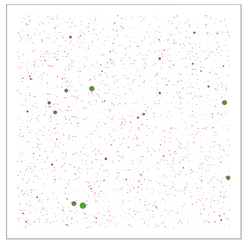
Applying Big Data Techniques for Rapid Analysis on Data Leaks

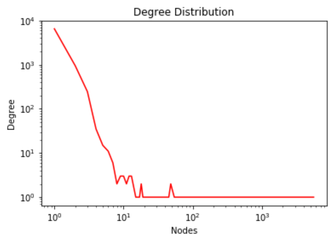
Background

Corporate and government organizations are large targets for malicious cybersecurity actors. Through the identification of an unprotected networked device (ie email server) an adversary could potentially have the capability of gain access to confidential information to subsequently leak to the public, rival state actors, or business competitors. Network scientists routinely work with large datasets and have developed techniques for gaining insights into a group of information that is structured and complex. This work applies tools such as community detection and PageRank calculations to gain insights into a large structured network. For example, analyzing the connectivity of nodes within a graph to identify nodes with similar properties and classify them as communities. Also, determining nodes with high degrees and resulting PageRanks to classify nodes with possible higher influence in the graph. The intent is to gain insight into a structured network to be able to develop a course of action following a data leak. For example, determine what high-level personnel were impacted or targeted.

Dataset

The raw dataset was retrieved from kaggle.com. This experiment used three of the available four published documents of the widely publicized Hillary Clinton Wikileaks email leak. The dataset includes emails by IDs, email aliases by ID, and a list of recipients for each email. The fourth committed document contains the content of each email which I considered outside of the scope of this work.



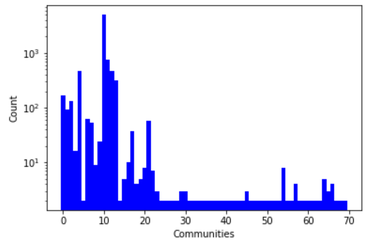


Methodology

This work leverages the NetworkX Python package to generate the graph data structures from the dataset. NetworkX is also used to calculate PageRanks. The Community package was used to calculate community data for the graph. This experiment was conducted on an IPython notebook running on the UCSD Jupyterhub website on a server instance that is allotted 8GB of RAM.

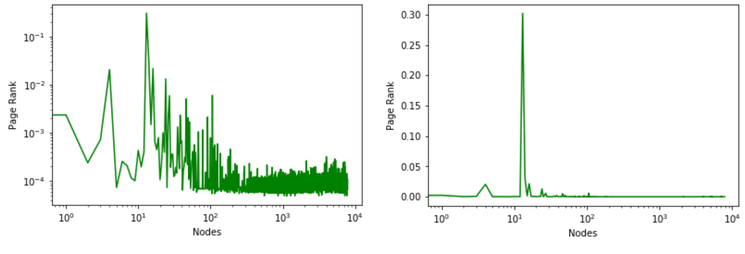
Louvain Algorithm for Community Detection

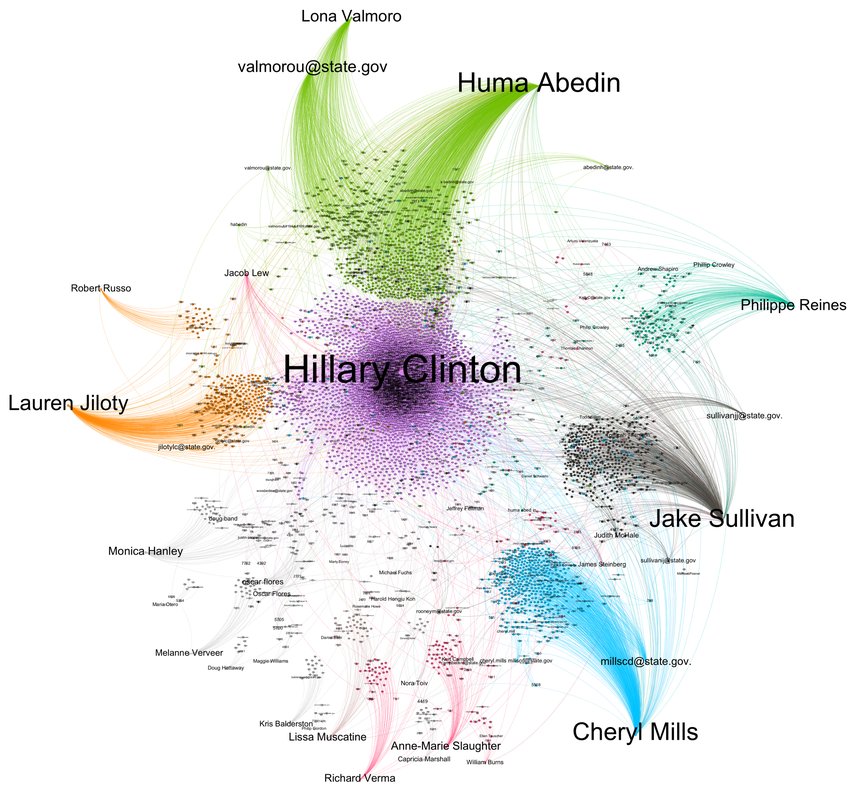
The Louvain algorithm generates several communities within the network graph. The histogram below shows that just under 70 communities were detected overall. However, the nodes are note evenly distributed through the generated communities. The majority of the nodes belong to only a handful of communities. Since the graph contains both aliases and emails as nodes, I attribute the sparsely populated communities to having the emails included as nodes instead of just edges.



PageRank Algorithm to ID important nodes

The figure below demonstrates the PageRanks for all the nodes in the graph. It is apparent that only a portion of the nodes have a significant PageRank in respect to the overall number of nodes in the graph. I attribute this result to having emails as nodes within the graph. Just as in the community detection results. The large number of email nodes skews the results of the network calculations causing me to use logarithmic scales to account for this effect. However, omitting the logarithmic scales for the PageRank values we can discover that there are only a small amount of nodes with significantly large ranks in comparison to the entire graph.





Gephi Visualization

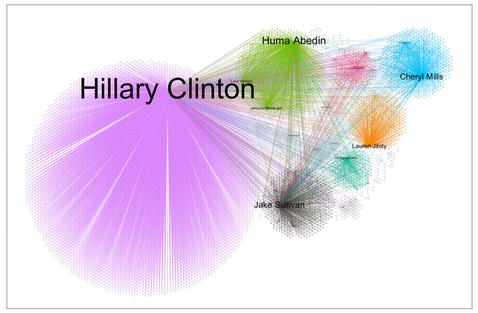
The visualization displayed above applies a combination of filters based on various calculations made from the dataset. While the node size is constant for all the nodes within the graph. The label for each node is resized to correspond to the calculated PageRank of that node. Therefore, the nodes with the higher PageRank obtain a larger label size. Additionally, the node color corresponds to an assigned community ID. The nodes within the same community obtain the same node color. This results in the apparent division within the graph of nodes by color. The layout used for the graph is a combination of Yifan Hu and ForceAtlas 2.

The table below depicts the Top 20 nodes in the graph by PageRank.

	Alias	Comm. ID	PageRank
1	Hillary Clinton	8	0.303201
2	Huma Abedin	10	0.02963
3	Jake Sullivan	5	0.021565
4	Cheryl Mills	9	0.020492
5	Lauren Jiloty	1	0.01315
6	valmorou@state.gov	10	0.006025
7	Lona Valmoro	10	0.005915
8	Philippe Reines	4	0.005096
9	Anne-Marie Slaughter	6	0.002353
10	Monica Hanley	11	0.002328
11	Lissa Muscatine	15	0.002137
12	millscd@state.gov.	9	0.002128
13	Richard Verma	6	0.002054
14	Robert Russo	1	0.001692
15	Jacob Lew	6	0.001485
16	Melanne Verveer	16	0.001323
17	sullivanjj@state.gov.	5	0.00115
18	oscar flores	7	0.00105
19	Kris Balderston	14	0.001008
20	Judith McHale	5	0.000799

Results

The figure below shows the graph with the nodes colored by their respective community. As expected the resulting visualization lists the top ranked officials operating within Hillary Clinton's presidential campaign. The graph also groups the highly ranked nodes with their respective communities. For example, (depicted in orange) Lauren Jiloty and Robert Russo belong to the same community and both worked in similar areas: Directors of Communications for Hillary Clinton. In green, Huma Adedin and Lona Valmoro belong to one of the largest communities. This can be attributed to them both working as top executive aides to Hillary Clinton. Interestingly, it can be observed that Jake Sullivan and Cheryl Mills belong to different communities that are adjacent to each other. Perhaps, the similarity of their profession as policy advisers and lawmakers manifests itself though the closeness of the two communities. Lastly, the nodes labeled in red towards the bottom of the visualization group together people that all worked together at the US Dept. of State or in various positions in the White House.



Future Work

Using this experiment as a proof of concept, the next task is to choose a model that best fits the properties we observed in this study. By selecting a model, we can generalize the techniques shown in this work to and apply them to a different dataset. Further work should be done on generating a dataset as well. A mechanism that accepts a data dump (ie email leak) and generates a graph would greatly improve the time spent on data formatting for a graph. It would then be possible for comparisons to be made between multiple models to search for the best fit.

Limitations

This work only provides high-level observations into the dataset. A real-world adversary would be interested in linking those nodes (people) with high PageRanks with possible salacious or confidential emails. Thus, a victim organization would need to take subsequent steps to identify the specific content associated with the important nodes detected in this experiment. In other words, this work provides a possible "where to look first" for someone seeking to classify severity of an email leak. This work did not consider the dataset file containing the content associated with the emails included in the overall graph.

Conclusion

Developing a graphical representation of a dataset provides a helpful reference for isolating related nodes (clustered information) for follow-on analysis. As a greedy approach, a security investigator could determine the influential nodes and focus on applying security mitigations towards those nodes and then work down to other nodes in similar communities. Additionally, the observations gathered can be very helpful to an organization wanting to determine any potential sensitive data loss like: intellectual property or personally identifiable information.

References

[1]. Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

[2]. Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[3]. https://www.kaggle.com/kaggle/hillary-clinton-emails/data

[4]. https://pypi.python.org/pypi/python-louvain/0.3

[5]. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

[6]. B. Hajek, Y. Wu, and J. Xu, "Information limits for recovering a hidden community," in Proc. IEEE Int. Symp. Inf. Theory, Barcelona, Spain, Jul. 2016, pp. 1894–1898.

Alan Adame (a2adame@ucsd.edu)
ECE289: Big Network Data
Presented 5 December 2017