Russian Ads Network Investigation and Classification using SNA and GCNs

Harika Naidu Etha[1] and Chaitanya Mundle[1]

Clemson University

Abstract. Understanding hidden patterns of complex real-world data is a challenging task. However, the recent advancements in Social Network Analysis (SNA) techniques have shown promising results for investigating real-world complex data. Here, we investigate the Russian Ads Network (RAN) using SNA approaches to identify hidden patterns from the network. Our study reveals interesting findings include 1), each topic including but not limited to racism, islamophobia, stop-AI and Muslimvoice represent different communities in the network. This implies the topic-based community structure of the RAN network. 2) There exist some communities which indicate conflicting topics such as stop refugees (no more mosques) and stop islamophobia which influences the whole network. 3) High impression rate does not necessarily lead to high clicks however, high cost has greater chances for getting more clicks. 4) Titles of high influenced URLs and most influenced URLs are identified. Based on these findings, we conclude that conflicting topics lead to high influence because of involving many individuals belong to different communities. In addition, we employ graph convolutional networks to predict ads' click rate. The model has shown good accuracy results over the RAN network.

Keywords: Social Network Analysis · Online Ads · Graph Convolutional Networks · Community detection · Centrality Measure.

1 Introduction

Networks are the backbone of various real-world complex systems including but not limited to online social networks, biological systems, transportation systems, citation networks, and collaboration networks [4]. Investigating the behaviors and interactions between entities in these complex systems provide an understating of various hidden characteristics that provide the ability to enhance the functionality of these systems. For example, understanding the likes of online social network users helps to recommend likely friends and utilities. Moreover, learning the molecular structure of toxic and non-toxic molecules help to predict whether the patient is healthy or unhealthy [15]. Regardless of investigating such complex networks, recent deep learning-based approaches such as Graph Convolutional Networks (GCNs) [9, 6] have shown promising results in learning over graph-structured data.

Online social networks provide various benefits to users such as real-time news and information discovery, easy and instant communication, friends and

family connectivity, and availability of enormous material for fun and enjoyment. However, they have many disadvantages, among them, the privacy and misinformation issues are the top in the list [14, 20]. Effecting users' behavior with misinformation using online social networks have been widely reported and highly studied the issue of online social networks [8]. Twitter and Facebook are vital sources to disseminate information and change people's opinions. In May 2018, congress published the Russian Facebook Ads data that were used to try and influence the US 2016 presidential election [1]. These were political Facebook ads purchased by Russian groups hoping to sow discords before and after the US election.

Having the advanced tools of network science and machine learning, it is worthwhile to investigate this data at micro and macro levels [15, 7, 16]. In this work, we aim to investigate the released image social data by creating a social network (denoted as RAN) and Social Network Analysis (SNA) and GCNs to highlight the following hypotheses.

- Explore the community structure of the network
- Identifying the most influential ads
- Does the community structure of the network provide insights concerning ads and sources of the ads?
- Exploring the nature of the ads. As they are reported to be racial and hated
- Exploring the network in micro-level to highlight significant sources of the ads (Facebook pages, URLs etc.)
- Is there any correlation between impression rate, cost and click rate?
- Apply Graph Convolutional Networks to predict the clicks rate

We investigate the Facebook ads network created from reverse image search results using advanced SNA tools to answer the above hypotheses. The tightly coupled communities of the network motivated us to deploy GCNs for ads clicks rate prediction. The main findings of the study are very interesting and as follows.

1. The network has tightly coupled community structure, and each community represent a specific type of people, which clearly show that such a community of people is targeted
2. There have been influential ads in each community which depicts the same type corresponding to their community
3. Many of these ads are racial
4. There is a strong correlation between impression rate, cost and clicked rate
5. GCNs show promising results on the classification task

This document can be organized is as follows.

2 Network Construction and Characteristics

The network consists of 45643 nodes and 66909 edges where nodes represent ads (images) and a link exist between two nodes if they appear on the same source. Note that this network has been made from the provided ads images

and then reverse image search has been performed on each image to find the sources where they appeared. And then a link between two images is created if they appeared on the same source. For example, if two ads appear on the same Facebook page, then there would be a link between them. The resultant network is a simple directed graph with no multi-edges and thus will be considered as a simple directed graph throughout the analysis We show the overall statistics of the network in table 1 to understand the structure of the network. The network characteristics include the total number of nodes, the total number of edges, the average clustering coefficient, average path length, network diameter, average degree, modularity score, and network density. Based on these statistics, we can see that our network is quite sparse but suitable for applying SNA techniques to explore and extract hidden information. Note that $|V|$ represents the total number of nodes, $|E|$ represents total edges and Avg CC represents the average clustering coefficient of the network. Figure 1 shows macro visualization of the network.

$\|V\|$	$\|E\|$	Avg. degree	Avg. CC	Avg. path length	Diameter	Denstiy	Modularity
45643	66909	1.466	0	0	1	0	0.939

Table 1. Russian Ads network characteristics

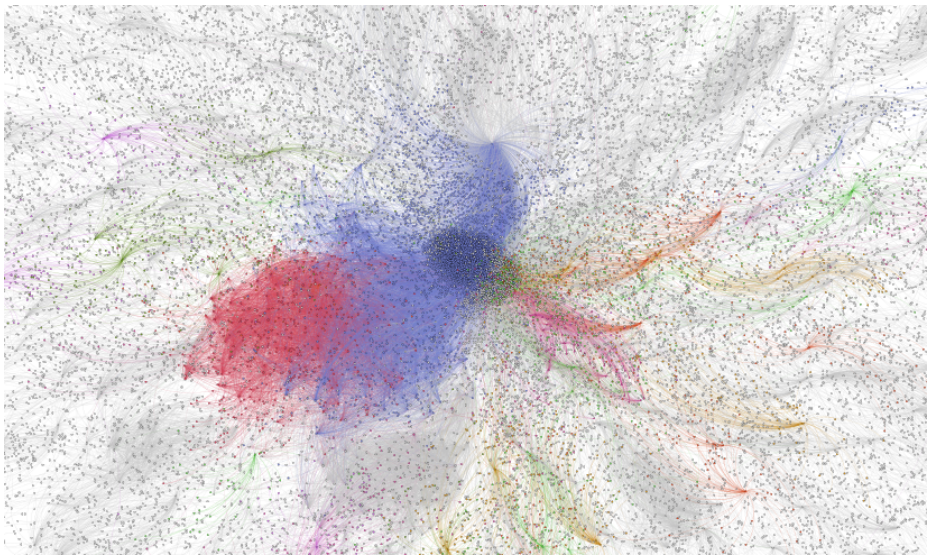

Fig. 1. A snapshot of the Russian ads network. Different colors represent communities in the network. The visualization has been made in Gephi

3 Exploring the community structure of the network

A community structure has no universal definition, but it is widely accepted that communities are sub-groups of network nodes that are densely intra-connected and sparsely interconnected [15]. Some networks may have well-defined communities, while in other networks, communities need to be discovered through suitable algorithms. Identifying community structure reveals abundant hidden information from the network [11].

Here, we explore the community structure of RAN network using Blondel algorithm [3]. This finds closely connected components by using the modularity function [12]. Modularity is a well-known measure of the quality of the communities discovered in a network; it evaluates the extent of the network's division into modules. The denser the connections within the modules and the sparser the connections between them, the higher the modularity value [10, 5]. The visualization of the network community structure is shown in figure 2 while figure 3 presents the tagged communities of the network.

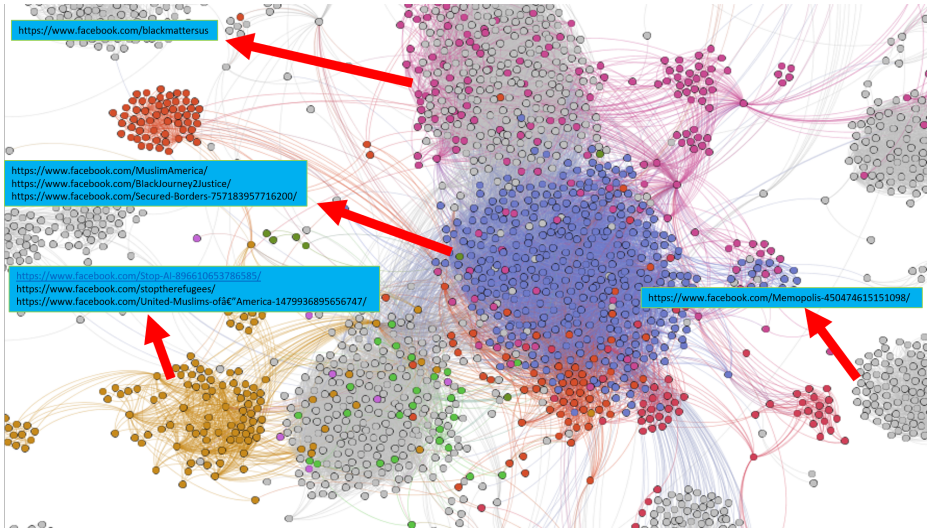

Fig. 2. Community structure of Russian Ads Network. The corresponding highly used source urls are highlighted against each community.

To extract information from the community structure of the network, we performed the following steps.

1. We load the network into Gephi
2. We apply modularity algorithm
3. In the appearance window, we select a partition and then choose modularity class from the drop-down menu

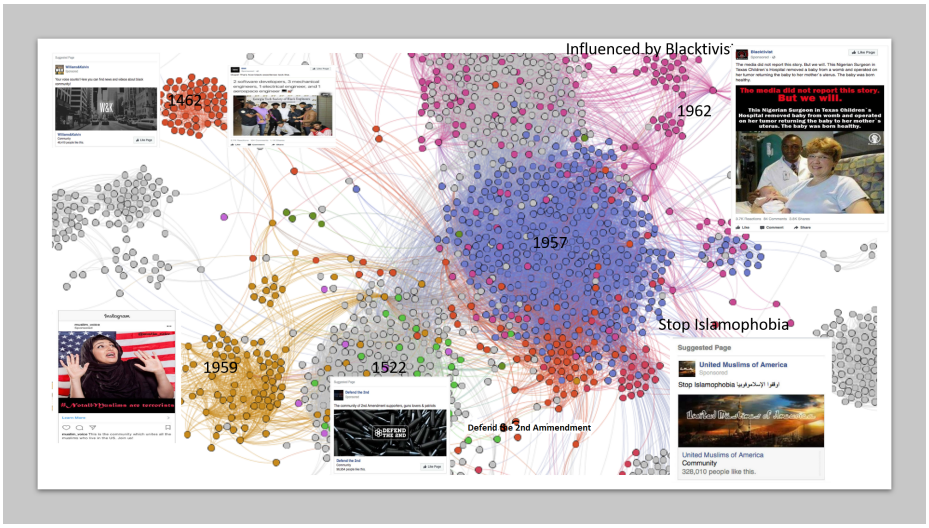

Fig. 3. Few labelled larger communities of RAN

4. We choose the color palette for coloring the communities
5. apply changes

We can see from the network statistics that our network is a quite sparse and low degree is abundant or isolated nodes, therefore, we filter the network and remove nodes having less than 4 degrees. Filtering the network able us to deep dive into the network to explore hidden information. We found 5 major communities: 1462,1522,1959,1957 and 1962 (labeled by Gephi automatically) in which the community no.1957 covers 10% of the whole network while the remaining covers 1,1% part of the network. We search for the node Ids of each community (from communities.csv file) and find their links, sources and ads title in ads.csv file to figure out the nature of the ads. For the larger community, we found that major ads of this community are related to the Muslim-voice topic. We have shown the image over the figure of this community which is a Facebook page of United Muslims of America. Based on a large number of users related to the topic of Muslim-voice, we named the community as stop-islamophobia. This is an interesting findings of this study.

The second community of the network which we named is 1462. We trace the titles and sources of this community and found that major posts of this community related to the black community like W&K and BM page (shown images) which are highly sourced from this community. In the community 1522, the majority of ads related to "defend the second amendment" while community 1922 influenced by blackactivist Facebook page. Community 1959 is found very interesting because it is influenced by contradicting topics and includes ads from Muslim-voice, stop refugees and stop-AI topics.

Overall, we conclude that communities of RAN represent different regional-wise topics in which many similar kinds of ads posted and thus have made the community. Thus, based on these findings, one can estimate the number of active topics on social topics by just finding the number of communities within the network. In summary, the Russian targeted different kinds of people like black people, Muslims, AI, and migrants.

4 Influential Ads

To find the influential ads from the RAN network, we employ PageRank algorithm [13]. PageRank is one of the widely used algorithms used in many recommender systems, search engines, and Social Network Analysis. In the context of SNA, influential users are those who are involved in many interactions in different contexts with other users. figure 4 presents the influential users of RAN network identified through PageRank algorithm. Note that we have shown the visualization of the filter network to show a clear picture of the network [18, 19].

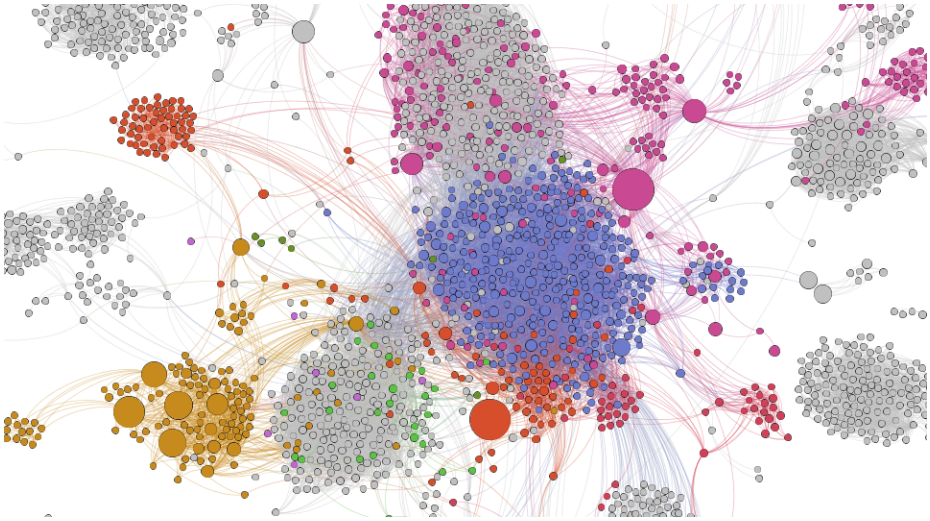

Fig. 4. Influential users of RAN identified through pageRank algorithm

We can see that each community contains several influential users (nodes with greater size) which indicate that there are some active groups or sources which spread a large amount of information with high influence. This section will study and highlight those adds and influential communities identified through PageRank algorithm.

Influential Ads community In this section, we present the influence of the community 1959. This community contains several influential ads like node id

2h6fhvw7ia7rq, 2h6fhvw7ia8lf, 2h6fhvw7ia8le and 2h6fhvw7ia7rs. We further investigated that to figure out the reason for the high influence of this community. Major topics discussed in this community also highlighted in Figure 5 are as follows.

Fig. 5. Influential and conflicting topic RAN community

- Stop-AI
- Muslim-voice
- Stop the refugees
- Awful police brutality-black man calls
- Heart of Texas

We can see that all the above topics are either racial like Muslim-voice, stop the refugees and Awful police brutality-black man calls, while the remaining are politically influencing topics.

5 Ads analysis

In this section, we explore the ads network to answer the following hypotheses.

Hypothesis 1: What is the correlation between impression rate and click rate?

Hypothesis 2: What is the correlation between cost, impression rate and click rate?

To answer these hypotheses, we plot impression vs clicked rate and impression vs clicked vs cost as shown in figure 6 and figure 7. In figure 6, we can see that highly impression Ads not always get the high clicked rate. On the other hand,

we can see in figure 7 that clicked rate costs are quite consistent with each other. These results highlight two important findings: a). Highly impression rate ads not always get more clicked although they have a higher chance of getting more clicked. 2) highly cost ads are more likely to get more clicks. Interestingly, the bars also show that very high costs above than the normal always do not lead the highest number of clicks, but overall the pattern is clear (high cost, high clicks).

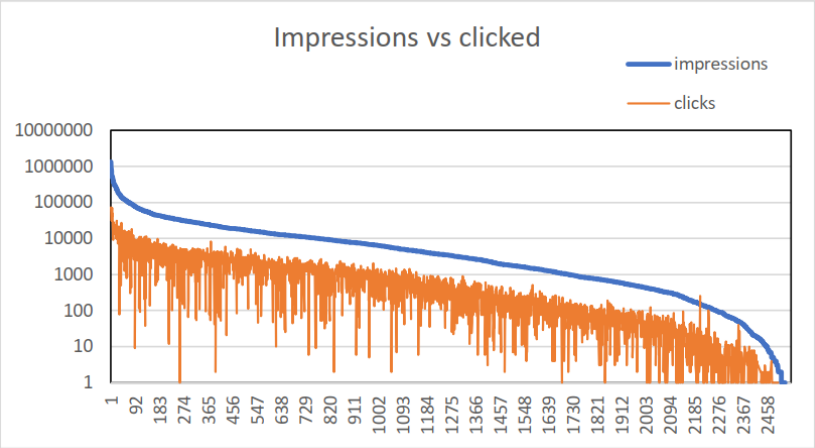

Fig. 6. Impression vs clicked rate

To get the idea that which kind of ads lead to high clicked. Here. we presented a few ads title, however, for complete titles and node IDs, please see impression Vs clicked excel file. From the tile of these ads, we can see that ads belong to different communities and they are making them influenced together.

- Community of people who support our brav
- Brown Power is a platform designed to ed
- Staying Woke,bui—ding-up our communities
- Say it loud: I'm black and I'm proud!
- United We Stand! Welcome every patriot w
- Heritage, not hate. The South will rise
- This video may bring tears to your eyes.
- We are proud to be black and stand for o

Finally, here we show some influenced URLs from the data. For the complete list, high PageRank URLSs form all excel files can be investigated.

- https://blackmattersus.com/ (The most influential url)
- https://www.facebook.com/BIacktivist-128371547505950/

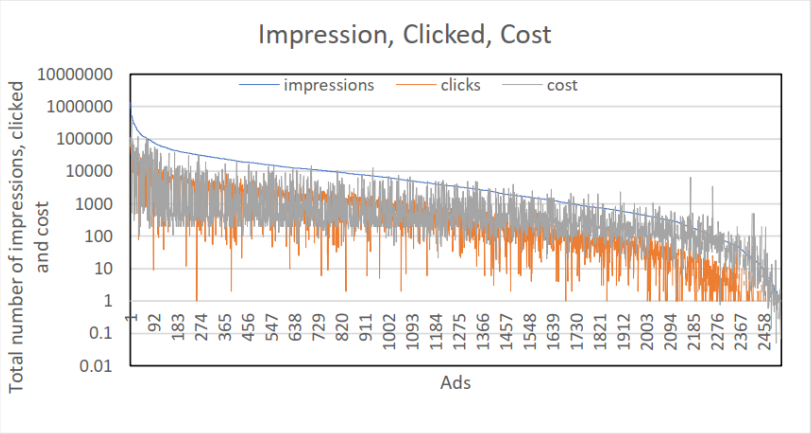

Fig. 7. Impression vs cost vs clicked rate

- https://www.facebook.com/savetheZa/
- https://www.facebook.com/Heart-of-Texas-761114847343674/timeline/
- https://www.facebook.com/events

6 Employing GCNs

The enormous applications of learning over graph-structured data have attracted the research community a lot in the last couple of years[7]. There has been a surge of approaches on representation learning in the last couple of years where it maps the problem as a supervised or semi-supervised classification problem [2, 21]. A well-known model known as GCNs uses a first-order approximation in the Fourier domain [9]. The main idea behind this model is that, for a given undirected and unweighted graph, the class of a node can be predicted by its local neighborhood. The authors propose a layer-wise propagation rule for semi-supervised classification. Additionally, they employ first-order approximation for spectral graph convolution. However, the proposed algorithm requires a full-graph Laplacian during the training phase. The model has achieved promising results with several benchmark datasets.

In the setting to supervised learning on graphs, a graph $G = (V, E)$ consisting of nodes (V) representing ads in the network and edges E indicates if they are shared on the same source. The network is undirected in the sense that an ads x and y shared on the same source are indistinguishable from the ads y and x shared from the same source. If there are n ads in the network then the corresponding network can be represented by $n \times n$ adjacency matrix A, where each value A_{ij} indicated whether they have shared from the same source or not. Each ad has associated an integer number representing the number of clicks which is our desired Y label to train the model and learn the structure of the

network concerning the corresponding label y. Thus, we have a set of nodes $\{v_1, v_2, ..., v_n\}$ and their labels $\{y_1, y_2, ..., y_n\}$.

6.1 Regression to classification learning problem

We transform the problem from regression to classification and made four classes: very-high, high, medium and low. Each class representing the clicked rate of the ads. However, as we have integer values, thus we need a proper way to perform proper binning. We employed Quantile-based discretization function which discretizes variable into equal-sized buckets based on rank or based on sample quantiles. Note that quantiles are just the most general term for things like percentiles, quartiles, and medians. This function divides the whole dataset into the given number binns with the same number of samples. For example, if we have 100 samples and want to bin them into five categories then each bin will have 20 samples.

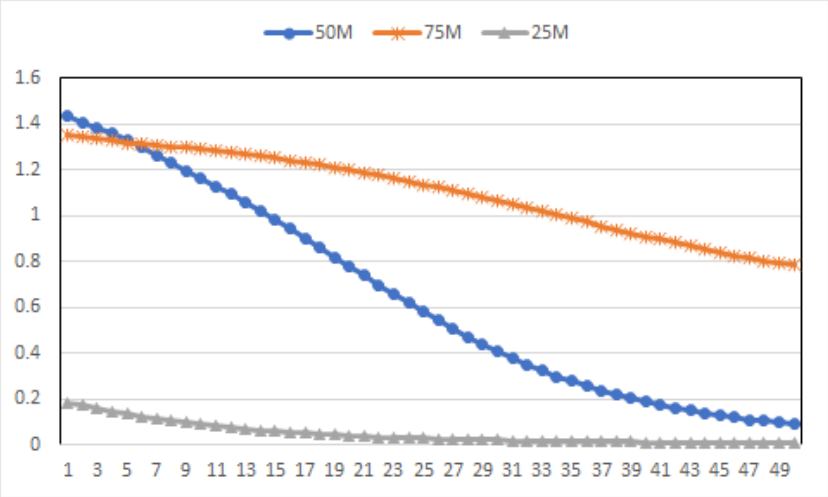

Fig. 8. GCNs results on ads network with different number of labelled nodes

6.2 Experimental setup

The input of our model is an adjacency matrix $A^{n\times n}$, a feature matrix $X^{n\times d}$ which is an identity matrix in our case, as we have no features associated with each node and the label set $Y^{|V|}$ with each node. To train the model, we use the same layer-wise propagation rules presented by Kipf & Welling, 2016 [9] which can be formulated as shown in Equation 4:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{l}W^{l}) \tag{1}$$

Here, $H^(l)$ is the activation at the l th layer, σ is a non-linear activation function, $A = A + I^n$, is the adjacency matrix augmented with self-loops, $\tilde{D}$, is the degree matrix of the graph represented by $\tilde{A}$, and W are the learning parameters. We used the same loss function presented in the original paper [9].

We set the number of layers equal 2 and the number of neurons in the first layer is set to $|V|$. The number of neurons in the second layer is set to 16 and 4 in the final layer. The learning rate is set $1e^{-2}$ and the number of epochs is set to 50. As the model is semi-supervised learning so we use a different set of nodes having $\{5, 25, 50, 75\}$ percent of nodes of training to see the effect of labeled nodes. the results are presented in figure 8.

To evaluate that our code, we also run the model on state of art Cora citation dataset [17] to reproduce the results. The plot is presented in figure 9. We can see that the model performs quite smoothly and accurately.

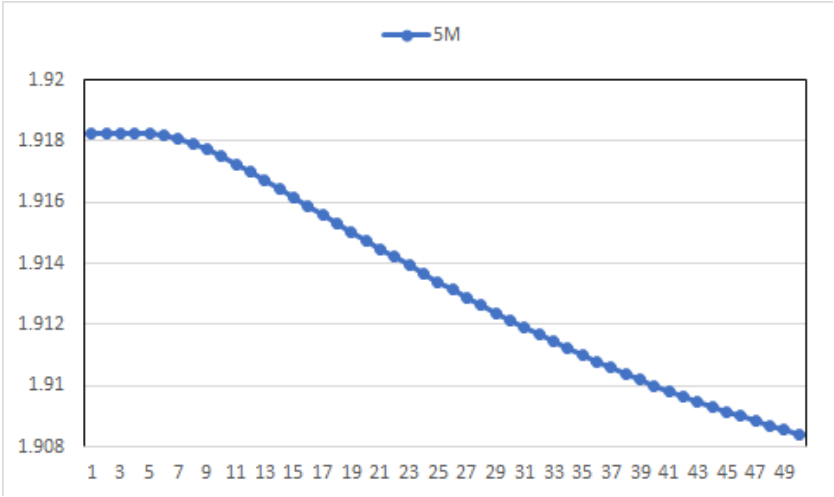

Fig. 9. GCNs results on Cora network with 5% labelled nodes

7 Conclusion

Social Network Analysis and Graph Convolutional Networks have proven to be a good candidate for understanding and learning graph-structured data. In this work, we employed community detection and centrality measures to explore Facebook ads networks. We also performed binning to predict the number of clicks of an ad using GCNs. We found several interesting results like major communities including black, Muslims and AI have been targeted using different sources and different kinds of information have been disseminated. We also observed that a high impression rate does not necessarily lead to high clicks

however, high cost has greater chances for getting more clicks. The classification accuracy of the GCNs also shown encouraging results.

References

1. Congress-social media advertisement data, https://democrats-intelligence.house.gov/facebook-ads/social-media-advertisements.htm
2. Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., Alemi, A.A.: Watch your step: Learning node embeddings via graph attention. In: Advances in Neural Information Processing Systems. pp. 9180–9190 (2018)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
4. Cai, H., Zheng, V.W., Chang, K.C.C.: A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering **30**(9), 1616–1637 (2018)
5. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
6. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems. pp. 1024–1034 (2017)
7. Hamilton, W.L., Ying, R., Leskovec, J.: Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584 (2017)
8. Jung, J., Shim, S.W., Jin, H.S., Khang, H.: Factors affecting attitudes and behavioural intention towards social networking advertising: a case of facebook users in south korea. International journal of Advertising **35**(2), 248–265 (2016)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
10. Leicht, E.A., Newman, M.E.: Community structure in directed networks. Physical review letters **100**(11), 118703 (2008)
11. Malliaros, F.D., Vazirgiannis, M.: Clustering and community detection in directed networks: A survey. Physics Reports **533**(4), 95–142 (2013)
12. Newman, M.E.: Modularity and community structure in networks. Proceedings of the national academy of sciences **103**(23), 8577–8582 (2006)
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
14. Panahi, S., Watson, J., Partridge, H.: Social media and physicians: exploring the benefits and challenges. Health informatics journal **22**(2), 99–112 (2016)
15. Said, A., Abbasi, R.A., Maqbool, O., Daud, A., Aljohani, N.R.: Cc-ga: A clustering coefficient based genetic algorithm for detecting communities in social networks. Applied Soft Computing **63**, 59–70 (2018)
16. Said, A., Bowman, T.D., Abbasi, R.A., Aljohani, N.R., Hassan, S.U., Nawaz, R.: Mining network-level properties of twitter altmetrics data. Scientometrics pp. 1–19 (2019)
17. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. AI magazine **29**(3), 93–93 (2008)
18. Tang, X., Yang, C.C.: Identifing influential users in an online healthcare social network. In: 2010 IEEE International Conference on Intelligence and Security Informatics. pp. 43–48. IEEE (2010)

19. Trusov, M., Bodapati, A.V., Bucklin, R.E.: Determining influential users in internet social networks. Journal of Marketing Research **47**(4), 643–658 (2010)
20. Tucker, C.E.: Social advertising: How advertising that explicitly promotes social influence can backfire. Available at SSRN 1975897 (2016)
21. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)