomalies, it is important to provide the user of the system with information about why the selected clusters were flagged as being anomalous. The method for linking anomalies to source data depends upon the type of feature-vector encoding that was employed for the sets of anomalous FTLEs.

For example, if the singular value decomposition of the adjacency matrix of a directed network was used to encode feature vectors, then a network visualization tool could be applied to highlight the links that were added, deleted, or changed within the selected cluster during the corresponding time interval.

If Paragon Science's joint network-and-attribute method had been used to encode the feature vectors, then the abovementioned network visualization could then be used in conjunction with an inverse transformation in order to highlight the discrete or continuous attributes that changed significantly and contributed to the peak value of the FTLE or its time derivative. For example, if discrete text attributes had been incorporated in the analysis, the specific attributes that contributed to the anomalies could be highlighted within the source data, as will be shown in Section 3.

In its unfiltered mode of operation, our software runs in an unsupervised fashion on the available input data sets using all defined values of the discrete or continuous attributes. However, if the user wishes to reduce the scope of the analysis, then arbitrarily general filters can be applied at the data input stage and/or during a post-processing step. Additionally, the system could be configured to accept positive or negative feedback by the user on the anomalies identified in order to tune the system's performance in a semi-supervised mode.

3. TEXT TERM ANOMALIES IN THE ANSAR1 DATA SET

As a first example, we used our anomaly detection software to characterize the time evolution of the text attributes in the Ansar1 forum messages in the CSI-KDD Challenge data set. In that forum, there are 29,056 messages posted by 379 members, with most messages in English, some in Arabic, and a few with Russian. The messages span a date range from 08 December 2008 to 02 January 2010.

3.1 Setting up the analysis

We used the WVTool software package [8] associated with the open-source RapidMiner data mining toolkit [9] to extract the terms from the forum messages. Without any pruning or stemming of terms, WVTool produced a word list of 215,933 terms, ranging in number of occurrences from 1 (131,431 such terms) to 14,337 (for the word "the").

The default mode of our software is to incorporate all extracted text attributes in the analysis and to add new terms to the word list as they are found. In that mode, the system can automatically detect the sudden appearance of new sets of terms as well as unexpected shifts in the usage of terms by groups or individuals. If there are certain terms of interest, the user can also opt to filter the analysis based on the selected terms.

In many cases, it is standard practice in natural language processing to prune the unabridged word list by frequency, and we performed a few initial tests with a pruned Ansar1 word list that included terms having frequencies between 200 and 2000 occurrences within the forum messages. We are also currently considering several automated approaches for selecting or weighting text dimensions for analysis.

However, because the stated aim of the CSI-KDD Challenge is to identify threatening or radical members and messages, we opted to edit the frequency-pruned word list manually to focus on 370 terms in both English and Arabic that relate directly or indirectly to threats, violence, or conflict. An excerpt of our threat keyword list is given below in Table 1, and the full list is available online in UTF-8-encoded format for use by other researchers (http://www.paragonscience.com/data/ansar1-wordlist-threats.txt).

Table 1: Excerpt of threat keyword list

Term	Frequency
just	1894
army	1846
ةقطنم (area)	1825
ودهاجم (Mujahideen)	1786
fighting	1785
brothers	1692
militants	1680
killing	1589
ودعلا (enemy)	1571
Mujahideen	1566
…	
ةحلسأب (arms)	203
نمأ (security	203
harm	203
volatile	202
opportunity	202
Jewish	202
حارجب (injured)	201
حلسم (armed)	201
zippyshare	201
illegal	200

3.2 Performing the FTLE calculations

Using the focused word list of 370 threat-related terms, we analyzed the time evolution of the distribution of text attributes used in the forum postings of each of the 379 listed members using an analysis time interval of one (1) week.

Figure 2 is a 3-D plot that displays the scaled deviation from the moving mean of the FTLE of the text attribute distribution, with red areas indicating the greatest deviation above the mean. The horizontal axis is the node ID of each forum member, and the vertical axis is the Unix timestamp, which is the number of seconds since January 1, 1970. (In the figure titles in this paper, the phrase "Change Metric" refers to the corresponding type of FTLE.)

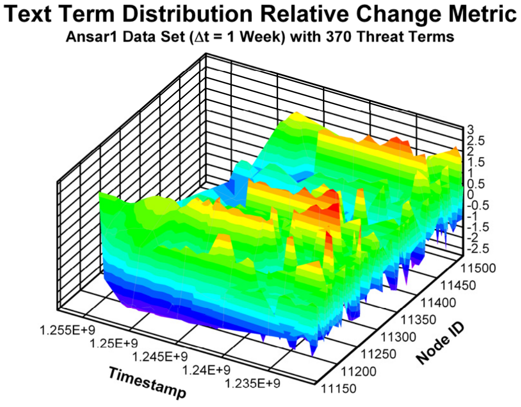

Figure 2

Figure 3 is a detailed view of the same plot, this time from above and enlarged to show one of the FTLE peaks in Figure 2. FTLE values that exceed a user-specified threshold can be classified as potential anomalies and then subjected to further analysis.

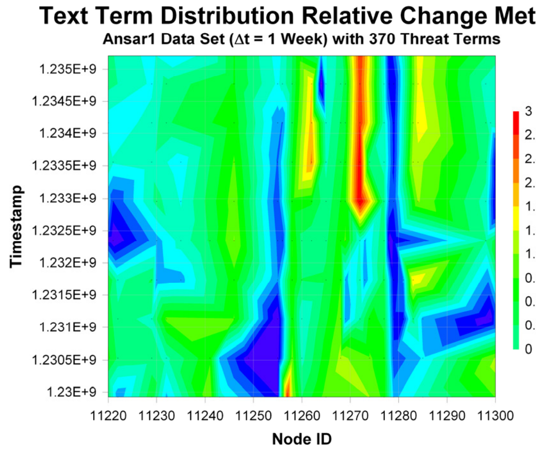

Figure 3

As mentioned earlier, prospective anomalies can also be obtained by looking for peaks in the time derivative of the FTLEs. Figure 4 shows a heat map of the calculated time derivative of the FTLEs of Figure 2.

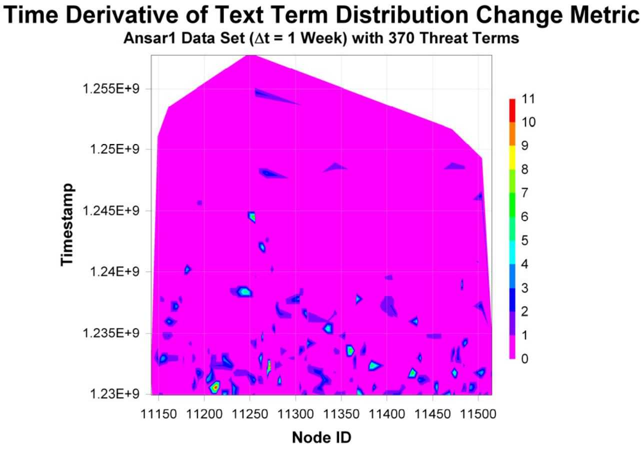

Figure 4

The sets of prospective anomalies identified using the approaches above can be compared in order to focus subsequent analyses on the most-anomalous time periods and nodes. In general, when multiple types of feature vectors are used, FTLE results across the multiple feature-vector types could be weighted so that only those anomalies that exceed a user-defined threshold are considered for reporting and further analysis.

3.3 Examining the corresponding term distributions

When an anomaly is selected for a particular time interval, we can re-examine the source data that contributed to the corresponding feature vectors, which represent the state of the system during that time interval. Figure 5 is a 3-D plot that shows the amplitude of the text attribute distribution as a function of the local node ID (within the set of members who posted messages during the selected analysis time period) and local term ID (within the set of terms that appear during the same period). To calculate the term weights that contribute to each member's feature vector, we used the default WVTool setting, which is the TF-IDF (term frequency–inverse document frequency) weighting, a standard statistical measure employed in information retrieval and text classification.

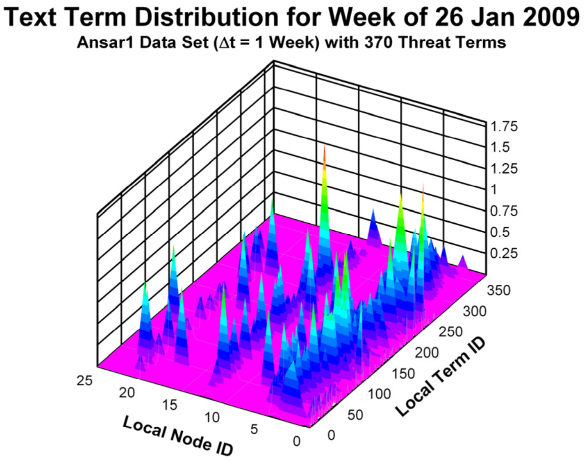

Figure 5

Once the text term distributions that relate to the anomalous FTLE values have been retrieved, the non-zero components can be analyzed to determine which terms either contributed the most or changed the greatest extent from the previous time period.

For the period containing the anomalous FTLE peak shown in Figure 3, the following terms had the largest contributions or greatest changes.

Table 2: Terms with largest contributions (Case #1)

Term	Total Contribution
troop	5.10
insurgent	3.64
fight	3.62
Israeli	3.62
shot	3.28

Table 3: Terms with greatest changes (Case #1)

Term	Change in contribution
shot	1.86
patrol	1.74
huge	1.50
firing	1.44
fight	1.42

3.4 Linking to the corresponding source forum messages

As noted earlier, one of the primary goals of our unsupervised (or semi-supervised) dynamic anomaly detection system is to provide a robust focus-of-attention mechanism to highlight anomalous entities or events without requiring the user to define in advance what constitutes normal vs. abnormal behavior.

Once we have identified the prospective anomalies, it is important to provide as much contextual information as possible about why those data elements were selected as anomalous. In general, this process depends upon the types of feature-vector encoding used. In this case, because the feature vectors were generated using a list of threat-related terms as a basis set, we can relate the anomalies to the terms with the greatest contributions or changes (see Table 2 and Table 3) and then highlight those terms in the forum messages that were posted during the specific time period. Table 6 in Appendix A shows the five top-ranked forum messages related to the FTLE peak under consideration. Most of the tagged messages are reposted new articles, primarily related to insurgents' activities and their tactics, as reported by the news media. One message, excerpted below, that does stand out somewhat is a opinion piece written by Aqil Abdul Razzaq Khan and posted by Nasrullah on 28 January 2009 that exhorts Muslims to stop financing the American economy and hence its support for Israel.

> "The bottom line is that the Muslims have played a significant role in keeping the American economy alive and have provided the cash-strapped American government the money it needs to finance the Israeli atrocities against the Palestinian people. … I salute the efforts of those Muslims who are boycotting American products, but I have to say with great regret that until we have people in our Ummah who are willing to earn money at the expense of the blood of their brothers, there is not much we as individuals can do to stop the Israeli crimes. We all know the obvious enemy, it's important we know the enemy within us."

At this point in the workflow, the results could be presented to an intelligence analyst for manual investigation, and then subsequent rounds of FTLE calculation and dynamic anomaly detection could be performed with alternative feature-vector encoding methods. The user can modify the filtered list of text terms, adding or removing words as desired. Additionally, the Paragon Science software results could be supplied as input to other machine learning or data mining tools for complementary analysis.

4. JOINT NETWORK-TERM ANOMALIES IN THE ANSAR1 DATA SET

As a second example, we used our anomaly detection software to characterize simultaneously the time evolution of the text attributes and of the sub-networks centered on each of a set of specific members of interest.

4.1 Performing an initial *k*-core decomposition

The *k*-core of graph is a maximal subgraph in which each vertex has at least degree *k*. The coreness of a vertex is *k* if it belongs to the *k*-core but not to the (*k*+1)-core. The *k*-core decomposition is performing by recursively removing all the vertices (along with their respective edges) that have degrees less than *k*. Recent research [10] has suggested that the *k*-core decomposition of a network can be very effective in identifying the individuals within a network who are best positioned to spread or share information.

We used the LaNet-vi tool developed by Prof. A. Vespignani of Indiana University and his collaborators [11] to perform a *k*-core decomposition of the bipartite network formed by the links among members and threads in the Ansar1 data set. In this undirected network representation, there is a link between a member and a thread for each post a member makes to a thread. Figure 6 shows the resulting decomposition. The members and threads with the highest coreness are displayed at the center of the figure. The size of each vertex reflects the degree of that vertex, which is either a forum thread or member.

There are 90 forum members who have the highest coreness value of 8 for the Ansar1 network. We suggest that these are the mostly highly connected and/or influential individuals within that forum community and thus merit further investigation.

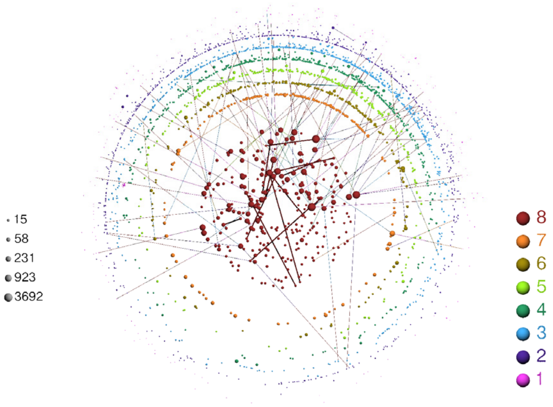

Figure 6: *k*-core decomposition of Ansar1 network

4.2 Calculating the joint network-term FTLEs

For this second round of dynamic anomaly detection calculations, we chose to focus on the 90 forum members in the 8-core of the Ansar1 network. For each of those members, we computed feature vectors that simultaneously encoded information about (1) the distribution of text terms in that member's postings and (2) the one-hop sub-network formed by treating each forum posting as a link between that member and the destination forum thread.

Figure 7 is a 3-D plot that displays the scaled deviation from the moving mean of the FTLE of the joint network and text attribute distribution as a function of cluster ID (corresponding to each of the members in the 8-core) and time. Peaks in this plot can be considered as potential anomalies.

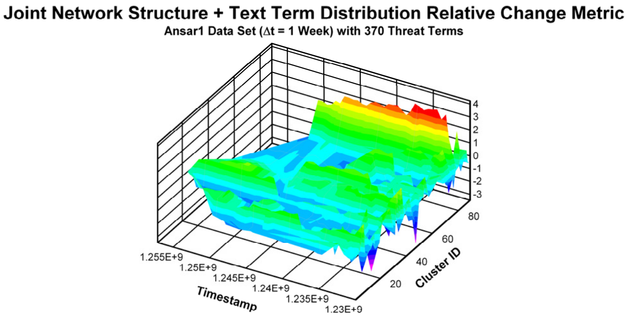

Figure 7

As before, prospective anomalies can alternatively be obtained by searching for peaks in the time derivative of the FTLEs. Figure 8 shows a 3-D plot of the calculated time derivative of the FTLEs of Figure 7.

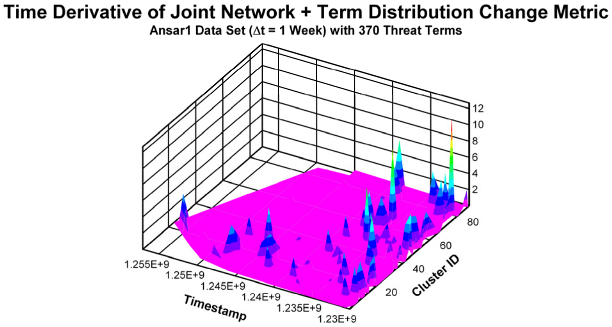

Figure 8

Figure 9 is a detailed view of the same plot, this time from above and enlarged to show one of the FTLE peaks in Figure 8.

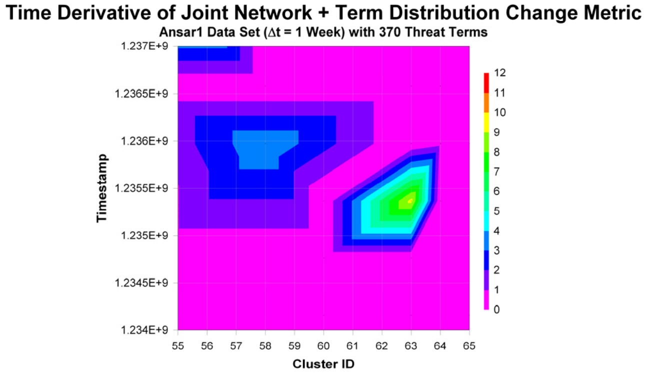

Figure 9

4.3 Examining the corresponding term distributions

We follow our earlier approach of re-examining the source data that contributed to the feature vectors that correspond to the anomalous FTLE instances. In Figure 10, the highest peak corresponds to the term "troop" in the threat word list (local term ID=53).

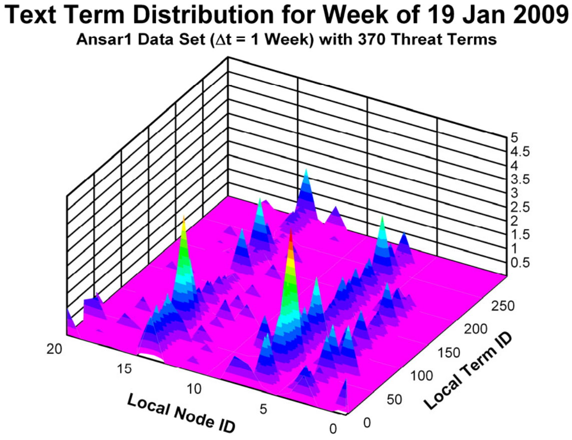

Figure 10

As before, we examine the non-zero components to determine which terms either contributed the most or changed the greatest extent from the previous time period.

For the period containing the anomalous FTLE peak shown in Figure 9, the following terms had the largest contributions or greatest changes. We note that because this second set of FTLEs was calculated using a different type of feature vectors (network + text attributes), we can discover different anomalies. In these results, we note in particular the appearance of two keywords ("rapidshare" and "megashares") that correspond to online file-sharing websites used by forum members.

Table 4: Terms with largest contributions (Case #2)

Term	Total Contribution
troop	9.88
rapidshare	8.90
online	7.33
bomb	6.47
megashares	6.08

Table 5: Terms with greatest changes (Case #2)

Term	Change in contribution
troop	4.05
death	3.49
just	3.42
bomb	3.24
online	3.09

4.4 Linking to the corresponding source forum messages

As noted above, the anomalous peak shown in Figure 9 is related to the terms in Table 4 and Table 5, and the corresponding source forum messages contain a significant number of links to external file-sharing websites and could be indicative of a sudden shift during that period for forum members to read and download those external files. An excerpt of one of the top-ranked related messages is provided below.

> بسم الله الرحمن الرحيم مؤسسة الأنصار الإعلامية تقدم الحَمدُ للهِ ناصِرِ المؤمِنين وهازِمِ
> وقائدِ المُجاهدين, والصَلاة والسَلام على نبينا مُحمد إمامِ المُتقين, الكَفَرةِ والمُشركين
> دْ جَاءَتْكُم مَّوْعِظَةٌ مِّن رَّبِّكُمْ يَا أَيُّهَا النَّاسُ قَ}:قال تعالى :أمابعد ,وعلى آلهِ وصَحبهِ أجمعين
> هللا نم ديدستب [57:سنوي]{وَشِفَاءٌ لِّمَا فِي الصُّدُورِ وَهُدًى وَرَحْمَةٌ لِّلْمُؤْمِنِينَ
> راصنالا دوسأ نيدهاجملا مكتوخإ نكمت ،هدنع نم رصنو ريدقلا يلعلا
> ةجيوحلا قوس نم برقلاب ةيكيرمأ ةيلآ باطعا نم كوكرك ةنيدم يف
> طباورلا دحا مادختسا ىجري ةيلمعلا ريوصت ةدهاشملو ،ريبكلا
> ةيلاع :ميركلا مكئاعد صلاخ نم نيدهاجملا مكتوخا اوسنت الو ،ةيلاتلا
>
> [Translation from Google:
>
> In the name of God the Merciful Al-Ansar Media offer praise to God Nasser believers and Conqueror of the infidels and idolaters, and blessings and peace upon our Prophet Muhammad Imam of the pious, the commander of the Mujahideen, and his family and companions, Omapad: The Almighty said: (O people, may have come to an admonition from your Lord and a healing for what is in the breasts and guidance and mercy for the believers) [Yunus: 57] by Allah the Almighty and a victory from him, enabling your brothers the Mujahideen black supporters in the city of Kirkuk to damage mechanism American market near Hawija great, to see the filming process, please use one of the following links, and not forget your brothers the Mujahideen from pure Karim allegations: High]

4.9 MB http://fyad.org/xyl7 http://url.file.am/?HJe94 http://ansariw.notlong.com/ http://uploaded.to/?id=weqyat
http://ww2.megashare.com/615699
http://www.damfile.com/?d=65F59EDD1
http://www.badongo.com/vid/1043899
http://www.sendspace.com/file/viy5gt
http://www.fileflyer.com/view/t99dVBp
http://www.fileflyer.com/view/5Lxx2Bs
http://www.fileflyer.com/view/DMejQB6
http://www.xtraupload.de/?d=4D8281004
http://www.xtraupload.de/?d=0536410E4
http://www.megaupload.com/?d=Q55N38C3

5. FUTURE RESEARCH

Due to the limitations of both publication space and research time prior to the submission deadline, the results presented here are preliminary and are intended primarily to provide an initial indication of how our dynamic anomaly detection methods can be applied. There is much follow-on research work that we plan to undertake in this area.

Because there are several tunable parameters and user-selectable thresholds in our analysis framework, we would like to determine optimal settings for the analysis of various types of source data. To date, we have found it very challenging to obtain publicly available real-world data with both timestamps and labeled anomalies. Having ground-truth information about known anomalies would allow us to investigate precision/recall performance tradeoffs for our technique.

More specifically regarding the CSI-KDD Challenge data set, we plan to study the distribution of URLs within the posted forum messages in order to characterize the dynamics of the online resources outside of the forum that are used by forum members to disseminate information and to recruit and train new members. We anticipate that that line of research will be particularly fruitful. It would also be interesting to analyze two-hop sub-networks centered on each of the 90 members in the 8-core of the Ansar1 forum network. Extending the sub-networks to two hops (member ↔ forum thread ↔ member) will yield another way to find infectious members and threads. Finally, we have not yet had time to study the other extremist forums available in the Dark Web portal, so it will be interesting to see what new insights can be gained by doing so.

6. SUMMARY

In this paper, we have briefly described our approach to dynamic anomaly detection in time-dependent data sets, initial research results on the Ansar1 forum in the CSI-KDD Challenge data set, and our planned directions for future research. In contrast to alternative methods based on explicit rules or probabilistic models, our dynamical systems technique has the advantage of being able to characterize the time evolution of the data sets under investigation in such a way that anomalous changes can be identified without requiring the user to state in advance what constitutes normal vs. abnormal behavior. Because our framework assumes that all entities can change in an arbitrarily general way over time, it can identify entities that have shifted in behavior unexpectedly or that are evolving in ways that are significantly different from other entities in the data set. We are hopeful that our software can function as a robust focus-of-attention mechanism as a complement to other anomaly detection methods that are currently being used.

7. ACKNOWLEDGMENTS

We gratefully acknowledge the generous assistance of Prof. Chen and his Artificial Intelligence Laboratory research team at the University of Arizona for providing access to the CSI-KDD Challenge data sets in their Dark Web portal collection.[12],[13] We also thank Prof. Vespignani and his collaborators for making their LaNet-Vi network visualization and analysis program available. Finally, we are grateful to the RapidMiner development team for releasing an open-source version of their flexible data-mining platform.

8. REFERENCES

[1] Kramer, S., Systems and Methods for Dynamic Anomaly Detection, U.S. Patent Application #12046394, March 11, 2008.

[2] Ziehmann, C., Smith, L.A., and Kurths, J., 2000. Localized Lyapunov exponents and the prediction of predictability, Phys. Lett. A, 4, 237-251.

[3] J. Shi, J. and J. Malik, J., 1997. Normalized cuts and image segmentation, Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), 731.

[4] Kubica, J., Moore, A., and Schneider, J., 2003. Tractable group detection on large link data sets, The Third IEEE International Conference on Data Mining.

[5] Palla, G., Barabasi, A-L. , and Vicsek, T., 2007. Quantifying social group evolution, Nature, 446, 664-667.

[6] Stewart, G.W., 1993. On the early history of the singular value decomposition, SIAM Review, 35, Issue 4, 551-566.

[7] Lekien, F., Shadden, S., and Marsden, J., 2007. Lagrangian coherent structures in n-dimensional systems, J. Math. Physics, 48, 065404.

[8] Wurst, M., The Word & Web Vector Tool, http://nemoz.org/joomla/index.php.

[9] RapidMiner Data Mining Tool, www.rapidminer.com.

[10] Kitska, M., *et al.*, 2010. Identifying influential spreaders in complex networks, arXiv:1001.5285v1 [physics.soc-ph].

[11] Alvarez-Hamelin, I., Dall'Asta, L., Barrat, A., and Vespignani, A., LaNet-vi: Large Network visualization tool, http://xavier.informatics.indiana.edu/lanet-vi.

[12] H. Chen, *et al.*, 2008. Uncovering the Dark Web: A Case Study of Jihad on the Web, Journal of the American Society for Information Science and Technology, 59, No. 8, 1347-1359.

[13] H. Chen, *et al.*, Dark Web Forum Portal, http://ai.arizona.edu/research/terror/.

9. APPENDIX A

Table 6: Five top-ranked Ansar1 forum messages related to anomaly case #1

Member	Thread	Posting Date	Score	Message
ShabaabQoqaz	909	2009-01-28	22	***Israeli*** jets have carried out fresh air raids on tunnels under the Gaza-Egypt border, reportedly sending hundreds of people fleeing their homes in panic. Local residents and Hamas security officials said three air attacks took place before dawn on Wednesday, but no casualties have yet been reported. The raids came just hours before the newly appointed US Middle East peace envoy was due to arrive in Israel, and after an attack on an ***Israeli*** army ***patrol*** that killed an officer and wounded three other soldiers. "IAF [***Israeli*** Air Force] aircraft recently hit a number of Hamas smuggling tunnels on the southern border of the Gaza Strip," a statement released on Wednesday by the ***Israeli*** army said. "This was in response to the attack against an IDF [***Israeli*** army] force in the area of Kissufim on the morning of January 27th, in which one IDF warrant officer was killed and three other IDF personnel were wounded, including one severe injury, when Palestinians detonated an explosive device against an IDF force patrolling on the ***Israeli*** side of the Gaza Strip security fence." Tunnel raid Israel says the attacks on the Rafah tunnels are aimed at stopping alleged weapons smuggling into the Gaza Strip by Hamas fighters. The tunnels are also used to smuggle food, fuel and consumer goods from Egypt and are considered a life-line for thousands of ordinary Gazans. Amid the continuing tension, George Mitchell, the US envoy to the Middle East, is on a tour of the region aimed at promoting a durable peace between Israel and the Palestinians. He began a series of meetings with regional leaders on Wednesday, holding talks with Hosni Mubarak, the Egyptian president, in Cairo. Mitchell and Mubarak discussed an Egyptian initiative aimed at restoring relative calm between Israel and the Palestinians and the re-opening of Gaza's border points. "The United States is grateful to Egypt for its leadership in bringing about a ceasefire. It is of critical importance that the ceasefire be extended and consolidated," Mitchell said after the meeting. "The United States is committed to vigorously pursuing a lasting peace and stability in the region. The decision by President Obama to dispatch me to this region less than one week after his inauguration is clear and tangible evidence of this commitment." Mitchell is set to meet Shimon Peres, the ***Israeli*** president, and Ehud Olmert, the ***Israeli*** prime minister, in Jerusalem later on Wednesday. He is due to travel to the Palestinian West Bank on Thursday. Al Jazeera's Barnaby Phillips, reporting from Jerusalem, said: "Mitchell has turned into a little bit more of a firefighter than he originally thought when he scheduled this tour. "The original intention, as US president Barack Obama said yesterday, was that Mitchell would go to the Middle East to listen and to learn - to show that the United States is not going to dictate terms in the Middle East." "Having said all that, the situation on the ground here has deteriorated in the past 24 hours." Hamas member killed The fresh jet raid on Gaza's tunnels came despite fragile ceasefires declared by Israel and Hamas last week, ending a 22-day ***Israeli*** military campaign on Gaza in which 1,300 people were killed. Another air raid shortly afterwards killed a Palestinian on a motorcycle whom an ***Israeli*** army spokesman identified as the planner of the roadside bomb attack. Hamas confirmed that one of its members riding a motorcycle was injured in the attack, which occurred in the town of Khan Younis. Neither Hamas nor any other group has claimed responsibility for Tuesday's bomb attack targeting an ***Israeli*** army ***patrol*** along the Gaza border. After the blast, ***Israeli*** soldiers opened fire, killing a Palestinian farmer, Palestinian medical workers said. Ehud Olmert, the ***Israeli*** prime minister, said late on Tuesday that the killing of the man on the motorcycle was only an initial reaction and that Israel's full response was still to come, ***Israeli*** media websites reported. Egyptian mediation Despite the latest violence, Egyptian mediators are continuing efforts to persuade Israel and Hamas to negotiate a more permanent ceasefire. Hamas wants the border crossings into Gaza reopened, including the Rafah checkpoint bordering Egypt, to end the ***Israeli*** blockade in the territory. Israel wants to stop the rocket fire and prevent Hamas fighters from using smuggling tunnels under the border with Egypt to rearm themselves with weapons. "The Israelis' position is extremely tough," Phillips said. "They are determined to show that the policy of deterrence - which they believed justified the recent attacks on Gaza - worked ... It makes it a very difficult situation for Mr Mitchell."
Abu Mu'aad	880	2009-01-27	18	Palestinian militants detonated a bomb next to an ***Israeli*** army ***patrol*** along the border with Gaza on Tuesday, killing one soldier and wounding three in the first serious clash since a cease-fire went into effect more than a week ago. ***Israeli*** soldiers briefly crossed the border in search of the attackers, and Israel's defense minister, Ehud Barak, called an urgent meeting of Israel's top defense officers, saying Israel "cannot accept" the attack. "We will respond, but there is no point in elaborating," Barak said in comments released by his office. The explosion jolted the calm that has largely prevailed since Israel ended a devastating three-week offensive on Jan. 17. Since withdrawing its ***troop***s, Israel has threatened to retaliate hard for any violations of the truce. The flare-up came as Gazans struggle to resume normal life after the fighting, and as international donors discuss how best to help the territory rebuild. Gaza's Hamas leader said Tuesday the group -- which is boycotted as a terrorist organization by the U.S. and European Union -- would not try to claim any of the reconstruction funds, an announcement that appeared aimed at clearing the way for money to start flowing. The announcement from Ismail Haniyeh, who remains in hiding because of fears he could be assassinated by Israel, appeared directed at donors who concerned their funds could end up in Hamas' hands. "Our aim now is to ease the suffering of our people and to remove the aftermath of the aggression in Gaza," the statement said. "Therefore we emphasize that we are not concerned to receive the money for rebuilding Gaza and we are not seeking that." After Tuesday's bomb blast, heavy gunfire was heard along the border in central Gaza and ***Israeli*** helicopters hovered in the air firing machine gun bursts, Palestinian witnesses said. An ***Israeli*** jet set off a loud sonic boom over Gaza City not long afterward, possibly as a warning. The ***Israeli*** military said the bomb targeted an ***Israeli patrol*** near the border community of Kissufim. It

Member	Thread	Posting Date	Score	Message
				was not clear if it was planted after the cease-fire, or whether it was an older device. There was no claim of responsibility. Not long after the bombing, a 27-year-old Gaza farmer was killed by ***Israeli*** gunfire along the border several miles (kilometers) away, according to Dr. Moaiya Hassanain of Gaza's Health Ministry. Two other Palestinians were wounded. The military had no immediate comment, and it was unclear if the two incidents were related. Israel closed its crossings into Gaza to humanitarian aid traffic after briefly opening them Tuesday morning. Gaza border official Raed Fattouh said ***Israeli*** officials informed him the closure was due to the attack. Israel and Gaza militants have been holding their fire since Israel ended its offensive, which was aimed at halting rocket fire from the territory. Israel announced a unilateral cease-fire on Jan. 17, and that was followed by a similar announcement from Gaza militants. In the days immediately following the cease-fire there was shelling by ***Israeli*** gunboats and some gunfire along the border -- including the killing of two men Palestinian officials identified as farmers -- but there were no serious clashes until Tuesday. Although there was no claim of responsibility, Mushir al-Masri, a Hamas leader, said Israel was to blame for continuing to fire into Gaza. Al-Masri said his group had not agreed to a full cease-fire but only to a "lull" in fighting. "The Zionists are responsible for any aggression," he said. Egypt is currently trying to negotiate a longer-term arrangement to allow quiet in the coastal territory of 1.4 million people, which has been ruled by the Islamic militants of Hamas since June 2007. Local experts believe the fighting caused some $2 billion in damage. Israel wants an end to Hamas rocket attacks and guarantees that Hamas will be prevented from smuggling weapons into Gaza from Egypt. Hamas has demanded that Israel and Egypt reopen Gaza's border crossings, which have been largely closed since Hamas took power. The crossings are Gaza's economic lifeline. The ***Israeli*** offensive killed 1,285 Palestinians, more than half of them civilians, according to records kept by the Palestinian Center for Human Rights. Thirteen Israelis, including three civilians, were also killed during the fighting. Source: ASSOCIATED PRESS
ShabaabQoqaz	870	2009-01-26	15	Analysis shows ***insurgent***s are increasingly confronting NATO ***troop***s in open warfare, rather than relying on bombings, suicide strikes GRAEME SMITH From Monday's Globe and Mail January 26, 2009 at 2:23 AM EST Taliban fighters are increasingly hitting their targets directly instead of relying on bombs, according to a year-end statistical review that contradicts a key NATO message about the war in Afghanistan. Public statements from Canadian and other foreign ***troop***s have repeatedly emphasized the idea that the ***insurgent***s are losing momentum because they can only detonate explosives, failing to confront their opponents in combat. But an analysis of almost 13,000 violent incidents in Afghanistan in 2007 and 2008, prepared by security consultant Sami Kovanen and provided to The Globe and Mail, shows a clear trend toward open warfare. By far the most common type of incident, in Mr. Kovanen's analysis, is the so-called “complex attack,” meaning ambushes or other kinds of battle using more than one type of weapon. The analyst counted 2,555 such attacks in 2008, up 117 per cent from the previous year. Bombings also increased, but only by 63 per cent year-on-year for a total of 2,384 successful and attempted strikes in 2008. Mr. Kovanen has spent years tracking the conflict in Afghanistan, first as a NATO officer and most recently at the newly established Kabul-based consultancy Tundra Strategic Security Solutions. The latest trends are disturbing, he says, because the Taliban need more manpower to launch complex ambushes. “Clearly they are not as weak as the military claims,” Mr. Kovanen said. The numbers for Kandahar province, where Canadian ***troop***s have responsibility, do not show the same trend. In that province, the number of bombing incidents has grown more quickly – up 141 per cent – than complex attacks – up 83 per cent. Kandahar remains the most violent province in the country, however, with 1,090 incidents of all types last year, up from 697 in 2007. The ***insurgent***s appear to have refined their strategy last summer, Mr. Kovanen said, as his database started to show the Taliban using more bombs against foreign ***troop***s and saving their guerrilla fighters for strikes on easier targets such as the Afghan army and police. That may explain why the pattern of attacks was different in Kandahar, with its concentration of international forces. At the same time, Mr. Kovanen said the threat against foreign ***troop***s is growing because the Taliban's bombs are getting bigger and the ***insurgent***s have proved they can briefly – and for now, occasionally – overwhelm the international forces with ambushes such as the attack in Surobi district that killed 10 French ***troop***s last summer. Besides the trends, what stands out in Mr. Kovanen's figures is the sheer number of incidents. At the beginning of 2008, Canadian and NATO officials were confidently predicting a halt to the increases in violence, and even senior UN analysts were forecasting only a slightly worse conflict. Mr. Kovanen's final tally shows a 52-per-cent increase in violent incidents in 2008 – more than predicted, even by his own analysis. The geographical spread of the violence is also raising concern, he said, as the ***insurgent***s appear to be gaining control in new provinces such as Wardak, Ghazni, and Logar. “One big surprise was not the actual numbers of incidents, but more how the Taliban were able to gain strong influence or even complete control in so wide an area during 2008,” Mr. Kovanen said. One hopeful trend in Mr. Kovanen's statistics is the slight decrease in suicide attacks. While almost every other category of violence worsened sharply last year, the number of ***insurgent***s who detonated themselves using car bombs or suicide vests has fallen, from 135 in 2007 to 126 last year. Those figures include a decrease in Kandahar's suicide attacks, too, from 26 to 23. Among those suicide blasts were some of the worst explosions in Afghanistan's history, such as the bombing that killed more than 100 people on the outskirts of Kandahar city last February. Still, the ***insurgent***s' apparent lack of enthusiasm for suicide blasts may suggest that Arab extremists are not yet an important part of the Taliban ranks. The coming year may bring global jihadists to Afghanistan, however, as the United States shifts the focus of its military effort away from Iraq. “ Jihadists' focus is shifting from Iraq into Afghanistan,” Mr. Kovanen said, “and this year we are expecting to see more sophisticated attacks and better tactics.”
ShabaabQoqaz	869	2009-01-	13	Somalia confirms complete ***troop*** withdrawal by Ethiopia from Baidoa www.chinaview.cn 2009-01-26 16:03:55

Member	Thread	Posting Date	Score	Message
		26		Print BAIDAO, Somalia, Jan. 26 (Xinhua) -- The last Ethiopian soldiers have left their base in the southern Somali town of Baidoa, as forces of the Al-Shabaab Islamist movement threaten to attack the town, a senior Somali government official said on Monday. Baidoa is the seat of the Somali parliament and the last place in the country where the Ethiopian ***troop***s remained after two years in the war-wrecked Horn of Africa nation, "I can confirm you that there is no single Ethiopian soldier in Baidoa today (Monday) and if what the ***insurgent*** forces were fighting was the foreign forces, they have now left the town and I see no reason for further violence," Mohamed Ibrahim Habsade, a senior Somali cabinet minister told Xinhua in Baidoa, 245 km southwest of Mogadishu. He said he is aware that the ***insurgent***s, who are now surrounding Baidoa, have been threatening to enter the town since the withdrawal of Ethiopian ***troop***s in the early hours of Monday morning. He said that it is only the local people who will suffer most. Habsade revealed that there are indirect talks with the ***insurgent***s for a peaceful resolution of the standoff, saying that local elders are mediating between the two sides. The spokesman for Al-Shabaab, Sheik Muqtar Robow Abu Mansuur, said following the withdrawal of the Ethiopian ***troop***s, that his fighters, who control now most of the villages and towns around Baidoa will "peacefully take over the town" which is guarded by small number of local militia and Somali government forces. The local militias and the Somali government forces in battle wagons can be seen protecting government targets and taking positions around the town in preparation for a possible assault from the ***insurgent***s. "If the militias and the government forces in the town try to resist we will kill them all but if they do not ***fight*** us we have nothing against them," Abu Mansuur told reporters in Baidoa by phone. The move to withdraw from Baidoa came a week after the Ethiopian ***troop***s pulled out of their bases in the Somali capital Mogadishu. The withdrawal of the Ethiopian ***troop***s is part of a wide-ranging peace and power-sharing deal reached between the Somali transitional government and the main opposition coalition, the Alliance for the Re-liberation of Somalia (ARS). The two sides are now meeting in Djibouti City, the capital of the northwestern neighbor of Somalia, to work out a power-sharing arrangement stipulated in the agreement signed last year.