Network Software Modelling – MIS40550

ANDY MCSWEENEY – 16203177
PETER ADAM – 16201859
22/04/2017

All code and files used for simulations are available for download at:
https://github.com/Padam-0/facebook_viral_advertising

Introduction

The goal of advertisers on Facebook is to reach the largest number of people possible, who are both likely to click on their ad and subsequently, buy their product. Facebook currently has ways to target ads to users based on likes and interests, but it is unclear if Facebook utilises friendship information to optimise who is shown specific ads.

Based on the assumption that users are more likely to click on an ad that a connection of theirs has already clicked on, can the user network be utilized to 'spread' the ad in a viral manner to achieve a similar (or superior) numbers of clicks, while showing the ad to fewer people? This would create a two-fold advantage for Facebook; higher click through rates for their advertising customers, and less ads clogging up user feeds, thus improving the product user experience.

Two metrics are important to measure this; total number of clicks, and clicks per view (CPV). Without influence from the network, to maximise total clicks, the ad should be shown to all users. If each user has an existing probability of clicking on the ad (which is related to other factors), what is the expected number of clicks per view based on this strategy? This forms the benchmark for our analysis.

Conversely, if a user's underlying probability of clicking on an ad is influenced by connections that have already clicked on the ad, then it may make more sense to disseminate the ad throughout the network slowly, to target users when their probability of clicking is higher than in the base case.

To do this, several people are 'seeded' with the ad at the beginning of the simulation and are assumed to have clicked on the ad. From here the organic spread of the ad through the network can be measured. Which users are 'seeded' are based on two different theories and form the test cases for this assignment:

- Those who are most likely to click on the ad based on likes and interests; or
- The most connected people in the graph, or the 'influencers'.

If a user clicks on an ad, that ad is then only shown to a subset of their connections. The hypothesis is that if this subset is too small, the ad will not propagate through the network, while conversely if it is too large, the ad will spread too fast, catching connections at probabilities close to their initial value, and reducing the click through rates. These connections can be selected from one of two lists:

- Strong connections of the user; or
- Weak connection of the user

The total number of connections the ad is shown to, and the distribution of connections between the two lists above is hence referred to as Ad-Serve.

The assumption made is that everyone in the graph all like the same topic (they may all be members of a group devoted to this topic, for instance), and the more active a person is (how much they interact with the group, or other similar pages they like) the more they like the topic. This is information which isn't available to us, hence the use of a random variate to simulate initial probabilities.

If a user clicks on an ad, it is suggested that those users they connect with are now somewhat more likely to click on that ad too. Strongly connected users more-so than others.

The simulation takes this assumption and attempts to propagate an ad through a network slower and thus increase the overall probability of a user clicking on an ad when it is shown to them. While all users will eventually be shown the ad (so Facebook would not lose revenue from ad views), more users may click on the ad because of their increased probability (so Facebook would increase revenue from ad clicks).

Model Design

The graph used for this simulation is undirected and is based on the Facebook 'Social Circles' dataset which consists of 4,039 nodes and 88, 234 edges.

Each node in the graph was assigned a base case probability of clicking on an ad. This probability was determined one way for the benchmark and Highest Click Probability trials, and another for the influencers trial (explained below).

Each edge is assigned a strength, which represents the strength of the connection between the two nodes. This is calculated as the number of shared connections between two nodes, and then if one user clicks on a node, all their connections have their probability is boosted in proportion to the strength of the connections. Strong nodes were deemed to have more than 50% of shared neighbours, and weak nodes less than 50%.

Trials begin by finding the set of the people who are most likely to click on an ad (the initial seed group), and then the ad is served to a subset of their connections (based on Ad-Serve composition). The simulation was run several times with different size seed groups and Ad-Serves.

The initial seed groups ranged in size from 10 to 40 people, and the total size of Ad-Serve also changed throughout this range. If Ad-Serve or initial seed is too large, the ad would propagate through the network too quickly, catching users before their probability had a chance to increase. If these values are too small, the 'virality' of the ad might not be enough to sustain its growth. Additionally, under the assumption that the users in the seed group are those 'guaranteed' to click the ad, and increasing this number too much stresses that assumption.

The simulation concluded when 4, 000 users had seen the ad, or when no new clicks were generated in an iteration. The stopping reason and output statistics were recorded.

Simulations

Benchmark Trial

In this simulation, all users are served the ad at once, and independent Bernoulli trials are conducted to establish how many people in the network click on the ad. For each trail, a new random set of probabilities are added to the Facebook graph, where the probability corresponds to the probability of clicking on an ad, based on an exponential distribution (a few users quite likely to click on the ad, a lot unlikely to click). The results are as follows:

Number of Trials	1000
Number of Clicks	80.538
Standard Deviation	8.68
Base Case CPV	0.0198

Highest Click Probability Trail

The first simulation used the 'viral' methodology described in Model Design. 20 graphs were generated, each with node probabilities from an exponential distribution (mean 0.03). Each Ad-Serve composition was run on each graph 15 times (each time with a different number of users 'seeded') to gain an average that accounts for the stochastic variation in graph composition.

When a user clicked on an ad, all their connections had their probability increased to reflect the assumption that they were now more likely to click on the ad based on their relationship. The maximum increase in probability was 10% for two nodes with strength 1 (all neighbours of both nodes are shared), and decreased linearly in line with the strength of connection.

Two insights are offered from the results. The first is the Ad-Serve composition and seed size which gives the highest possible clicks per view and total clicks, and the second is exploring the relationship between seed size and clicks per view.

The graph below plots the best clicks per view for each Ad-Serve, along with total clicks for that composition:

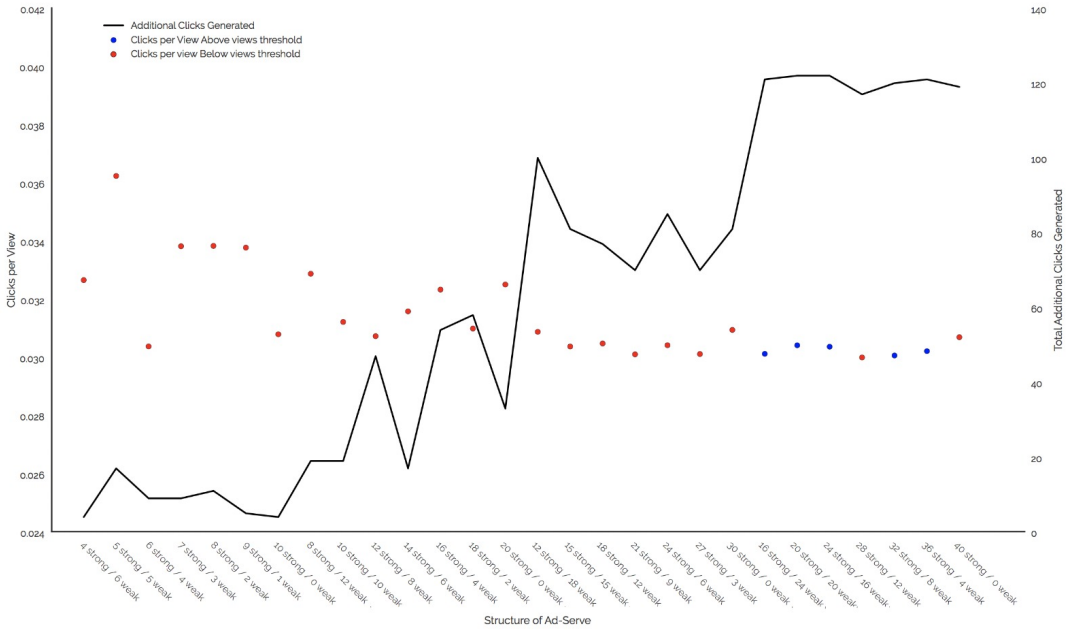


Figure 1

This graph confirms the hypothesis that only showing the ad to 10, 20 or even 30 connections of a user who clicks the ad, causes the viral spread of the ad to die before breaking the 4000 views threshold. Ad-Serve combinations that showed the ad to 40 users all reached the threshold and produced strong numbers of additional clicks (total clicks excluding those seeded for the first iteration).

Interestingly, despite the additional increase in probability that strong connections serve to neighbours, both additional clicks and CPV were only slightly influenced by Ad-Serve composition. This could be because of the strength of weak ties, as while weak connections were less likely to click the ad, if they did, they allowed access to large new areas of the network for the ad to propagate, previously not exposed.

From the graph above, the best performing Ad-Serve ratio was 20 Strong / 20 Weak, seeded with the 12 initial highest probability users. This produced on average 122.45 additional ad clicks from 4025 views, at a CPV ratio of 0.0304, an increase over the base case of 53%.

Following these simulations however, we began to challenge some of the underlying assumptions we had made. We assumed that if an ad was presented to the people who had the highest probability of clicking the ad, that they would click the ad. As there is no extra incentive given to this group, it may be incorrect to assume that if served the ad they would unquestionably click on it. Also, these users were selected randomly based on an exponential distribution (hence the requirement to test each across 20 random graphs), where there was a better way to understand the most important users in a network.

Influencer Trial

The next trial aims to bridge the gap in the 'seeding' assumption and model the real-world trend of social media 'influencers'. Over the past few years, the rise of the 'influencer' on social media has seen massive growth. An influencer can be defined as a someone who has a large social media following and is paid by companies to promote products to their network.

The underlying probability of all nodes was based on their degree, relative to the degree of the node with the largest degree. This produced a somewhat skewed distribution of probabilities, but scaling down by a factor of 0.7 bought the distribution into line with the exponential (mean 0.03) used above. (Appendix, Graph 1) The key difference now is that all nodes had a deterministic probability that was related to their status in the network.

Again, this simulation chooses the most connected people in the graph and 'seeds' them with the ad. While the likelihood that these people will click on the ad is high, their probability is still not 1, which gives the same problems as the trail before. To circumvent this problem however, the assumption is made that these people could have been paid to click the ad, and all other assumptions then hold.

The other key difference from the Highest Click Probability Trial, is that when a user (or in the first iteration, influencer) clicks on the ad, the probability of their connections changes based on 2 factors. The first, the strength of the connection between the two users, is the same but the maximum increase in probability is decreased to 5%. The second is that probabilities are also adjusted based on the influence of the user who served them the ad. This assumption seems to hold in the real world, as intuitively users buy more products advertised by Kim Kardashian than someone they really know, despite the difference in strength of connection. The degree of the user that served them the ad can increase a user's probability by at most 15% for the most connected users, and decreases in proportion to influence.

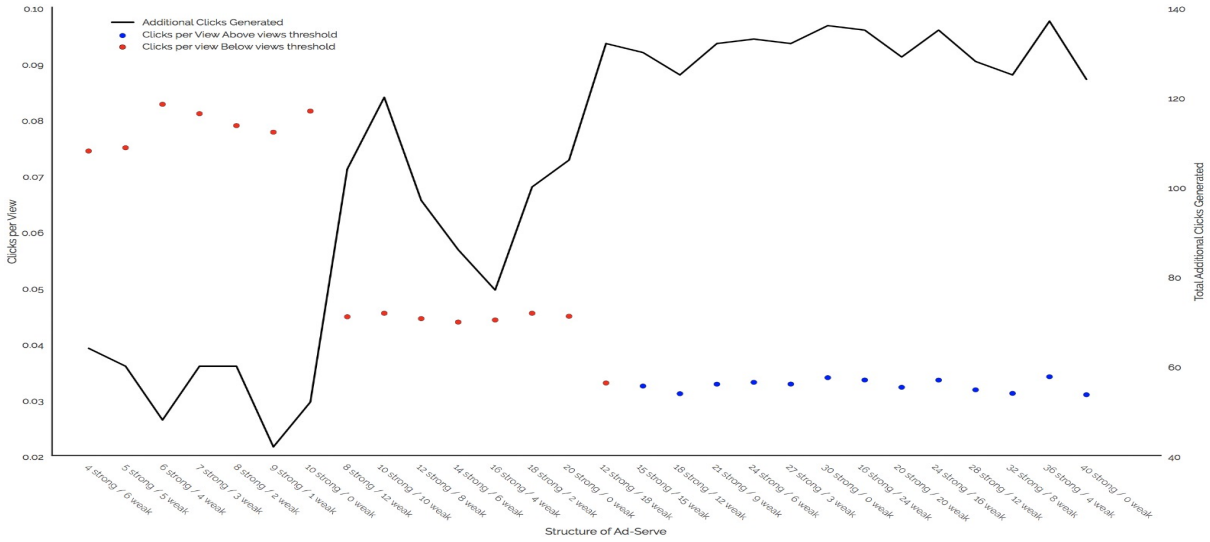


Figure 2

The highest CPV generated was 0.0341 using an Ad-Serve composition of 36 strong and 4 weak, seeded with 24 influencers. This generated an additional 137 clicks. However, the Ad-Serve composition of 16 strong and 24 weak generated a CPV ratio of 0.0335 using only 14 influencers. Depending on the cost per influencer, it may be more profitable to use an Ad-Serve composition that produces a slightly lower CPV but from fewer influencers.

Scalability

To test the scalability of the 'influencer' model created above, the same model was run on larger graphs. To produce a similarly distributed graph structure, the Albert-Barabási preferential attachment model was used to randomly generate graphs with 4,039, 10,000 and 20,000 nodes. This model produced various similar graphs, with degree distributions which closely matched the original Facebook graph (Appendix, Graph 1).

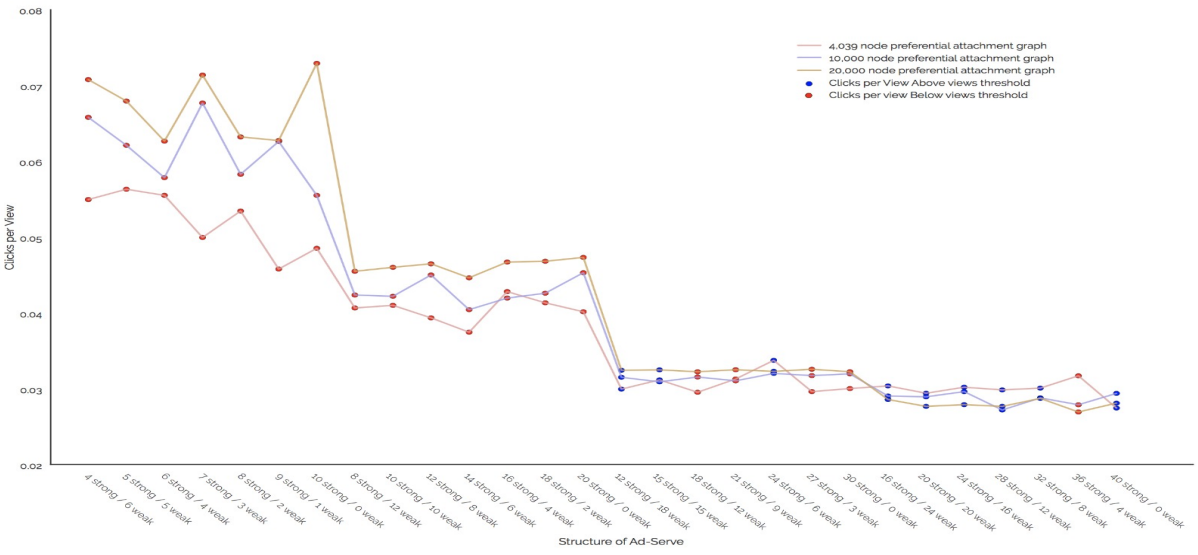


Figure 3

As Figure 3 shows, the results from the preferential attachment graphs closely match the original Facebook graph, thus giving confidence that results for this model is reproducible at scale.

Conclusion

Social media yearly advertising budgets have seen massive growth in the last few years, and are projected to reach $17.34 billion by 2019 in the US alone (LePage, E. 2016). The ability to target huge markets directly, demonstrated by social media companies such as Facebook, Twitter and Snapchat, has given marketing agencies a completely new, and more powerful way to reach their customers.

The primary goal of advertising is to reach the largest group of people who are likely to click their ad and further buy their product. The methods social media companies use to effectively distribute ads through their network need to be beneficial to both themselves and the advertisers. Facebook charges advertisers 'per view' and 'per click', so advertisers must balance their costs against the number of users they expose.

This project aimed to test several theories of how to effectively distribute ads through an undirected social media graph, asking if doing so slowly can produce higher CPV rates than blindly showing an ad to the whole population at once.

Our base case proved to be a good benchmark (0.0198 CPV) to compare our test results with. Our initial test case seeded a set of people with an ad and let the ad naturally spread around the graph based on an Ad-Serve model, and what combinations of 'seed' size and Ad-Serve composition produced the highest CPV. This simulation showed that Ad-Serve combinations with low total degree (each ad shown to 10 connections of a user who clicked it) produced very high CPV (0.06) but low total clicks generated, as the ads propagation died out quickly.

This highlighted the importance of balancing CPV with total additional clicks generated (TACG), as neither Facebook (who would miss out on revenue) nor advertisers (who miss out on ad clicks) would prefer this method. When filtering Ad-Serve compositions to only those that propagated throughout the entire graph, the best case produced a CPV ratio of 0.0304, an increase of 53% over the base case, and 122.45 TACG. The Ad-Serve that produced these results, 20 Strong / 20 Weak shows the importance of weak ties in viral propagation.

While these results were very encouraging, we began to question our initial assumptions and examined whether they would hold true to a real-life scenario. This probing allowed us to find a stronger case to trail our hypothesis on, targeting influencers, or highly connected individuals to act as 'seeds' (instead of random seeds). The rise of social media influencers has been gathering momentum over the past few years but it can be hard to quantify their value to an advertising agency. The second model attempted to recreate this real-world model and study whether ads generating from these people would truly increase CPV.

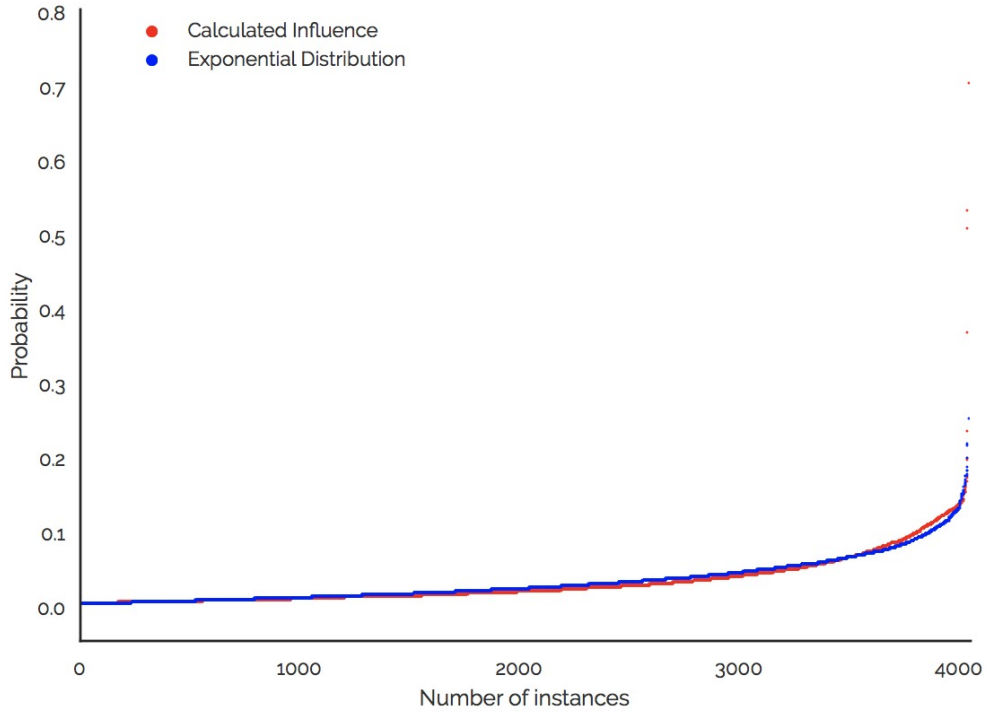
Two optimum cases stood out as excellent candidates for this model. Using an Ad-Serve composition of 36 Strong / 4 Weak, seeded with 24 influencers produced a CPV ratio of 0.0341 and 137 TACG. This was promising, but did use many influencers. Upon exploration, an Ad-Serve composition of 16 Strong / 24 Weak generated a CPV ratio of 0.0335 (135 TACG) using only 14 influencers. As it is assumed a cost will be associated with ensuring each influencer clicks on the ad, the slight reduction in CPV and TACG could be offset by the reduced initial payment to the influencers.

These results were extremely encouraging, showing that Facebook and advertisers could increase ad clicks by 72% by slightly changing the way users are served ads. However, this was on a small-scale network (4039 users), and it remained to be seen if the model was scalable. This was tested using large artificial graphs, generated using the Albert-Barabási preferential model, which closely mimic the degree distribution seen in the original Facebook graph (Appendix, Graph 2). Reassuringly, the model produced very similar results on larger graphs (Figure 3) and gave confidence the model could be scaled to far larger, real world graphs developed by Facebook or Twitter.

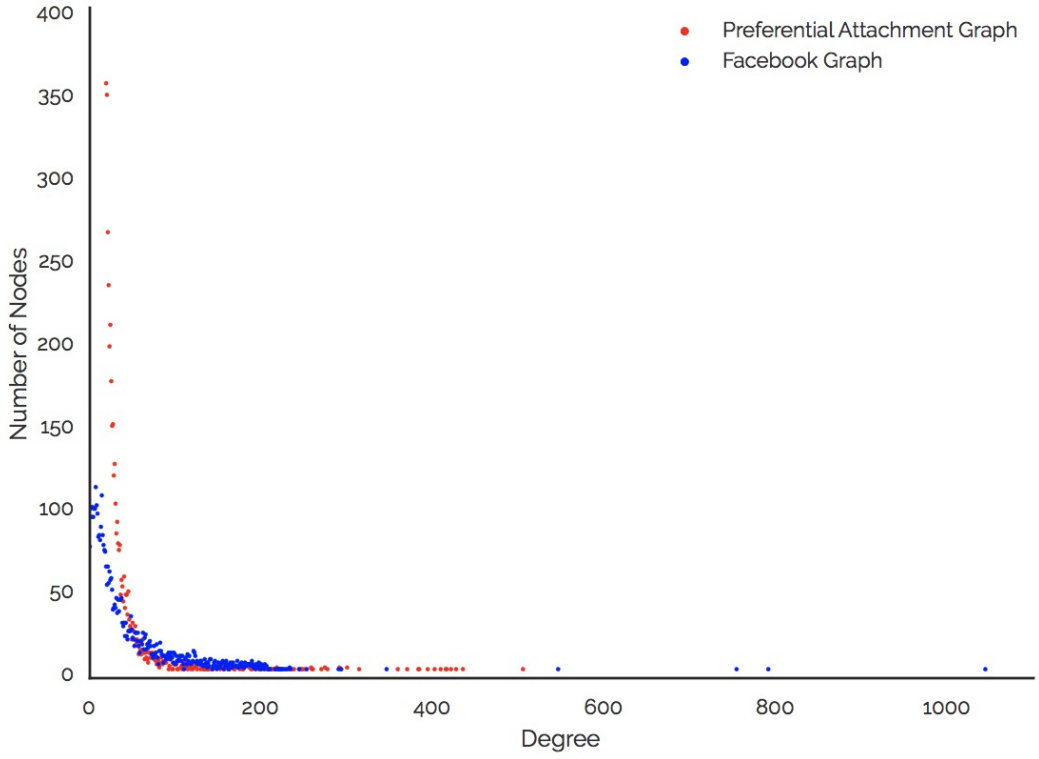
The success of the both the Highest Click Probability and Influencers models rests on the assumption that a user's probability of clicking on an ad increases based on the strength of the connection, or the network influence, of another user who did click (or was paid to advertise) the ad. The values attributed to these factors (10% and 0% in the Highest Click Probability model, and 5% and 15% in the Influencers model) are purely assumptions, and had an impact on the size of CPV increases shown over the base case. The most interesting and important insight from the model is instead, that regardless of what these values are in the real world, if they are positive, using a viral marketing model will generate CPV and TACG increases, and hence revenue increases for Facebook and advertisers.

Appendix

Graph 1



Graph 2



References

LePage, Evan. 2016. "All The Social Media Advertising Stats You Need To Know". *Hootsuite Social Media Management*. N.p., 2017. Web. 21 Apr. 2017.