Application of Schneiderman's mantra to identify malicious behavior in network graphs

Nicholas Spyrison* Miji Kim† Ha Nam Anh Pham‡

Monash University, Faculty of Information Technology

ABSTRACT

VAST Challenge is an annual contest that provides a series of challenges. Teams are prompted with hypothetical scenarios and synthetic data set to address. The 2020 Mini-Challenge involves comparing and contrasting large, complex, multi-modal network datasets and comparing and contrasting template with suspect sub-networks. Our approach follows Shneiderman's mantra; namely 1) overview, 2) zoom-in, 3) details on demand. We capture an overview of the data with heatmap and network visualization across the different data channels. We zoom in by exploring the network visualization and t-SNE embedding of the candidate subgroups. Lastly, we produce highly customized animations across time to try to follow a narrative lead.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Information visualization Human-centered computing—Visualization—Visualization application domains—Visual analytics Human-centered computing—Visualization—Visualization techniques—Graph drawings

1 INTRODUCTION

The IEEE Visual Analytics Science and Technology (VAST) is a competition held annually, with the goal is to contribute to the data analytics field through competition, in addition to helping researchers understand how they can use analytics tools through a series of challenges, as well as a benchmark to demonstrate capabilities of the tools used.

The aim of the VAST Challenge in 2020 focuses on a hypothetical cyber-attack that caused wide-spread internet outages. The first mini-challenge provides a template network domain experts have verified contain action indicative of the malicious behavior we are looking for. Five suspect networks are also included to compare and contrast with the template. Together these 6 networks are relatively small selections of the full set of network interactions provided in the challenge for optional consumption.

2 APPROACH

Our approach follows Shneiderman's mantra for extracting information via data visualization (Shneiderman, 1996 [5]); namely, the structure of our paper and workflow followed 1) overview, 2) zoom-in, 3) details on demand.

3 APPLICATION

3.1 Heatmap

To get a higher level picture of the distribution of observations across the data sources and transaction channels. From figure 1 we see that the distribution between sources is consistent. The demographic channel has the bulk of the observations while the co-authorship channel is not always used. We conclude that the suspect networks are of similar distribution to the template network and have a better handle on which channels we are visualizing.

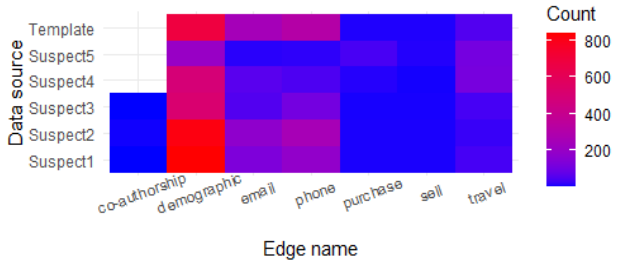

Figure 1: Count of observations within each data subset across the edge types. The bulk of the observations happens in the demographic channel. Co-authorship not used in all networks. Distributions are similar between data sources.

3.2 Network

To identify and contrast specific features across different networks, network layouts were designed. As seen from figure 2, faceting techniques were applied within data sources and node types to better compare and contract each subset network. It is noticeable that compared to the suspects, the template network has a massive number of transactions moving toward the inside of the densest region. On the other hand, Suspects 4 and 5 are having more outgoing data instead of incoming data. It is concluded that Suspects 1, 2, and 3 exhibits more similarities in the flow of transactions with the template network.

3.3 tSNE

T-distributed stochastic neighbor embedding (tSNE) is a non-linear dimensionality reduction technique purposed by van der Maaten, 2008 [6]. The application of nearest neighbors preserves local structure in the full dimensionality while distorting the global structure to fit into an arbitrary 2D embedding. We perform Barnes-Hut tSNE on each of the suspects against the template networks with the same hyperparameters; namely, PCA initialization, 500 iterations, perplexity $= 1/3 * \sqrt{\#\ observations}$, theta = .5.

3.4 Animating across time

To further examine the network and identify suspects which most closely resemble a template profile, animations across the time were produced based on the procurement transactions. The name of each node was changed to the letter for better readability, followed by extra letters to indicate whether its edge type was sell or purchase and whether it was the template or the suspect.

The animated bar chart (figure 4) aggregates weight over time and shows the breakdown of the weight per node in descending order. It was designed to be a diverging bar chart, with bars that race to the

*e-mail: nicholas.spyrison@monash.edu
†e-mail: mkim0021@student.monash.edu
‡e-mail: andrew.pham@monash.edu

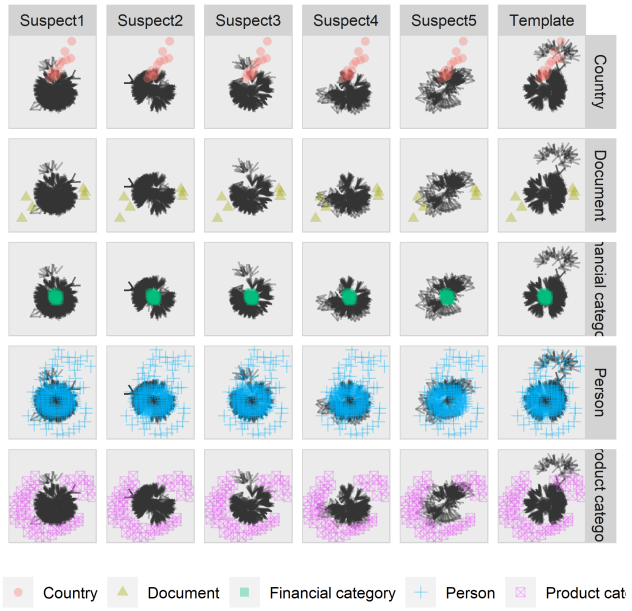

Figure 2: Network graph of each subset network across node type. This layout was generated from the 'largely' layout of the {igraph} R package. Click the image for a larger variant on our GitHub repository. The template network has the bulk of transactions happening in the densest region of the network pointing inward while having some outward-facing transactions going to nodes in low-density regions. Suspects 4 and 5 are comprised mostly of outwards facing transactions they look quite different from the template network. Suspects 1, 2, and 3 have mostly inward trending transactions. They also exhibit a smaller fraction of outward transaction except for suspect 2.

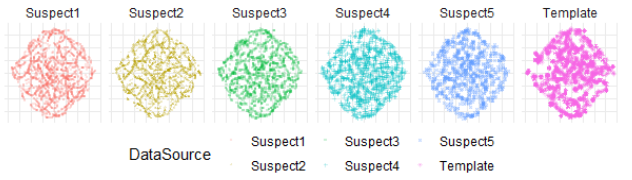

Figure 3: tSNE non-linear embedding of each subset network. Template network has medium length strings and clusters of points. No suspects have such clustering of points. Suspects 1, 2, and 3 have suitable stringiness, while suspects 4 and 5 strings are short and choppy. Click on the image for a larger version.

top based on ranks. As per bar, dark green represents last weight, while light green represents incremental weight.

The scatter plot (figure 5) shows a cumulative weight for the template and the suspect with a timeline element. It was designed to identify and visualize similarities between each suspect and the template profile shown over time. Faceting was used to make each suspect and the template easily distinguished. In the same context, each data point has a different shape and color according to its data source. The grey bar at the bottom illustrates the progress of time.

4 Conclusion

In VAST Challenge 2020 we are asked to compare and contrast 5 suspect networks with a template network. This template network has been verified by domain experts to contain true malicious behavior of the type we are trying to identify in such a cyber-attack. Our approach applies Shneiderman's mantra.

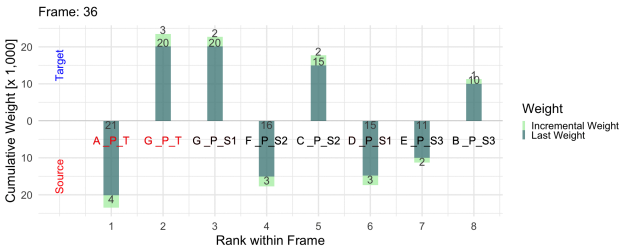

Figure 4: The final frame of the cumulative and incremental summation of procurement weights of the target, suspect-1, -2, and -3 networks. Each bar is a node within the top 10 sum of procurement weights within the given frame. Click on the figure to view the full .gif animation across time.

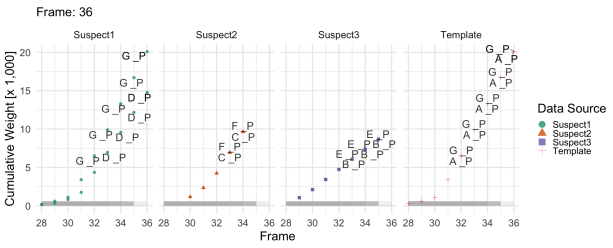

Figure 5: NEED TO REVIEW. Procurement scatter plot demonstrates a cumulative weight over time showing the label of the node when its cumulative weight is above 6.

First, we take a higher-level overview in figure 1, we find the distributions of the data sources are similar enough for like-comparisons. Zooming-in with figures 2 and 3, both the network visualization and tSNE embedding of the different networks corroborate that suspects 4 and 5 do not behave like the template network. We remove these candidates from our search. Figure 2 further shows that suspect 2 does not have the fraction of transactions coming outward from the dense center that is found in suspects 1, 3, and the template network. While figure 3 shows that clustered behavior of the template network is absent from the embedding for suspects 1, 2, and 3. To allow for details on demand we produced bar charts and scatter plots animated across time to take a closer look at the temporal behavior of individual channels or networks.

We recommend conferring with domain experts on a couple of fronts. First, to validate if the outward trend absent from suspect 2 is a necessary feature of such malicious behavior. If so this would preclude suspect 2 from the search. Secondly, we want to identify if the clustering behavior identified uniquely in the template network is necessary for target behavior. If so this lends evidence that while suspects 1 and 3 are most like the template, they too lack a necessary feature of the malicious behavior in question.

5 Software

The work for this paper was performed in R [4] using the packages {dplyr} [8], {gganimate} [3], {ggplot2} [7], {ggraph} [2], {Rtsne} [1], {tidyr} [9]. For larger variants of any of the figures, click on that figure to bring up a larger remote version. All code, figures and their variants can be found on our GitHub repository github.com/nspyrison/VAST_Challenge_2020/.

References

[1] J. H. Krijthe. *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*.

[2] T. L. Pedersen. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*.

[3] T. L. Pedersen and D. Robinson. *gganimate: A Grammar of Animated Graphics*.
[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
[5] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pp. 336–343. IEEE.
[6] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. 9:2579–2605.
[7] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
[8] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*.
[9] H. Wickham and L. Henry. *tidyr: Tidy Messy Data*.