International Journal of Cyber-Security and Digital Forensics (IJCSDF) 6(3): 121-125
The Society of Digital Information and Wireless Communications (SDIWC), 2017 ISSN: 2305-0012

Learning Maliciousness in Cybersecurity Graphs

Connor Walsh
Uplevel Security
New York, NY 10001
connor@uplevelsecurity.com

Akshay Rangamani
ECE Department
Johns Hopkins University
rangamani.akshay@jhu.edu

Sam Gottlieb
Uplevel Security
New York, NY 10001
samantha@uplevelsecurity.com

Liz Maida
Uplevel Security
New York, NY 10001
liz@uplevelsecurity.com

ABSTRACT

Statistical relational learning is concerned with inferring patterns from data explicitly modeled as graphs. In this work, we present an approach to learning latent topological and attribute features of multi-relational property graphs in settings where a fraction of node attributes are missing. This work draws upon prior work based on tensor factorization. We demonstrate how learned latent embeddings can be used to approximate the missing attributes. The methods explored are applied to the problem of detecting malicious entities in a novel cybersecurity ontology in which emails are explicitly modeled as graphs.

KEYWORDS

Cybersecurity, data visualization, graph theory, malicious code, malware, machine learning

1 INTRODUCTION

While traditional machine learning provides ways of extracting invariant patterns from sets of statistically variant examples, the semantics describing how such examples are related to one another are only implicitly exposed by the distributions from which they are drawn. Yet, in a broad spectrum of settings—statistical mechanics [19, 2], biology [8, 17, 20], social science [14, 10, 21]—entities are inextricably relational by nature, fitting into an ontological paradigm intuitively composed of nodes and edges. In these cases, failures to consider the rich relational structure of data by our learning algorithms severely constrain the depth and completeness of concepts we can capture. Statistical relational learning, conversely, offers a class of methods for inferring patterns from data which are modeled explicitly as graphs [15]. In particular, relational learning has been used for a variety of learning tasks including link prediction, entity resolution, and entity attribute prediction.

Prior work has been devoted largely to the problem of link prediction [3, 4, 6, 22], where new relationships are learned given those that exist. Comparatively, however, there has been a lack of focus within the relational learning literature on the task of predicting unknown entity attributes. More concretely, consider a K-relational property graph, $G = (V, E)$, where each element of the vertex set, $v_i \in V$, maps to a set of attributes, D_i, and

$$E = \cup_{i=1}^{K} E_i \text{ s.t. } E_i \cap E_j = \emptyset \text{if } i \neq j. \quad (1)$$

The task is to compute an approximate attribute set, $\hat{D}_i$, for the i^{th} node if it is unknown. In

essence, this problem can be characterized as inferring the attributes of a node from a combination of topological features and known attributes of neighboring nodes. Previous work has explored the similar problem of jointly learning attributes and topology, though for the task of link prediction [16]. Building upon this prior work, we use a latent feature model based on tensor factorization to learn latent relationship, node, and attribute embeddings with which to approximate unknown/missing attributes. In particular, we apply our methods to the domain of cybersecurity for predicting the malicious components of phishing emails.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 RELATIONAL LEARNING IN CYBERSECURITY

As cybersecurity attacks have become increasingly common, the focus in enterprise security circles has been on attack detection, deploying firewalls, malware detection software, and other solutions that issue alerts in response to suspicious activity. Incident responders, tasked with investigating and remediating these alerts, are typically inundated by alerts and must comb through them laboriously in order to ascertain the truly malicious events. Unsurprisingly, post-hoc analysis of security breaches often reveals missed alerts in the early stages of the attacks [18]. Moreover, since alerts are investigated and retained as independent, isolated occurrences, it is difficult for incident responders to discover associations and systematize findings.

In contrast, our approach centers on preserving the context of events by decomposing their constituent observable data into a novel graph ontology of cyber artifacts. In this work, we define emails as graphs composed of typed nodes (e.g. email address, subject, body, domain name) and the typed relationships (e.g. from, to, cc, bcc, link) connecting the observable nodes.

3 ALGORITHM

As described in the previous section, our data model consists of N cybersecurity observables in a graph, connected via K different kinds of relationships. This sort of multi-relational data is modeled by an $N \times N \times K$ adjacency tensor $\underline{\mathbf{X}}$. This means $\underline{\mathbf{X}}_{ijk} = 1$ if observables i and j are connected by relationship k. The frontal slice $\mathbf{X}_k$ describes the adjacency matrix for relationship k. The problem of learning representations for entities in such graphs has been explored previously in approaches like RESCAL [16] and TransE [4]. While Nickel et. al. in [16] presented an extension of RESCAL to include attributes for the entities in the graph, the relational learning applications described were restricted to prediction of triples and collective learning. In the RESCAL framework, the tensor is factorized as $\mathbf{X}_k \approx \mathbf{A}\mathbf{R}_k\mathbf{A}^{\top}$, where $\mathbf{A}$ is an $N \times r$ matrix that contains the latent embeddings of the entities while $\mathbf{R}_k$ models the interactions between the entities for relationship k.

In this paper, we tackle the problem of using relational learning to predict the *maliciousness* of observables in the graph, modeling *maliciousness* as an attribute of the observables. Typically, we know the *maliciousness* of a fraction of the nodes in the graph, and need to predict it for the rest of the observables.

We modify the RESCAL setup to include missing attributes in our data. If $\mathbf{D} \in \mathrm{R}^{N \times D}$ is the matrix of D dimensional attributes, and $\mathbf{M} \in \{0,1\}^{N \times D}$ is a mask indicating which attributes are known, then we can formulate the task of learning unknown entity attributes as

$$\underset{\mathbf{A},\mathbf{R}_k,\mathbf{V}}{\operatorname{argmin}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{A}\mathbf{R}_k\mathbf{A}^{\top}\|_F^2 + \|\mathbf{M} \odot (\mathbf{D} - \mathbf{A}\mathbf{V})\|_F^2 + \lambda_A\|\mathbf{A}\|_F^2 + \lambda_R \sum_{k=1}^{K} \|\mathbf{R}_k\|_F^2 + \lambda_V \|\mathbf{V}\|_F^2 \quad (2)$$

Here $\mathbf{A}$ and $\mathbf{R}_k$ are as described above, while $\mathbf{V}$ is a mapping from the latent embeddings to the attributes. To solve this problem, we use the familiar alternating least squares method, with appropriate adjustments to the RESCAL update steps to account for missing values. After obtaining $\mathbf{A}$, $\mathbf{V}$, we can predict the unknown

attributes by simply computing the matrix product $\hat{\mathbf{D}} = \mathbf{AV}$.

4 EXPERIMENTS

Using a modified formulation of RESCAL, as defined above, we learn latent topological and feature embeddings which are used to predict missing node attributes in cybersecurity observable graphs. While previous studies have applied relational learning to knowledge graphs (KGs) such as *Freebase* [22], *WordNet* [4], *DBpedia* [9], *YAGO* [16], our approach to modeling cybersecurity data relies on a novel ontology for which no canonical data sets are readily available. Therefore, we take two well-known corpora of emails, one of legitimate emails [7] and the other phishing [13], and transform them into cybersecurity knowledge graphs fitting our ontology. Furthermore, since the legitimate and phishing subgraphs are disjoint, we explore the effects of introducing synthetic connectivity between both email sets.

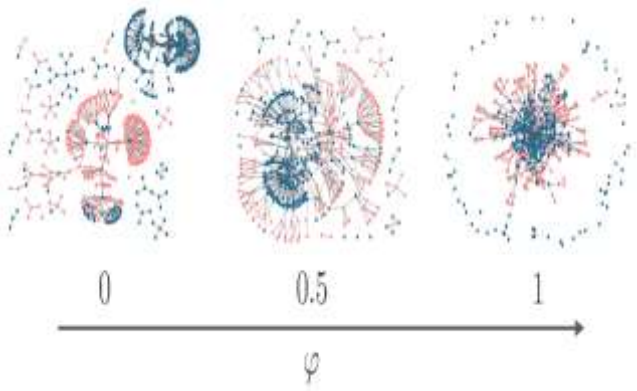


Figure 1. Visualizing the effects of introducing connectivity between the phishing and legitimate email corpora. Red and blue nodes correspond to data characterized as malicious and benign respectively. The recipient mixing factor, ϕ, is varied from 0 to 1 demonstrating a transition from topologically partitioned to enmeshed maliciousness classes.

4.1 Data sets

Enron Email Corpus Initially published by the Federal Energy Regulatory Commission during legal investigations of the Enron corporation, it has since been widely used in natural language processing [7] and social network research [5]. It is among the largest corpora of human generated email messages available, consisting of ~ 0.5M messages and around 150 users.

Nazario Phishing Email Corpus This phishing email corpus is a public collection of 4558 messages [13] and has been used as a training set in several phishing email classification systems [12, 1]. Notably, these emails have been cleaned of any sensitive information concerning the recipients. Consequently, the original recipient email addresses have been modified to resemble user@domain.com. Hence many emails within this corpus will be topologically connected despite the fact that this may not have been the case prior to cleaning.

4.2 Topological synthesis

As previously described, the data set used in our experiments is a composite of two email corpora, each distinct in their spatial-temporal origination. Consequently, once translated to fit our ontology, the constituent observables (e.g. senders, recipients, subject lines, embedded domains, etc.) generated for each set of emails are unlikely to overlap. However, in the real world, phishing emails do not occur in isolation, but within the context of a body of legitimate messages. For this reason, we investigated the effects of introducing synthetic connections between the two original data sets. We incorporate the *recipient mixing factor*, ϕ,

$$\varphi = \frac{\sum_{r \in R_n} \delta_r}{|R_p|}$$
$$\delta_r = \begin{cases} 1 & : r \in R_\ell \\ 0 & : r \notin R_\ell \end{cases} \quad (3)$$

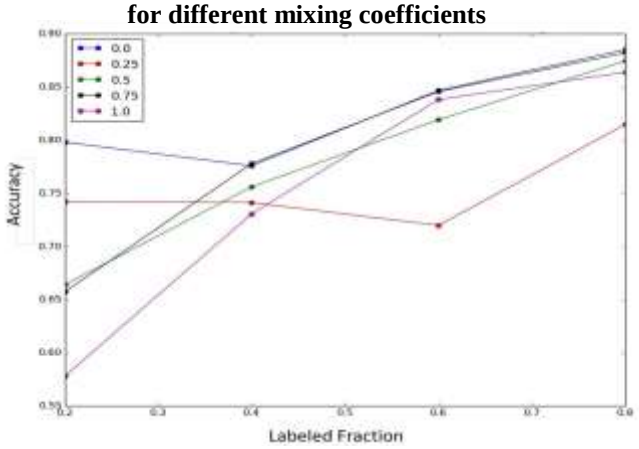
where R_p and R_ℓ are the sets of recipients linked to emails from the phishing and legitimate corpora respectively. This measure tracks the percent of recipients of phishing emails which have been replaced with recipients of legitimate emails. This approach to synthesis explicitly assumes that email recipients in both corpora are non-malicious and that re-assigning recipients introduces negligible effects on the plausibility of how the data is structured. We believe this approach to combining two distinct data sets introduces fewer limiting assumptions, as compared to generating synthetic data using Kronecker graphs [11] for example.

In our experiments, we varied ϕ from 0 to 1 (Fig. 1), where ϕ = 0 is the trivial case in which phishing/legitimate emails are disjoint subgraphs and ϕ = 1 is the case which simulates a fully integrated email corpus where users are receiving both legitimate and phishing email. We hypothesized that increasing ϕ is equivalent to increasing the difficulty of the task at hand since the learning latent features of a node would rely more on the attributes of its neighbors as opposed to just detecting topological communities.

(a) Accuracy vs fraction of labeled nodes

(b) Precision-Recall curves for graphs with ϕ =0,0.75

Modified-RESCAL accuracy vs fraction of labeled nodes for different mixing coefficients



Modified-RESCAL precision-recall curve for different mixing coefficients

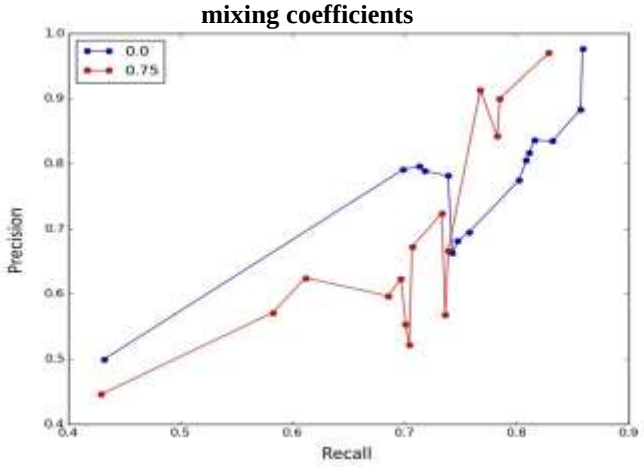


Figure 2. Results from experiments run using RESCAL to predict unknown attributes features of a node would rely more on the attributes of its neighbors as opposed to just detecting topological communities.

4.3 Setup

In our experiments, we randomly set a certain fraction of the nodes in the graphs (generated as described above) as having known *maliciousness* $\in \{+1,-1\}$. +1 indicates that a node is malicious, and −1 indicates that it is benign. Using our modified RESCAL model we try to predict the maliciousness of the rest of the nodes in the graph. We repeat the experiment with different missing fractions and report the accuracy on the test set. The accuracy is obtained by averaging the accuracy over 5 replicate sets. In addition, we used graphs of three different sizes, 622 nodes & 697 edges (small), 2825 nodes & 3961 edges (medium), and 22812 nodes & 45092 edges (large). All graphs have 14 different kinds of relationships. The results we report are for the medium size graph. We selected the best performing hyper-parameters (λ_V and, a binary threshold) with a small validation set.

4.3.1 Results

The accuracy results of our experiments are shown in Figure 2a. We see that our prediction accuracy improves as the fraction of labeled instances increases. When the *maliciousness* of 80% of the observables is known, we can predict the labels of the remaining observables with nearly 85% accuracy. We can also see that as ϕ increases, the performance at lower labeled fractions drops to around 65%. This is because there is not as much topological separation between the clusters of malicious and benign nodes in the graph at higher values of ϕ. We also show some precision-recall curves in Figure 2b. These were generated by sweeping the labeled fraction from 0.2% to 90%.

5 DISCUSSION

In this paper, we presented an approach to learning latent topological and attribute embeddings in the presence of unknown attribute information. We've demonstrated an application of this method to the problem of learning the maliciousness of cyber artifacts within a novel graph data model. Our results show relatively high accuracy for settings in which many node attributes are known.

Additionally, we've explored how increasing the connectivity within the graph appears to lead to lower accuracy, possibly because higher connectivity may increase the difficulty of the task. Since our current approach treats all entities in a graph as static, in future work we aim to introduce dynamics.

REFERENCES

1. Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014. 10.1155/2014/425731.
2. Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. Rev. Mod. Phys. Vol. 74, pp. 47–97, Jan 2002.
3. Antoine Bordes, Xavier Glorot, Jason Weston, Yoshua Bengio, Antoine Bordes, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2):233–259, 2013.
4. Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pp. 2787–2795, 2013.
5. Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. Communication networks from the Enron email corpus "it's always about the people. Enron is no different". Computational & Mathematical Organization Theory, 11(3):201–228, 2005.
6. Rodolphe Jenatton, Nicolas Le Roux, Antoine Bordes, and Guillaume Obozinski. A latent factor model for highly multi-relational data. pp. 1–9.
7. Bryan Klimt and Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research, pp. 217–226. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
8. Nevan Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron Tikuisis, Thanuja Punna, José P. Alvarez,
   Michael Shales, Xin Zhang, Michael Davey, Mark Robinson, Alberto Paccanaro, James Bray, Anthony Sheung, Bryan Beattie, Dawn Richards, Veronica Canadien, Atanas Lalev, Frank Mena, Peter Wong, Andrei Starostine, Myra Canete, James Vlasblom, Samuel Wu, Chris Orsi, Sean Collins, Shamanta Chandran, Robin Haw, Jennifer Rilstone, Kiran Gandi, Natalie Thompson, Gabe Musso, Peter S. Onge, Shaun Ghanny, Mandy Lam, Gareth Butland, Amin A. Ul, Shigehiko Kanaya, Ali Shilatifard, Erin O'Shea, Jonathan Weissman, C. Ingles, Timothy Hughes, John Parkinson, Mark Gerstein, Shoshana Wodak, Andrew Emili, and Jack Greenblatt. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440:637–643, March 2006.
9. Denis Krompaß, Stephan Baier, and Volker Tresp. Type-Constrained Representation Learning in Knowledge Graphs. 2015 (http://arxiv.org/abs/1508.02593).
10. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. Science, Vol. 323, pp. 721–723, 2009.
11. Jure Leskovec and Christos Faloutsos. Scalable modeling of real graphs using kronecker multiplication, Proceedings of the 24th International Conference on Machine Learning, ICML '07, pp. 497–504, New York, NY, USA, 2007. ACM.
12. Gaston L'Huillier, Richard Weber, and Nicolas Figueroa. Online phishing classification using adversarial data mining and signaling games, Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, CSI-KDD '09, pp. 33–42, New York, NY, USA, 2009. ACM.
13. Jose Nazario. Phishing corpus. 2004-2007 (http://monkey.org/∼jose/phishing/).
14. M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, Vol. 98, pp. 404–409, 2001.
15. Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs, Proceedings of the IEEE, Vol. 104, pp. 11–33, 2016.
16. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing YAGO, Proceedings of the 21st International Conference on World Wide Web, Nickel, M.:271—-280, 2012 (http://dl.acm.org/citation.cfm?doid=2187836.2187874).
17. David Papo, Javier M. Buldú, Stefano Boccaletti, and Edward T. Bullmore. Complex network theory and the brain. Philosophical Transactions of the Royal Society of London B: Biological Sciences, Vol. 369(1653), 2014.
18. Teri Radichel. Case study: Critical controls that could have prevented target breach. SANS Institute, 2014 (https://www.sans.org/reading-room/whitepapers/casestudies/case-study-critical-controls-prevented-target-breach-35412).
19. Herbert A. Simon. On a class of skew distribution functions. Biometrika, Vol. 42(3-4), pp.425–440, 1955.
20. Guopeng Wei, Chieh Lo, Connor Walsh, N. Luisa Hiller, and Radu Marculescu. In silico evaluation of the impacts of quorum sensing inhibition (qsi) on strain competition and development of qsi resistance. Scientific Reports, Vol. 6, p 35136, month 10, 2016.
21. Wayne W Zachary. An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research, Vol. 33(4), pp. 452–473, 1977.
22. Yu Zhao, Sheng Gao, Patrick Gallinari, and Jun Guo. Knowledge base completion by learning pairwise-interaction differentiated embeddings. Data Mining and Knowledge Discovery, Vol. 29(5), pp. 1486–1504, 2015.