Detecting Radicalization Trajectories Using Graph Pattern Matching Algorithms

Benjamin W. K. Hung, Anura P. Jayasumana
Department of Electrical and Computer Engineering
Colorado State University
Fort Collins, Colorado, USA
Email: {benjamin.hung, anura.jayasumana} @colostate.edu

Vidarshana W. Bandara
CA Technologies
Plano, Texas, USA
Email: vidarshana.bandara@ca.com

***Abstract*—This paper outlines our on-going efforts to address the *radicalization detection problem*, the automated or semi-automated task of dynamically detecting and tracking behavioral changes in individuals who undergo the process of increasingly espousing jihadist beliefs and transition to the use of violent action in support of those beliefs. Leveraging the notion that *personal trajectories* towards violent radicalization exist, we take a graph pattern matching approach to track individual-level indicators using data fused from available public and government/law enforcement databases. We show that our approach provides analysts with the ability to find full or partial matches against a query pattern of radicalization, and a means to quantify the pace of the appearance of the indicators that may help prioritize investigative efforts and resources to prevent planned attacks.**

I. Introduction and Motivation

Within the last decade, we have seen a rise in the threat of radicalized homegrown violent extremists who seek to commit acts of terrorism in the United States and abroad. Facilitated and inspired by extremist organizations utilizing the internet and social media for recruitment and radicalization, these individuals have presented significant challenges to law enforcement and signify a new nature of the terrorist threat.

While researchers generally concede that there are no universal profiles for people who seek to commit such violence [8], some have posited lists of activities as likely signals. Some of these include active online participation in blogs or chat rooms, communication with extremists, consumption of jihadist videos and propaganda, suspicious foreign travel or attempted travel, and expressions of acceptance or intent to conduct violent jihad or martyrdom operations [15]. Despite the identification of these indicators, the risk of threats going undetected to law enforcement remains high for a variety of reasons, to include limited personnel to track all those who may exhibit such indicators. The FBI Director recently requested help from local law enforcement, saying "It's an extraordinarily difficult challenge task to find... and then assess those who may be on a journey from talking to doing" [16].

To aid in this important effort to prevent future terrorist incidents, we seek to leverage the notions that *personal trajectories towards violent radicalization* exist [12] [19] and that some activities are "detectable and observable" [8]. The identification of radicalizing individuals, especially those with a sudden surge in indicator activity [14] may help prioritize investigative efforts and resources to interrupt planned attacks.

II. Problem Statement

We therefore define the *radicalization detection problem* as the automated or semi-automated task of dynamically detecting and tracking behavioral changes in individuals who undergo the process of increasingly espousing jihadist beliefs and transition to the use of violent action in support of those beliefs [1] [12]. The fundamental question we are asking is, "How can technologies better assist in tracking the individual-level signals or behaviors found in large, public and government/law enforcement databases in order to detect the radicalization of or threat from homegrown violent extremists?"

III. Proposed Methodology

We are developing approaches to track individual-level behavioral indicators of homegrown violent extremism using data fused from available public and government/law enforcement databases.[1] Our intuition is that when the fused data from such sources are represented as *dynamic heterogeneous graphs*, researchers can more readily identify suspicious behavioral trajectories and on-going radicalization through graph pattern matching. We first develop a query pattern of radicalization, and then use various graph pattern matching algorithms to find and track those who increasingly exhibit those indicators.

While graph pattern matching approaches are efficient and robust in application, most rely on the certainty of specific types of connections or attributes in the query pattern. In reality for radicalization, one may be much less certain about the query structure, or the entities of interest may not exhibit all of the possible behaviors or attributes. We address these aspects in the *investigative graph search* framework [9], which is the process of searching for and prioritizing entities of interest who *may exhibit part* or all of a pattern of attributes or connections. Some distinguishing characteristics of this search from other social or entity searches include:

- Query nodes are hypothesized indicators of a latent behavior of interest; not all indicators may appear in the matched result to make a partial match worthy of further investigation.

[1]Our approach is predicated on access to and proper classification/labeling data from open and restricted sources to produced large heterogeneous data graphs. For example, we envision that social media data (e.g, text and associations on Twitter and Facebook) would be fused with firearm background check databases and local/state/federal criminal and terrorist databases (including data from the FBI's Tripwire and 'FBI Tips' programs as well as the TSA's Automated Targeting System and Secure Flight programs.)

978-1-5090-3865-7/16/$31.00 ©2016 IEEE

- While the hypothesized indicators may or may not have a known sequence of occurrence, the rate and trajectory by which someone exhibits the indicators in a query pattern is often of interest to investigators.
- Ranking full or partial match results based on both the similarity and trajectory of the match to the query helps analysts prioritize among potentially many matches.

IV. Background and Related Work

Our work is related to advances in both threat-based social media monitoring and graph pattern matching.

Threat-based social media monitoring systems have been proposed in [18] for school shooters and [2] for lone wolf terrorists. The latter outlined a semi-automated approach to detect the weak signals of lone wolf terrorists by analyzing for intent, capability, and opportunity. Their proposed methodology employed a web-crawler to find extremist forums/websites and algorithms to identify the potential actors and aliases who are active on them. From there, they proposed methods to estimate the components of the lone wolf hypothesis through tailored natural language processing techniques.

Our work differs first by its focus: rather than designing the whole system (i.e., data collection and management, entity labeling/resolution, and detection), we scope our efforts on the detection mechanism. Moreover, we focus on the radicalization trajectory of individuals over time, something that was not previously addressed. Lastly, our heterogeneous data graph-based approach enhances detection capability by capturing the important context of behaviors and associations. Previous approaches only analyze social media text and ignore the information contained in the network and other databases.

Graph pattern matching. We focus our efforts on inexact matching due its flexibility for returning results in the presence of noise or errors in the data [7]. The recent notable works include [11] and [17]; the latter most closely relates to our work because it introduced a method that supports inexact queries on both node and edge attributes as well as wildcard matches. This offers flexibility in the query construction and allows intelligence analysts to explore the unknown or uncertain connections. However, this matching method still does not truly support uncertain or partial indicator-type matches, nor tracks the match scores for entities over time.

Equally important are simulation-based matching techniques, starting with bounded graph simulation [4] to find meaningful matches given a pattern graph with arbitrary or specified path lengths in the connections, and *dual and strong simulations* [13] by preserving query graph topology through enforcement of both parent and child relationships in the match (which we adopted) and imposing locality constraints. However, despite improvements to the simulation-based approaches developed in [9], the matching process is still based on graph traversal rather than more efficient vectorized methods.

There are several dynamic graph pattern matching methods worth mentioning for comparison. Researchers in [5] developed algorithms for incremental pattern matching over a series of relatively small changes in graphs over time, while others in [3] developed an exact subgraph incremental search algorithm for continuous queries. None of these works, however, explicitly track the full or partial entity-level match results over time. Lastly, [20] proposed a time-based extension of dual bounded simulation. The data graph was enriched with timestamped edges, which is a modeling practice we adopted. The query graph was also expanded to include a strict allowable edge sequence for the matches to occur. However, such strict edge constraints are too restrictive for noisy social graphs, when connections can often appear out of an anticipated sequence. Additionally, this approach is limited to finding exact matching subgraphs, and neither returns partial matches nor tracks the trajectory of the matches to a query over time.

V. Progress and Results

The following summarizes our progress to date. In [9] we developed the *investigative simulation* graph pattern matching technique, which comprised of several necessary extensions to the existing dual simulation graph pattern matching method in order to prevent over-matching and make it appropriate for intelligence analysts and law enforcement officials. Specifically, we devised and imposed a categorical label structure on nodes consistent with the nature of indicators in investigations.

- Query focus: The subject of the investigative query– namely, the people in the data graph.
- Individually innocuous but related activity: An activity that needs to occur in conjunction with other (more suspicious) indicators to be worth further examination.
- Indicator: A behavior that suggests a person is a threat.
- Red flag indicator: An activity that is *individually sufficient* to warrant further investigation.

We also developed pruning and match completion rules based upon these node labels on the search results to ensure sensibility and usefulness of partial matches to analysts. Lastly, we introduced a top-*k* ranking scheme that can help analysts prioritize investigative efforts. We demonstrated the performance of investigative simulation on a static real-world large dataset by successfully producing sensible matching results.

In [10] we focused on the determining the pattern matching trajectories of the entities of interest that may exhibit part or all of a pattern of attributes or connections for a latent behavior. Our methods were general enough for a variety applications, but we highlight here its use for detecting radicalization. Our approach provides analysts not only with the ability to find full or partial matches against a query pattern of radicalization, but also a means to quantify the pace of the appearance of the indicators. We proposed and implemented a vectorized graph pattern matching approach that calculates the multi-hop class similarities between nodes in the query and data graphs over time. By tracking partial match trajectories, we provide another dimension of analysis in investigative graph searches to highlight entities on a pathway towards a pattern of radicalization. Figure 1 depicts a small investigative graph search problem and the results of our approach. As expected through visual inspection of the original data graph G, Node (Person) 9 after 4 time steps exhibited all the radicalization

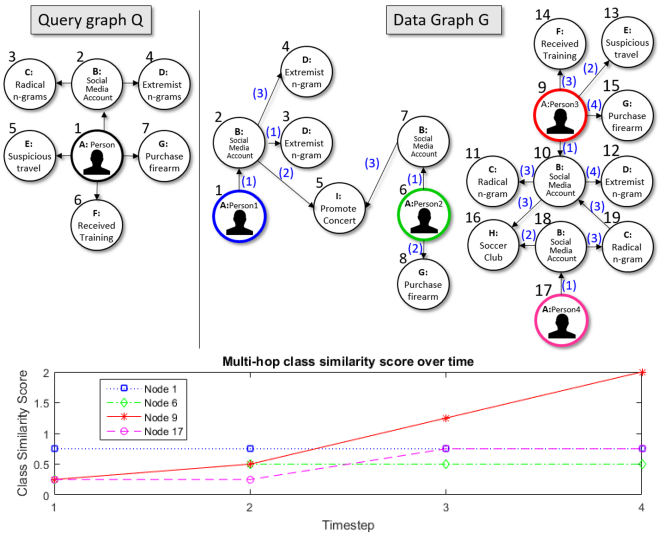

Fig. 1: An small example problem of investigative search for homegrown violent extremists. *Top left*: Query graph Q of some possible indicators of a radicalizing extremist. *Top right*: Fictitious data graph G of 4 people with indicators of on- and off-line activities. The node class labels are inside the node, and the node IDs are outside the node. Each edge has a time-stamp (in blue) that denotes when the edge was formed. Bottom: Plot of the multi-hop class similarity scores over time. For each of the timesteps, we show the changes in class similarity over 3 hops of each person of interest node.

indicators in query graph Q. Additionally, our approach returned the multi-hop class membership dynamics and shows the trajectory of each of the 4 persons towards exhibiting the indicators of homegrown violent extremism over time. This analysis could be useful for investigators who were concerned with the sudden surge in indicator activity by certain people [14] in an effort to interrupt any planned acts of targeted violence. We showed that our approach was scalable to a large proxy dataset of 470K nodes and 4 million edges after some pre-processing steps.

VI. Outstanding Issues

A key area of future focus is an incremental graph pattern matching approach like those found in [3] and [5] in order to dynamically and more efficiently update the multi-hop class similarity scores for new data in the heterogeneous graph in the form of added or deleted edges or nodes. We also intend to explore how to implement our technique on distributed systems.

A more important outstanding issue is the need to validate our approach with real, labeled data. All our large experiments thus far have been on the BlogCatalog dataset [21], which we tailored as a proxy for data on radicalization indicators. We are seeking data that contain time-based, labeled indicators of real cases of radicalization that both did and did not ultimately lead to violent activity.

VII. Conclusion

Our on-going research on dynamic graph-based pattern matching approaches to detecting radicalization trajectories of extremists is an important component in the effort to prevent future attacks. We show that our approach provides analysts with the ability to find full or partial matches against a query pattern of radicalization, and a means to quantify the pace of the indicators that may help prioritize investigative efforts.

References

[1] J. Bjelopera, "American Jihadist Terrorism: Combating a Complex Threat," *Congressional Research Service*, January 23, 2013.

[2] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Martenson, and P. Svenson, "Harvesting and analysis of weak signals for detecting lone wolf terrorists," *Security Informatics*, volume 2, pp 11-26, 2013.

[3] S. Choudhury, L. Holder, J. Feo, and G. Chin, "Fast Search for Dynamic Multi-Relational Graphs," *DyNetMM 2013*, Association for Computing Machinery, June 2013.

[4] W. Fan et al, "Graph pattern matching: From intractable to polynomial time," *Proceedings of the VLDB Endowment*, Vol 3, No. 1, 2010.

[5] W. Fan et al, "Incremental graph pattern matching," *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, p. 925-936, 2011.

[6] W. Fan, "Diversified Top-k graph pattern matching," *Proceedings of the VLDB Endowment*, Vol 6, No. 13, 2013.

[7] B. Gallagher, "Matching structure and semantics: A survey on graph-based pattern matching," *Journal of the American Association for Artificial Intelligence*, Fall Symposium Technical Report, January 2006.

[8] P. Gill, J. Horgan, P. Deckert, "Bombing Alone: Tracing the Motivations and Antecedent Behaviors of Lone-Actor Terrorists," *Journal of Forensic Science*, Vol. 59, No. 2, March 2014.

[9] B. Hung and A. Jayasumana, "Investigative Simulation: Towards Utilizing Graph Pattern Matching for Investigative Search," to appear in *Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI) 2016 Conference Proceedings*, preprint available online at http://arxiv.org/abs/1608.01760.

[10] B. Hung, A. Jayasumana, and V. Bandara, "Pattern Matching Trajectories for Investigative Graph Searches," to appear in *IEEE Data Science and Advanced Analytics (DSAA) 2016 Conference Proceedings*, 2016.

[11] A. Khan, Y. Wu, C. Aggarwal, and X. Yan, "NeMa: Fast Graph Search with Label Similarity," *Proceedings of the VLDB Endowment*, Volume 6, Issue 3, January 2013.

[12] J. Klausen, S. Campion, N. Needle, G. Nguyen, and R. Libretti, "Toward a Behavioral Model of 'Homegrown' Radicalization Trajectories", *Studies in Conflict and Terrorism*, 39:1, 67-83, 2015.

[13] S. Ma, Y. Cao, W. Fan, J. Huai, and T. Wo, "Strong Simulation: Capturing Topology in Graph Pattern Matching," *Proceedings of the VLDB Endowment*, Vol 5, No. 4, 2012.

[14] J. Meloy, J. Hoffmann, A. Guldimann, and D. James, "The Role of Warning Behaviors in Threat Assessment: An Exploration and Suggested Typology," *Behavioral Sciences and the Law*, Vol 30, p. 256–279, 2011

[15] National Counterterrorism Center, "Behavioral Indicators Offer Insights for Spotting Extremists Mobilizing for Violence," 2011.

[16] E. Perez and S. Prokupecz, "FBI struggling with surge in homegrown terror cases," CNN, May 30, 2015.

[17] R. Pienta, A. Tamersoy, H. Tong, and D. Chau, "MAGE: Matching Approximate Patterns in Richly-Attributed Graphs," *Proceedings of the IEEE Conference on Big Data*, October 2014.

[18] A. Semenov, J. Veijalainen, and A. Boukhanovsky, "A Generic Architecture for a Social Network Monitoring and Analysis System," *IEEE International Conference on Network-Based Information Systems*, 2011.

[19] M. Silber, and A. Bhatt, "Radicalization in the West: The Homegrown Threat," New York Police Department Intelligence Division, 2007.

[20] C. Song, T. Ge, C. Chen, and J. Wang, "Event Pattern Matching over Graph Streams," *Proceedings of the VLDB Endowment*, Vol. 8, No. 4, 2014.

[21] R. Zafarani and H. Liu. Social Computing Data Repository at ASU [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering, 2009.