Reaction Paper # 3: Information Diffusion

Mohammad Elkholy  
Computer Science and Engineering  
 The American University in Cairo  
 New Cairo, Cairo, Egypt  
moelkholy@aucegypt.edu

SUMMARY

Understanding how information spreads in a network is very important for many different domains that deal with complex networks like disease spread across communities, marketing, news spreading, etc... In marketing this is very important as it allows marketing teams in companies to target groups of people that are most likely to cause the highest information to spread across the network, thereby increasing sales and decreasing costs relative to marketing for all consumer segments. It is important in disease spread as finding sources of spread can allow the implementation of policies that minimize the spread as much as possible. The first paper *Maximizing the Spread of Influence through a Social Network* tries to find approximation guarantees for efficient algorithms that can help pick out a subset of individuals that will maximize the information spread across the network, as well as provide some data on their algorithms and how they fair against other known selection heuristics. The second paper *Network-Based Marketing: Identifying Likely Adopters via Consumer Networks* tries to establish the importance of taking into account network properties when trying to make selections for information spread through empirical evidence in a marketing environment. This however could be expanded to other domains.

**Maximizing the Spread of Influence through a Social Network**

The authors in this paper try to provide an approximation guarantee for already existing algorithms that helps tackle this problem. They consider two diffusion models, the linear threshold and Independent Cascade ones. In the linear model, inactive nodes are activated if the influence of its neighbors is greater then a prespecified threshold. This happens across timesteps, with a random choice of active nodes at the beginning. The Independent Cascade model on the other hand is a bit different in the sense that across different timesteps, given an initial set of active nodes, upon activation of a node, it tries to activate its neighbors exactly once with a certain probability. Upon success, the neighbor of this node is activated in the next timestep, and the cascading continues. Finding the optimal set of active nodes at the start is NP-hard for both algorithms. The authors attempt to provide a guarantee by trying to prove that the functions governing these two models are submodular functions, a family of functions with a known approximation using the hill-climbing method of about (1-1/e) of the optimal solution. A submodular function is one that maps subsets of a ground set U to the set of non-negative real numbers such that it satisfies the property of diminishing returns. More formally: where T is a superset of S. The function is monotone in the sense that adding an element to a set doesn’t cause the function to decrease. We want to find a subset S s.t size of S is k and is maximized. This is NP-hard, which is why the hill-climbing approximation is used. It is a greedy algorithm that starts with an empty set and adds nodes to it that will provide the most gain incrementally. Greedy algorithms do fall into the problem of reaching local optimums and not global optimums which is why it is an approximation. The authors then provide proofs for why the functions in these models are submodular and thus can be approximated. They then experiment with the greedy hill-climbing approach for node selection against other heuristics and random selection. It is evident that the hill-climbing method does better than the others as the target set size increases. The next part tries to find generalization of these two models, as well as try to address an assumption that they made at the beginning of the paper, that once a node is activated, it can not be deactivated. They proved that this assumption is not necessary, as even if the nodes were deactivated after becoming active, there are still approximation guarantees for it. The last part addresses the problem of marketing strategy in which we may want to perform some action M for a specific subset of nodes N. They provide approximation guarantees for that too.

**Network Based Marketing**

The authors in this paper try to provide some empirical evidence for taking advantage of links between consumers to increase sales. They start by defining the different modes that network-based marketing could take:

* Explicit advocacy: individuals recommending a product to their friends
* Implicit advocacy: individuals advocate by simply adopting the product
* Network targeting: firms targeting previous customers’ neighbors.

The authors then provide an in-depth literature review going over various popular model from different domains, their drawbacks and current research opportunities and challenges researchers could face. The main challenges were:

* Low incidence of response: customers do not adopt product so less data available for use in models, could be handled by Poisson regression.
* Separating word of mouth from homophily: like minded customers are more likely to buy the same products, so it is not necessarily information diffusion that is happening.
* Incorporating extended network structure: Using data about the network can help improve the models’ performances.
* Missing Data: a lot of datasets are missing data about segments of the population, particularly the ones that the company or firm has still not expanded to

The dataset they used was provided by a company that was running a marketing campaign for a new service. The marketing team labeled different segments according to different criteria to target differently. They believed that the different segments would have different response rates and so they separated the consumers accordingly. They derived a loyalty measure to account for customers’ history with the company. The authors then try to test their primary hypothesis: that the different segments are not independent of each other. They go about this by constructing a network neighbor attribute that indicates whether the targeted customer has been in contact with someone using the service being marketed for. Additionally, they constructed an extra segment that included people deemed not important by the marketing team. They then analyze the data they had and find that network neighbors positively affect the take rate of 20 of the 22 segments. The results were statistically significant. They found that over the entire dataset, the network-neighbors take rates were greater by about 3.4 times. They also highlighted that although the segments are significant on their own, their effect is negated by the network neighbors’ interactions. Thus, they conclude that being a network neighbor is also an important factor to be taken into consideration when constructing these segments. They also find that network neighbors in segment 22 (which was mostly people the company valued pretty low for adoption) have much greater take rates when accounting for network neighbors. Much more than non-network neighbors in the segments originally made by the marketing team. They then experimented by adding more network attributes to see their effect on the ROC-AUC scores. Many of the network attributes had great predictive power on their own and are much more powerful when combined.

**COMMENTS**

**Maximizing the Spread of Influence through a Social Network**

This paper was straight forward for the most part. The only downside was that they discussed two models that might not be very reflective of reality. Firstly, the Linear Threshold model’s problem lies in the very fact that it is linear, and interactions in a network need not be linear in nature and can be much more complex. On the other hand, the cascade model gives every node a single chance to affect a certain other node. This is also not reflective of interactions in real life, for example someone could keep recommending a product they adopted to their family members until some of them eventually cave in. While it is interesting that they were able to restate these models in terms of a submodular function, I would like to see other more complex models that might take such things into account being investigated.

**Network-Based Marketing**

I think this paper was my favorite reading so far. They managed to paint a very good picture by reviewing the literature, explaining their process and dataset properties in details; acknowledging the limitations of their dataset, and trying to extract as much information as possible to support their hypothesis, which they did very well. My only concern is that they did not state how much of the dataset was sourced from outside sources and not from the marketing team of the company. It would be very interesting to examine these properties on more modern networks such as the effect of targeted marketing on YouTube videos and ads on Facebook for example.

**CONCLUSIONS**

Information diffusion seems to be one of the most useful aspects of network analysis so far when it comes to applications in the real world due to its numerous use cases. It was also interesting to find that network properties could have such a significant impact on the flow of information.

REFERENCES

[1] Kempe, David & Kleinberg, Jon & Tardos, Éva. (2003). Maximizing the Spread of Influence through a Social Network. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 137-146. 10.1145/956750.956769.

[2] Hill, Shawndra, Provost, Foster, and Chris Volinsky. "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks." arXiv, (2006). https://doi.org/10.1214/088342306000000222.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00