Anomaly Detection with Joint Representation Learning of Content and Connection

Junhao Wang [1] **Renhao Wang** [2] **Aayushi Kulshrestha** [1] **Reihaneh Rabbany** [1]

Abstract

Social media sites are becoming a key factor in politics. These platforms are easy to manipulate for the purpose of distorting information space to confuse and distract voters. Past works to identify disruptive patterns are mostly focused on analyzing the content of tweets. In this study, we jointly embed the information from both user posted content as well as a user's follower network, to detect groups of densely connected users in an unsupervised fashion. We then investigate these dense sub-blocks of users to flag anomalous behavior. In our experiments, we study the tweets related to the upcoming 2019 Canadian Elections, and observe a set of densely-connected users engaging in local politics in different provinces, and exhibiting troll-like behavior.

1. Introduction

There have been multiple recent studies on how social networks have been manipulated to foster divisions among people, and the tactics used for information operations. Wilson et al. (2018) define Information Operations as the suite of methods used to influence others through the dissemination of propaganda and disinformation. The aim is to paralyze the decision making abilities of individuals, thus making the public vulnerable. With these operations at large, democracy stands threatened. A case in point is the 2016 US Presidential Elections which are believed to have been swayed through social media by interference from Russian trolls and bots (Badawy et al., 2018). This was further validated by Twitter in 2018, when they reported possible engagement of 1.4 billion users with suspected "trolls" from the Russian government funded Internet Research Agency (Policy, 2018).

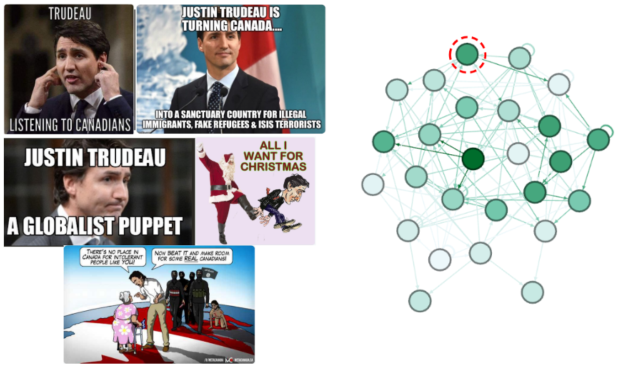

TRUDEAU
LISTENING TO CANADIANS
JUSTIN TRUDEAU IS TURNING CANADA....
INTO A SANCTUARY COUNTRY FOR ILLEGAL IMMIGRANTS, FAKE REFUGEES & ISIS TERRORISTS
JUSTIN TRUDEAU
A GLOBALIST PUPPET
ALL I WANT FOR CHRISTMAS
THERE'S NO PLACE IN CANADA FOR INTOLERANT PEOPLE LIKE YOU!
NOW BEAT IT AND MAKE ROOM FOR SOME REAL CANADIANS!

Figure 1. Sample Memes used by Suspicious Anomalous User in Cluster #8 and its Follower Network Location

Twitter is one of the most commonly used platforms to mobilize public at the time of political unrest. Therefore, it is imperative that we develop tools to identify Information Operations at an early stage, leading to a healthy democratic society. Towards this goal, we analyze the activity on Twitter related to the upcoming elections in Canada, to proactively monitor the interactions on this platform. The main contributions of the current work are:

- We create a joint autoencoder based solution to the problem by formulating Information Operation detection as dense sub-block detection on binary attributed graph, which encodes both the content of tweets and the connections in the Twitter follower network. The dense sub-blocks are detected using density-based clustering on learned node embeddings of the graph.
- We design an adaptive hyperparameter selection method by generating task-specific synthetic data, thus solving the problem of lack of objective evaluation standard for this unsupervised anomaly detection tasks.
- We demonstrate the application of our solution to real-world data by identifying a sub-block of suspicious and tightly connected users, as well as a suspicious account exhibiting behaviors related to Information Operations.

[*]Work done during an internship at Google AI [1]School of Computer Science, McGill University, Montreal, Canada [2]Department of Computer Science, University of British Columbia, Vancouver, Canada. Correspondence to: Junhao Wang <junhao@umich.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

2. Data Collection

In this paper, our main focus would be to study the political interactions of Twitter users surrounding Canadian 2019 Federal Election, and identify groups of users that seek to disseminate crafted messages (propaganda). The dataset comprises of 38,498 tweets from 7,298 distinct Twitter users. The tweets were collected using the Twitter Streaming API, using the following hashtags - #Trudeau, #TrudeauMustGo, #cdnpoli, #TrudeauResign, #LavScam, #SNCgate, #StandWithTrudeau. We also collected the list of followers for each of the 7,298 users in our dataset to construct a follower network, resulting in a total of 474,459 connections. The hashtags are further formulated as a vector of size 3,047 to represent the attributes of each node or user, denoting whether the user used a certain hashtag in the dataset. We represent the entire data as concatenated adjacency and attribute matrix as shown in Figure 4.

3. Method

Definition 1 (Binary Attributed Graph). *A binary attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ consists of: (1) the set of nodes $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$, where $|\mathcal{V}| = n$; (2) the set of edges $\mathcal{E}$, where $|\mathcal{E}| = m$; and (3) the binary node attribute matrix $\mathbf{X} \in \{0,1\}^{n\times d}$, where the i^{th} row vector $\mathbf{x}_i \in \{0,1\}^d, (i = 1 \ldots n)$ is the binary attribute information for the i^{th} node.*

Definition 2 (Binary Attributed Graph Dense Sub--block). *For a binary attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ and a density threshold $t \in [0, 1]$, t-induced dense sub-blocks are set of tuples $\{(\mathcal{S}_\mathcal{V}, \mathcal{S}_\mathbf{X})\}$ where $\mathcal{S}_\mathcal{V}$ is subset of nodes $\mathcal{V}$ and $\mathcal{S}_\mathbf{X}$ is subset of binary attributes of $\mathbf{X}$, such that: (1) subgraph induced by $\mathcal{S}_\mathcal{V} : \mathcal{S}_\mathcal{V}(\mathcal{V}, \mathcal{E})$ has network density above t, and (2) bipartite graph induced by top nodes $\mathcal{S}_\mathcal{V}$, bottom nodes $\mathcal{S}_\mathbf{X}$ and edges $\mathcal{E}_b$, has network density above t, where $\mathcal{E}_b$ correspond to activation of binary attributes. We denote adjacency matrix of $\mathcal{G}$ by $\mathbf{A}$, attribute matrix by $\mathbf{X}$.*

We represent our data as a Binary Attributed Graph with adjacency matrix $\mathbf{A}$ encoding follower relations and binary attribute matrix $\mathbf{X}$ encoding user hashtag usage. We are interested in detecting binary attributed graph dense sub-block. Our approach to tracking Twitter political interactions and potentially identifying Information Operations is a 3-fold process: (1) learn node embeddings using joint autoencoder to minimize reconstruction error of the adjacency matrix and attribute matrix and preserve pairwise Jaccard distance of input vectors, (2) apply density-based clustering on the node embeddings, (3) conduct manual inspection of filtered clusters.

3.1. Joint Autoencoder

Our joint autoencoder architecture is inspired by (Wang et al., 2018), where we extend their loss functions to deal with joint information of attribute and adjacency matrix.

Definition 3 (Joint Autoencoder)**.** *A joint autoencoder shown in Figure 5 is characterized by $(\phi^e_\mathbf{A}, \phi^e_\mathbf{X}, \phi^e_\mathbf{J}, \phi^d_\mathbf{A}, \phi^d_\mathbf{X})$ where ϕ can be a function represented by a single layer of neural network or composition of multiple layers, ϕ^e is encoding function and ϕ^d is decoding function. Subscripts of $\phi : \mathbf{A}, \mathbf{X}, \mathbf{J}$ denote the information ϕ^e encodes from or ϕ^d decodes into, where $\mathbf{A}$ and $\mathbf{X}$ are adjacency and attribute matrix of $\mathcal{G}$ defined previously, and $\mathbf{J}$ is concatenated latent representation of them: $concat(\phi^e_\mathbf{A}(\mathbf{A}), \phi^e_\mathbf{X}(\mathbf{X}))$. Decoders $\phi^d_\mathbf{A}$ and $\phi^d_\mathbf{X}$ transform joint latent representation $\mathbf{J}$ to approximation of $\mathbf{A}, \mathbf{X} : \hat{\mathbf{A}}, \hat{\mathbf{X}}$.*

The joint reconstruction error weighted by hyperparameters $w_\mathbf{A}$ and $w_\mathbf{X}$, with attention weights $\mathbf{W}^\mathbf{A}_{att}$ and $\mathbf{W}^\mathbf{X}_{att}$ is calculated by

$$\mathbf{H} = \phi^e_\mathbf{J}(\mathbf{J}) \tag{1}$$

$$\mathcal{L}^\mathbf{A}_{recon} = ||(\phi^d_\mathbf{A}(\mathbf{H}) - \mathbf{A}) \odot \mathbf{W}^\mathbf{A}_{att}||^2_F \tag{2}$$

$$\mathcal{L}^\mathbf{X}_{recon} = ||(\phi^d_\mathbf{X}(\mathbf{H}) - \mathbf{X}) \odot \mathbf{W}^\mathbf{X}_{att}||^2_F \tag{3}$$

$$\mathcal{L}_{recon} = w_\mathbf{A}\mathcal{L}^\mathbf{A}_{recon} + w_\mathbf{X}\mathcal{L}^\mathbf{X}_{recon} \tag{4}$$

Besides reconstruction loss, we define similarity loss as the discrepancy between pairwise Euclidean distance of $\mathbf{H}$ and pairwise Jaccard distance of $\mathbf{A}$ and $\mathbf{X}$, weighted by the same $w_\mathbf{A}$ and $w_\mathbf{X}$. In order to compare these 2 different distance metrics, we apply a logit transformation on the pairwise Euclidean distance to compress its range to $[0, 1]$, the same as the range of pairwise Jaccard distance. Let $\mathbf{S}^\mathbf{X}_{Jar}$ be the pairwise Jaccard distance of rows of $\mathbf{X}$, similarly $\mathbf{S}^\mathbf{A}_{Jar}$ for $\mathbf{A}$, and $\mathbf{S}^\mathbf{H}_{Euc}$ be the pairwise Euclidean distance for latent vectors $\mathbf{H}$, and choose $\lambda \geq 0$:

$$\mathcal{L}^\mathbf{A}_{sim} = ||exp(-\lambda\mathbf{S}^\mathbf{H}_{Euc}) - \mathbf{S}^\mathbf{A}_{Jar}||^2_F \tag{5}$$

$$\mathcal{L}^\mathbf{X}_{sim} = ||exp(-\lambda\mathbf{S}^\mathbf{H}_{Euc}) - \mathbf{S}^\mathbf{X}_{Jar}||^2_F \tag{6}$$

$$\mathcal{L}_{sim} = w_\mathbf{A}\mathcal{L}^\mathbf{A}_{sim} + w_\mathbf{X}\mathcal{L}^\mathbf{X}_{sim} \tag{7}$$

The joint loss to minimize is weighted combination of reconstruction loss and similarity loss with weights w_{recon} and w_{sim}, plus L2 regularization loss at every layer:

$$\mathcal{L}_{joint} = w_{recon}\mathcal{L}_{recon} + w_{sim}\mathcal{L}_{sim} + \mathcal{L}_{reg} \tag{8}$$

In practice, we train on sampled batches instead of the entire data matrix. For each epoch, we select node $v_i \in \mathcal{V}$ uniformly at random, and sample set of nodes $\{v_j : v_j \in \{\mathcal{V} - v_i\}\}$ according to some distribution D related to the similarity between v_i and v_j.

3.2. Density-based Clustering

We transform the problem of dense sub-block detection of the binary attributed graph to density-based clustering of latent embeddings of nodes in Euclidean space, thus enabling the use of established density-based clustering algorithm such as DBSCAN (Ester et al., 1996), as well as making the process more interpretable. We first apply dimensionality reduction using Uniform Manifold Approximation and Projection (McInnes et al., 2018) on $\mathbf{H}$ to get $\mathbf{H}_{reduced}$ with 2 dimensions, and then apply DBSCAN on $\mathbf{H}_{reduced}$ with parameters that cater to the size of the clusters that we are interested to study. Then we define dense clusters as clusters that induce subgraphs of Twitter follower network whose network density is above a specified threshold. In practice, we look at the top k densest clusters returned by DBSCAN. The motivation for using network density to indicate anomaly is related to Wilson et al. (2018)'s definition of Information Operations: we aim to study densely connected users exhibiting similar hashtag usage, thus potentially showing similar attitudes towards certain events or ideologies, which is indicative of the presence of Information Operation.

4. Experiments

We first conduct synthetic experiments to choose the best set of hyperparameters for exploratory analysis on real data, as well as to demonstrate the effectiveness of our algorithm on binary attributed graph dense sub-block detection compared to FRAUDAR (Hooi et al., 2016), a classical baseline for dense sub-block detection with only adjacency matrices, and DOMINANT (Ding et al., 2019), a Graph Convolutional Network (GCN) based approach that utilizes both adjacency and attribute matrices. We then create a joint "fingerprint" of identified clusters based on both the graph topology of cluster-induced subgraph, and attributes of nodes in the cluster, which could potentially be used to identify Information Operations in Canadian 2019 Federal Election. We also manually inspect the nodes in the three clusters with highest cluster-induced network density, and find some suspicious accounts that might have engaged in Information Operations.

4.1. Hyper-Parameter Tuning

We inject artificial dense sub-block anomalies into our Twitter data in order to tune our algorithm to perform well for the unsupervised anomaly detection task. With the injected data, we conduct a random search of the hyperparameter space and identify the best hyperparameter option by F-1 score, with labels being anomaly or non-anomaly. Then we use it to identify interesting dense clusters on the real data without dense sub-block injection. We show that our method outperforms both baselines across all injected sub-block densities in Figure 2.

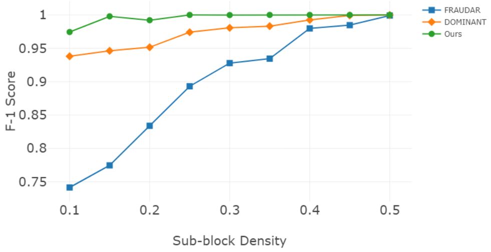


Figure 2. Synthetic Experiment Performance

4.1.1. Synthetic Data Generation

For adjacency matrix, we inject dense subgraph by injecting random dense graph with a specified density and size at sub-block indices. For attribute matrix entries, we create an empirical distribution of hashtag usage indicating how likely a random person from a sub-block would use certain hashtags, and apply add-k smoothing on this empirical distribution. Next, we sharpen the distribution by applying an exponential factor to it: $exp(\lambda \cdot)$ where λ controls for how concentrated the transformed distribution is. By sampling a certain number of hashtags from this distribution, we simulate the presence of Information Operations, where a group of highly connected users tweet a subset of hashtags. Finally we inject the bipartite graph with the specified density at these sub-block and attribute indices. For our experiment, we inject 3 dense sub-blocks of size 500, and use the same network density for both adjacency matrix and attribute matrix, from 0.1 to 0.5 with 0.05 interval.

4.2. Results

Using the best hyperparameter option, we create 10 clusters. For each cluster and its corresponding induced sub-graph of follower network, we generate hashtag fingerprint, which reflects user attribute information, as well as clustering fingerprint, which reflects key network topology information. We put our focus on 2 of the densest clusters (#9, #1), and report interesting exploratory findings.

For each cluster, we define hashtag fingerprint as the relative usage frequency of popular hashtags within cluster. Note that usage here refers to whether a hashtag is used or not in the dataset for a given user, and frequency refers to the number of users in a cluster using certain hashtag. A high relative usage frequency corresponds to highly used hashtag in a cluster. Hashtag fingerprints for cluster #9 and #1, and a randomly sampled set of users are shown in Figure 3. Similarly, for each cluster-induced subgraph of the follower

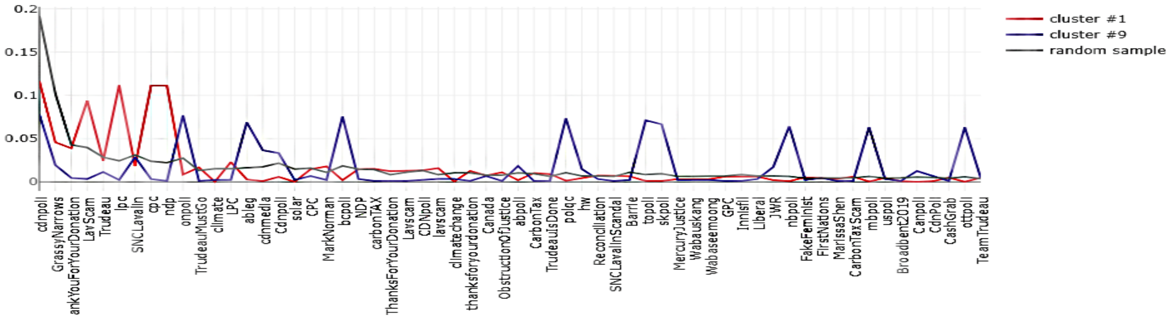


Figure 3. Hashtag Fingerprint: Per-cluster relative usage frequency for popular hashtags

network, we define clustering fingerprint (shown in Figure 6) by the probability density of node clustering coefficients in the subgraph. Clustering coefficient of each user captures the connected-ness of its neighbors.

By analyzing the hashtag fingerprint we note that cluster #9 exhibits interesting spikes on hashtags related to diverse locations: Alberta (#ableg), British Columbia (#bcpoli), Quebec (#polqc), Toronto (#topoli), Saskatchewan (#skpoli), New Brunswick (#nbpoli), Manitoba (#mbpoli) and Ottawa (#ottpoli). This is counter-intuitive; we normally assume people engage with each other locally, but we clearly see multi-regional cluster of users in close contact with each other.

For cluster #1, the hashtag usage centers around a recently heated political scandal related to government and corporate corruption (#LavScam), and a few prominant political parties: the Liberal Party of Canada (#lpc), the Conservative Party of Canada (#cpc), and the New Democratic Party (#ndp). This reflects the fact that an emergent cluster of users related to different political parties are talking about the recent scandal.

The hashtag fingerprint for both clusters identified through our algorithm reveals interesting insights that would otherwise be hard to obtain by going through the tweets manually. On the other hand, the hashtag fingerprint of a random sample is highly centered around the most popular hashtags and cannot yield much insight into the user group.

Inspecting clustering fingerprints in Figure 6, both cluster #1 and #9 exhibit a more spread-out distribution compared to random sample. In particular, cluster #9 where tweet hashtags are related to diverse locations exhibits clustering coefficient distribution centered around value (0.1 to 0.2) higher than what we would expect for such a multi-regional cluster. To investigate the reason for such phenomena is an interesting direction of future empirical study.

4.2.1. Sample User

We finally manually inspect the Twitter profile page of users in top 3 densest clusters (#8, #9, #1), and for each cluster, we visualized its cluster-induced subgraph and per-node HITS authority score (Kleinberg, 1999). We use darker green gradient to denote nodes with higher HITS authority score.

Shwon in Figure 1, we identify one Twitter account highlighted with a dotted red line that exhibits behaviors suspicious of Information Operations. The suspicious user account was created in August 2017. Since December 2017, the user started consistently creating and spreading divisive tweets and memes that demote Justin Trudeau and his administration. A sample of the political memes deployed by this user is shown in Figure 1. Furthermore, this user changed it user handle twice from mid-April to mid-May. Such high frequency of changing user handles might be related to malicious intent (Jain & Kumaraguru, 2016). Both the identified user's posted content and behavior are suspicious of engaging in Information Operations.

5. Conclusion

In this work, we proposed an embedding based solution to track Twitter political interactions and potentially identify anomalous user groups. This approach is built on the recent advances on representation learning for graphs and hence provides a scalable solution for this domain. We contribute three key insights: (1) Information Operations detection on Twitter can be formulated as a dense sub-block detection problem on binary attributed graphs that encode both network topology and user attribute information; (2) dense sub-block detection can be formulated as a density-based clustering problem on latent Euclidean embeddings of the nodes in the graph; (3) each cluster identified by the algorithm can be assigned a joint fingerprint that yields insight which is otherwise not possible through manual inspection of tweets. For code and more information see `https://sites.google.com/view/jointdetect/`

References

Badawy, A., Ferrara, E., and Lerman, K. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265. IEEE, 2018.

Ding, K., Li, J., Bhanushali, R., and Liu, H. Deep anomaly detection on attributed networks. 2019.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pp. 226–231, 1996.

Hooi, B., Song, H. A., Beutel, A., Shah, N., Shin, K., and Faloutsos, C. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 895–904. ACM, 2016.

Jain, P. and Kumaraguru, P. On the dynamics of username changing behavior on twitter. In *Proceedings of the 3rd IKDD Conference on Data Science, 2016*, CODS '16, pp. 6:1–6:6, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4217-9. doi: 10.1145/2888451.2888452. URL http://doi.acm.org.proxy3.library.mcgill.ca/10.1145/2888451.2888452.

Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

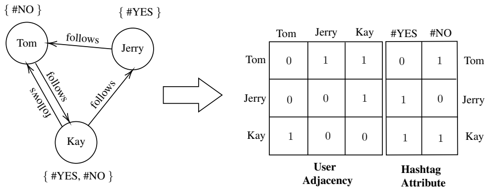
McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Policy, T. P. Update on twitters review of the 2016 us election. *Retrieved April*, 15:2018, 2018.

Wang, H., Zhou, C., Wu, J., Dang, W., Zhu, X., and Wang, J. Deep structure learning for fraud detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 567–576, Nov 2018. doi: 10.1109/ICDM.2018.00072.

Wilson, T., Zhou, K., and Starbird, K. Assembling strategic narratives: Information operations as collaborative work within an online community. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):183, 2018.

A. Plots and Images



	Tom	Jerry	Kay	#YES	#NO	
Tom	0	1	1	0	1	Tom
Jerry	0	0	1	1	0	Jerry
Kay	1	0	0	1	1	Kay
	User Adjacency			Hashtag Attribute		

Figure 4. Concatenated Adjacency and Attribute Matrix

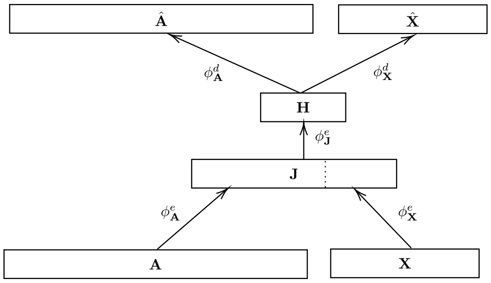


Figure 5. Joint Autoencoder Architecture

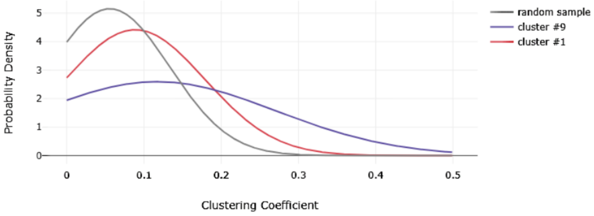


Figure 6. Clustering Fingerprint: Per-cluster probability density for node clustering coefficients