Quantifying the Influencers and Activists in the Online YouTube Marketplace

Emily Muller (emily@aims.ac.za)

September 27, 2017

1 Abstract

The study proposed will be applied to the YouTube network of *women of colour/black women hair* channels. The online haircare market for this demographic has gained significant momentum in recent years and this work wishes to quantify the network structure of the market. Although this work is generalisable for other markets spanning across various online platforms, the interest in the chosen market comes from the desire to create market knowledge for South African businesses wishing to compete with overseas markets.

The aim of this work is to determine who are the influential users in the YouTube haircare market for black women. This will be done by considering the network of user interactions based on one user commenting on another users material. The activists are considered users who are actively seeking the attention of those influential users greater up in the hierarchy. Additionally, it is to quantify the strategy of these users in assuming such influential positions.

This work will be applied to the particular case study of black female haircare, but nevertheless can be extended to various markets and platforms. The next section will introduce the concept from another perspective to exemplify the problem.

2 Introduction

Suppose you were to follow two scientific podcasts on twitter, for example: *Tell me something I don't know* and *The Infinite Monkey Cage*. It is assumed that each podcast has an overlapping audience and therefore is active in the same market or *network*. The purpose of this work is to address the following questions:

(a). What is the meaning of *Tell me something I don't know* making reference to *The Infinite Monkey Cage's* material?

(b). What is the meaning of *The Infinite Monkey Cage* engaging with their own material?

In order to address the first question, it is assumed that there exists some cost for one user to interact with the other and therefore some perceived utility. The strategic choice of interactions by users will therefore uncover the direction of influence, and the power structure of the network.

In a similar light, there is a perceived utility in responding to your own followers. The second part of this work then aims to determine the behaviours which lead to certain users gaining network dominance. As well as behaviours, such as (b), it will also be necessary to consider attributes associated to the user itself. Are they providing entertainment which is currently trending? What is the frequency of their activity as so forth.

3 Network

The sub-setted network of influential YouTube users will be modelled as a weighted directed network, where nodes are the users (or sellers) and an edge from user U_i to user U_j is a measure of interaction. The following notation will be used:

Notation	Description	Domain
U_i	users/sellers in the network	$i = 1, ..., n.$
I_{jk}	items in the market (videos) from user U_j	$k = 1, ..., \|I_j\|.$
e_ie_{jk}	$\begin{cases} 1 & \text{there exists interaction (comment) from user i to item k of user j} \\ 0 & \text{there is no such interaction} \end{cases}$	{0,1}

Consider first the set of users and videos. These can be modelled as a bipartite graph, with U_i on the left, I_{ik} on the right and edges, e_ie_{jk}, directed from left to right.

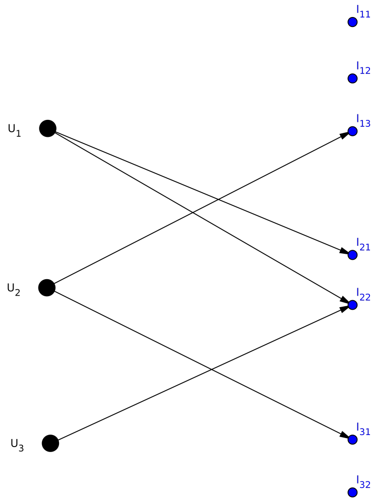

Figure 1: Example of bipartite graph representing network

Figure 1 shows an example of a network consisting of 3 users. Note that an interaction from U_i to I_{jk} is only counted once for each comment. The aim is to derive from the bipartite graph, a weighted directed graph depicting interaction activity between users. It is necessary to consider both, the number of possible interactions from user U_i to $\{I_j\}$ and the activity of user U_i. A method similar to that used in calculating recommender power for collaborative filtering recommender algorithms is used [Sawant(2013)]. The interaction from U_i to U_j is defined as,

$$I(U_i, U_j) = \frac{\#\text{ of edges from } U_i \text{ to } I_{jk} \text{ for all } k}{\text{total } \# \text{ edges sent from } U_i} \cdot \frac{\#\text{ of edges from } U_i \text{ to } I_{jk} \text{ for all } k}{\text{total } \# \text{ items } I_{jk} \text{ for all k}} \tag{3.1}$$

$$I(U_i, U_j) = \frac{\sum_{k=1}^{|I_j|} e_ie_{jk}}{\sum_{j=1}^{n}\sum_{k=1}^{|I_j|} e_ie_{jk}} \cdot \frac{\sum_{k=1}^{|I_j|} e_ie_{jk}}{|I_j|}. \tag{3.2}$$

Therefore, the interaction matrix for Figure 1 is,

$$I = \begin{bmatrix} 0 & 1 & 0 \\ 1/6 & 0 & 1/4 \\ 0 & 1/2 & 0 \end{bmatrix}. \tag{3.3}$$

It is also possible to consider interactions from U_i to I_{ik}, i.e., comments on own items. This measures the response of a user to comments on their own items and can be stored in a diagonal matrix, D, where D_{ii} = average fraction of comments on I_{ik} for which U_i has replied.

4 Analysis

In order to provide meaningful analysis of the interaction matrix, it is necessary to consider both the position of U_i within the network, and furthermore, the dynamic evolution of the network itself. From the perspective of a user, one would wish to participate in a strong marketplace whilst maintaining a strong position within that market. The analysis will therefore be separated in two: network topology, dynamic evolution of the network.

4.1 Network Topology

The aim is to analyse the topology of the network relative to individual actors, and deduce correlations between topological metrics and market competition. Topological metrics such as in and out-degree distributions, clustering and centrality measures such as eigenvector, closeness and betweenness will be considered. In particular, a method of PageRank sorting for weighted graphs will be investigated [Csendes and Antal(2010)] and furthermore dynamic computation of Katz centrality [Grindrod and Higham(2014)] extended for weighted graphs.

4.2 Dynamic Network

It is essential to consider the dynamic nature of the network. The online community is posting new material and engaging in new trends at a rate much faster than the retail market. The aim is to analyse the mechanisms of network evolution leading to certain network topologies, and further investigate the stability and success of such mechanisms. User attributes contained in a feature vector will also be considered in the model (and lend towards specific mechanisms such as homophily [McPherson et al.(2001)McPherson, Smith-Lovin, and Cook]). Examples of features include: videotags, location, interaction with own material. It is assumed that U_i chooses to interact with I_{jk} based on perceived utility in making that interaction. This analyse will be done using the Stochastic Actor-Oriented Model (SAOM) [Snijders(2005)]. The purpose of SAOM model is to determine which structurally dependent network effects lead individuals to select and de-select friendship nominations, thus the name: Actor-Oriented Model. The selection choice is captured by a utility function (the objective function): a linear combination of network effects, ρ_s, plus random residual influence. The probability of selection is a Discrete Choice model [Maddala(1983)]. The coefficients of network effects are estimated using a converging stochastic approximation algorithm based on iterative updates of the selective Markov process. This is an MCMC method. This model considers binary relations between nodes, and therefore, the theoretical framework of this work is to develop a model which will consider weighted relations between users.

5 Data Collection

Data mining can be sub-sectioned as follows:

- Subset the $\approx$ 1.3 million YouTube channels with the tag 'Natural Hair' to contain only channels that align with market criteria and definition of 'influencer'.
- Geomap the channel networks, with particular focus on the U.S.A, U.K, South Africa, Jamaica and Nigeria.
- Scrape the channel videos to collect comment data.
- Using comment data create bipartite network (as per Section 3).

The last two steps to be repeated periodically for time series data. The initial definition of an influencer is to subset the large dataset into something more manageable Here, features such as subscription count and number of video views and likes will be considered. Thereafter, once the dynamic interaction matrix has been built, the task is then to quantify which of these influencers are truly influential, or are actively seeking attention from those influential users.

6 Conclusion

This work aims to identify influencers and activist within online markets. It further aims to analyse the structure and stability of online markets based on the mechanisms of evolution. The application of this work can extend to various online markets. An interesting and viable extension of this work will be to consider the diffusion of trends within the network.

Work done with: Dr Vukosi Marivate, Nyalleng Moorosi, Franck Kalala M & Bubacarr Bah

References

[Csendes and Antal(2010)] T. Csendes and E. Antal. Pagerank based network algorithms for weighted graphs with applications to wine tasting and scientometrics. In *Proceedings of the 8th International Conference on Applied Informatics*, pages 209–216, 2010.

[Grindrod and Higham(2014)] P. Grindrod and D. J. Higham. A dynamical systems view of network centrality. In *Proc. R. Soc. A*, volume 470, page 20130835. The Royal Society, 2014.

[Maddala(1983)] G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press, 1983.

[McPherson et al.(2001)McPherson, Smith-Lovin, and Cook] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[Sawant(2013)] S. Sawant. Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp. In *Proceedings of the CS224W: Social and information network analysis conference*, 2013.

[Snijders(2005)] T. A. B. Snijders. Models for longitudinal network data. In *Models and Methods in Social Network Analysis*, pages 215–447. Cambridge University Press, 2005.